ECCB2022

21st European Conference on Computational Biology

Planetary Health and Biodiversity

BOOK OF ABSTRACTS

12-21 September 2022
Sitges, Barcelona

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Spanish National Bioinformatics Institute

ELIXIR SPAIN

eccb2022.org

Table of content

Poster presentations

Data

A benchmark of library-size bias in correlation for single-cell expression data

Suzanne Jin (Centre for Genomic Regulation (CRG)), Nacho Molina (University of Strasbourg, IGBMC), Ionas Erb (Centre for Genomic Regulation (CRG)) and Cedric Notredame (Centre for Genomic Regulation (CRG)).

Abstract:

Gene expression data can be considered compositional due to an artifact known as “constant-sum constraint”. Consequently, the expression of a gene in different samples needs to be compared with respect to an internal reference (e.g., a library-size normalization). Here we aim to understand repercussions for single-cell co-expression inference in scenarios where normalization assumptions break down. We use both synthetic and experimental data to create a framework in which we can exactly determine the compositional bias and then study how it affects gene-gene association measures. With simulations we create samples with varying mRNA content that preserve gene correlations. Our experimental data are scRNA-seq collected from an asynchronous population of cells that exhibit mRNA total changes by a factor of two along the cell cycle. Not knowing the total mRNA content leads to considerable distortions – a plausible scenario for relative gene expression data. With both the simulation and the experimental set-up, we can compare the amount of co-expression determined on biased data against the ground truth. We benchmark the performance of association measures like correlation and partial correlation in combination with different normalizations, regularization techniques, and zero-handling strategies. Our results show that Pearson correlation in combination with off-the-shelf normalizations can fail spectacularly. Partial correlations on log-ratio transformed data, on the other hand, are the least biased. Additional observations also provide insights on how to improve covariance regularization and shrinkage estimates for compositional data.

Data

A graph-based knowledge base of multi-layer clinical information for the heart failure disease

Irina-Afrodita Balaur (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg), Shaman Narayanasamy (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg), Nikolaus Berndt (Charité Universitätsmedizin Berlin), Muhammad Shoaib (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg), Sarah Nordmeyer (Charité Universitätsmedizin Berlin), Titus Kühne (Charité Universitätsmedizin Berlin) and Venkata Satagopam (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg).

Abstract:

Heart failure, one of the leading death cause in EU, is a complex progressive cardiovascular disease characterized by multiple aetiologies (e.g. volume-pressure overload, ischemic heart disease) and forms - with preserved and with reduced ejection fraction (HFpEF and HFrEF) - in addition to asymptomatic manifestation in some patients, leading to limitations in disease detection and treatment during incipient phases.

Graph databases (GDBs) are widely being used in systems biomedicine due their capacity i) to represent naturally the biomedical information, characterized by large heterogeneous datasets (including omics, clinical, imagining, sensor etc.), ii) to capture complex data inter-relationships and iii) to provide graph-based support for network-based analysis and modeling of biomedical data [e.g. Lysenko et al. (2015) PMID: 27462371; Balaur et al. (2016) PMID: 27627442, Wang et. al (2022) PMID: 35444317].

We will describe the HeartMed GDB, which is a graph-based knowledge base developed in the translational HeartMed EU project that combines (pre-)clinical data from multiple experimental studies for improving patient-specific diagnosis and enabling clinical decision-making in cardiovascular medicine. Specifically, the HeartMed GDB integrates biomedical data corresponding to multiple biomedical layers, including disease biomarkers, drug targets, pathway involvements, patient characteristics, where data (concepts) are represented as nodes and their inter-relationships as edges in the underlying graph. We will present how we use the HeartMed GDB for contextualisation of the heart failure-associated proteins (selected from the initial HeartMed proteomics dataset) and how we transformed the heterogeneous graph into an homogeneous graph focusing on the proteins level to facilitate network-based exploration of key proteins.

Data

A linear optimization approach for normalization of SARS-CoV-2 vaccination data

Simon Cyrani (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Lukas Wenner (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Ferdous Nasri (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam) and Bernhard Y Renard (Data Analytics & Computational Statistics, Hasso Plattner Inst

Abstract:

Having fine-grained vaccination data enables authorities to discover and target deficiencies in vaccination endeavors more easily and focused, leading in the best case to better containment of outbreaks. In many countries there are two factors that hinder more detailed breakdowns of SARS-CoV-2 vaccination data: Firstly, vaccinations are registered solely by place of vaccination, instead of actual residency of the vaccinated person. Additionally, significant proportions of the population in rural areas travel to neighboring, more urban counties, to receive vaccinations. Together, this yields highly distorted vaccination numbers on county level, with some counties having theoretical vaccination rates exceeding 100%. Secondly, these countries lack a vaccination registry which records exact days and vaccine types for each person. It is well established, though, that the effectiveness of SARS-CoV-2 vaccinations decreases over time. Therefore, exact calculations of actual effectiveness depending on the day and type of the last vaccination are not possible, either.

Our work presents a two-part algorithm tackling this problem and delivering more reliable county-wide data, using Germany as an example. In the first step, we reverse the aforementioned effect of vaccination tourism using a flow-based linear programming model. In the second step, we approximate a vaccination registry per county using daily vaccination data of each county's health authority. Combining recent insights in effectiveness of SARS-CoV-2 vaccines regarding contagiousness over time, we calculate lower and upper bounds of the current effectiveness of vaccinations per county via linear programming, and show that these are in fact close to each other.

Data

A Named Entity Recognition Pipeline for Custom Anonymisation of Clinical Free-text Notes

Andreia Rogerio (Genomics England Limited), Christopher Boustred (Genomics England Limited), Anna Need (Genomics England Limited) and Francisco Azuaje (Genomics England Limited).

Abstract:

Sharing clinical data leads to faster collaboration in medical research and ultimately contributes to disease prevention and improved treatments. Unfortunately, ensuring personal identifiable data (PID) is anonymised and compliant with the relevant governance is still challenging, especially in unstructured data.

Here we apply natural language processing (NLP) models for named entity recognition (NER) to detect person names in the Clinical Variant Ark (CVA) of Genomics England. This database contains free-text boxes filled in by clinicians. Upon manual review, 210 records were found to contain PID, 95 in the genomic interpretation (GI) content and 115 in variant comments (VC), which together consist of the evaluation set.

We compare the recall obtained by different open-source NER models, integrated in the anonymisation pipeline tool Presidio. We further add NLP techniques to improve the precision of our solution: 1) we developed a RegEx pattern-matching recogniser that detects literature authors fully integrated in Presidio; 2) we included the pre-trained model ner-disease-ncbi-bionlp-bc5cdr-pubmed to recognise disease names and remove them from final predictions.

The most performant model was Flair with 100% and 85% recall for the VC and GI subsets respectively. These results confirm that Flair outperforms BERT models in NER tasks and show that the precision can be increased by removing person names wrongly detected as PID, from 58% to 80% in VC and 60% to 67% in GI subset.

These results show high potential for Flair applicability in other unstructured clinical datasets, while reinforcing the need for a human-in-the-loop to refine the results' quality.

Data

A simulation-based approach to normalize county-level vaccination data

Lukas Wenner (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Simon Cyrani (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Simon Knott (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Ferdous Nasri (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam) and Bernhard Y Renard (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam).

Abstract:

SARS-CoV-2 vaccinations are known to result in individuals becoming infected less frequently and having milder cases. Therefore, vaccination coverage is critical in assessing which interventions are most appropriate in specific areas at a given time. At present, some countries are missing a central vaccination registry. As a result, vaccinations are reported based on vaccination center locations and not the individual's residence. For example, in Germany, this circumstance leads to counties with over two hundred percent vaccination rates, while others have correspondingly low rates. The data, therefore, requires pre-processing before we can undertake further analysis.

To address this problem, we calculate a daily immunity index at the county level. The index represents an assumed protective rating against infection and consists of two main components. The first component is normalized vaccination data. To normalize, we first rebalance the vaccinations between counties. To achieve this, we created a flow-based graph that distributes vaccinations between nearby counties, observing differences in vaccination rates, population, and distance. Then we simulate the vaccination status of each individual. Based on the vaccine manufacturer, vaccination status, and vaccination age, we determine the protection against infection at any given time. We add the natural protection of recovered people and obtain the immunity index. To verify the result of the vaccination normalization, we look at countries that reported vaccinations by residency and artificially distort these data. Afterward, we use the method described above and compare these results with the original data.

Data

A sustainable workflow for MPRA data analysis: MPRAsnakeflow

Pyaree Mohan Dash (Berlin Institute of Health at Charité – Universitätsmedizin Berlin), Martin Kircher (Berlin Institute of Health at Charité – Universitätsmedizin Berlin) and Max Schubach (Berlin Institute of Health at Charité – Universitätsmedizin Berlin).

Abstract:

Gene expression and its phenotypic effects are regulated by nucleotide variations in so-called regulatory elements. These variants, usually localized in the non-coding parts of the genome, are found to be the cause of human diseases. Hence, the need for accurate quantification of variant effects has given rise to high throughput technologies such as Massively Parallel Reporter Assays (MPRAs) that enable simultaneous testing of thousands of candidate regulatory elements (CREs) and their variants.

We have built a computational tool based on the workflow manager Snakemake, called MPRAsnakeflow, for MPRA data analysis. Our tool overcomes the limitations of custom MPRA data-analysis pipelines tailored to a specific type of MPRA experiment or their specific computing environments. MPRAsnakeflow is an open-source, fast, reproducible, user-friendly tool available on GitHub at https://github.com/kircherlab/MPRAsnakeflow for measuring CRE activity and its variant effects. It is compatible with various MPRA experimental designs including lenti- and episomal-based MPRAs. It covers the inference of a barcode to reporter-sequence assignment alongside count sequencing steps and computes variant effects by contrasting reference and alternative sequences. Moreover, it encapsulates rich quality measurements and additional analyses, for example, barcode overlap statistics between replicates and various downsampling approaches to assess confidence in CRE activity. MPRAsnakeflow assists in deriving a better understanding of MPRA experiments and helps to identify bottlenecks and experimental errors.

Data

A Text Mining approach for retrieving gene::drug::cancer associations from full-text biomedical literature

Elissavet Zacharopoulou (Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly, 35131 Lamia, Greece), Giorgos Skoufos (Dept. of Electrical and Computer Engineering, Univ. of Thessaly, 38221 Volos, Greece), Spiros Tastsoglou (Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly, 35131 Lamia, Greece) and Artemis Hatzigeorgiou (Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly, 35131 Lamia, Greece).

Abstract:

The volume of produced biomedical scientific literature, and the highly diverse and unstructured nature of publications, emphasize the importance of automated information extraction computing. Methodologies of natural language processing (NLP) permit a system to extract information from the available bibliography in literature services such as PubMed Central (PMC). In this study, we describe a text-mining approach to extract the relations between variants and cancer progression. The algorithm that was used implements the creation of a dictionary containing main keywords, such as gene names, variant names, and the different types of cancer. Subsequently, it retrieves publicly available full-text biomedical articles that potentially contain experimental confirmation of a variant's impact on cancer development. These articles are analyzed to the sentence level, to determine entity correlations and features and generate a final database consisting of the exported sentences with additional metadata.

This work displays sets of gene::cancer or variant::cancer association pairs and supplemental drug response information. Out of 19,497 full-text articles – searched and exported with an advanced query - 7,545 publications yielded 73,253 records/sentences, of which 641 were associated with drug response. The final database contains information regarding 31 different cancer types and 3,713 genes. The extent of overlap between this dataset and an in-house collection of variants identified using publicly available whole genomes from ICGC will also be assessed. The text mining pipeline can be utilized for the construction of databases containing gene::drug::cancer associations, providing valuable biomedical assets that facilitate cancer-related research and clinical diagnosis.

Data

A tissue specific post-translational modification (PTM) map of human proteome

Pathmanaban Ramasamy (VIB-UGent Center for Medical Biotechnology), Hanne Devos (UGent / VIB), Lennart Martens (UGent / VIB) and Wim Vranken (Vrije Universiteit Brussel).

Abstract:

Post-translational modifications (PTMs) are covalent modifications that happen on the amino acid residues of proteins after their biosynthesis. These site specific modifications of proteins such as phosphorylation, acetylation, glycosylation, and ubiquitination are involved in almost all signaling pathways that regulate cellular processes. Any deregulation of the proteins involved in such regulatory networks might lead to several diseases such as cancer. PTMs heavily influence the protein structure, folding, function, interaction partners and the sub-cellular localization of proteins. PTMs and enzymes catalyzing these modifications are therefore emerged as prominent therapeutic targets. Although several studies have shown the difference in composition and expression of genes in different tissues/organs through genome and transcriptome profiling, defining tissue specific proteome is still challenging. Given the importance of PTMs in protein function and regulation, it is important to identify the tissue specific and global PTMs to understand the normal development, function and underlying mechanisms for various diseases. In this work, we identify the tissue-specific PTM map of human proteome by large scale re-processing of available public proteomics datasets. We show how re-processing and re-using existing data can add insights on differential PTM sites across tissues and in different pathological states when available.

Data

Agnodice: indexing experimentally supported bacterial sRNA-RNA interactions

Vasiliki Kotsira (DIANA-Lab, Department of Computer Science and Biomedical Informatics, Univ. of Thessaly, Lamia 35131, Greece), Giorgos Skoufos (Department of Electrical & Computer Engineering, Univ. of Thessaly, Volos 38221, Greece), Athanasios Alexiou (DIANA-Lab, Department of Computer Science and Biomedical Informatics, Univ. of Thessaly, Lamia 35131, Greece), Filippos S. Kardaras (DIANA-Lab, Department of Computer Science and Biomedical Informatics, Univ. of Thessaly, Lamia 35131, Greece), Maria Zioga (Department of Computer Science and Biomedical Informatics, Univ. of Thessaly, Lamia 35131, Greece), Spyros Tastsoglou (DIANA-Lab, Department of Computer Science and Biomedical Informatics, Univ. of Thessaly, Lamia 35131, Greece), Theodosia Charitou (Department of Computer Science and Biomedical Informatics, Univ. of Thessaly, Lamia 35131, Greece), Zacharopoulou Elissavet (DIANA-Lab, Department of Computer Science and Biomedical Informatics, Univ. of Thessaly, Lamia 35131, Greece), Nikos Perdikopanis (DIANA-Lab, Department of Computer Science and Biomedical Informatics, Univ. of Thessaly, Lamia 35131, Greece) and Artemis G. Hatzigeorgiou (DIANA-Lab, Department of Computer Science and Biomedical Informatics, Univ. of Thessaly, Lamia 35131, Greece).

Abstract:

We present Agnodice (https://dianalab.e-ce.uth.gr/agnodice/#/), a curated collection of interactions between sRNAs and other bacterial RNAs provided as an online database. The content of the database is exclusively experimentally supported and incorporates Low- and state-of-the-art High-Throughput methods.

Bacterial small regulatory RNAs (sRNAs) are non-coding RNAs, typically ranging between 50-500 nucleotides. Their main function is the positive/negative post-transcriptional regulation of gene expression, achieved through imperfect base-pairing with their mRNA targets, in response to stress or environmental stimuli. Recent studies demonstrate the importance of sRNAs during bacterial infection, in the modulation of antibiotic resistance and in other crucial processes. In order to gain ground on understanding bacterial gene regulatory mechanisms, the function of sRNAs and their targeting repertoires need to be further investigated and properly catalogued.

We designed and implemented a database comprising more than 20,000 experimentally supported sRNA-RNA interactions, linking more than 250 small RNAs with the regulation of 4,700 RNAs at strain level resolution, by manually curating the existing literature and performing re-analysis of publicly available Next-Generation Sequencing datasets from relevant techniques (e.g., RIL-Seq, CLIP-Seq, CLASH). In an effort to cater consistently high-quality results throughout the database, interactions were meticulously annotated using the same reference resources (NCBI Taxonomy). Studies containing low-throughput experimental evidence were derived by manually querying PubMed, using relevant keywords (e.g., “sRNA”, “Hfq”, “interaction”), while a text mining approach was also deployed to maximize the initial volume of the curation set.

We aspire that Agnodice will serve as a valuable resource for microbiologists and scientists from other relevant fields.

Data

AN INTEGRATIVE BIOINFORMATICS PIPELINE FOR THE ANALYSIS OF A COMPREHENSIVE NGS DIAGNOSTICS PAN-CANCER PANEL

Raúl Marín (Bellvitge Biomedical Research Institute (IDIBELL) - Catalan Institute of Oncology (ICO)), Ania Alay (Bellvitge Biomedical Research Institute (IDIBELL) - Catalan Institute of Oncology (ICO)), Sara Hijazo-Pechero (August Pi i Sunyer Biomedical Research Institute (IDIBAPS) - Hospital Clínic de Barcelona), Ernest Nadal (Bellvitge Biomedical Research Institute (IDIBELL) - Catalan Institute of Oncology (ICO)), Víctor Moreno (Bellvitge Biomedical Research Institute (IDIBELL) - Catalan Institute of Oncology (ICO)) and Xavier Solé (Hospital Clínic de Barcelona).

Abstract:

Gene sequencing panels have emerged as a key diagnostic tool for the application of precision oncology in the clinical practice. However, commercial panels often provide proprietary and non-customizable tools which generate numerous output metrics without visual reports, thus hindering the subsequent interpretation of results. By using tumor sequencing data from the Illumina TSO500 panel as input, here we present an open and customizable bioinformatics pipeline to (i) identify tumor mutations and copy number (CN) variations, (ii) measure immuno-oncology biomarkers: tumor mutational burden (TMB) and microsatellite instability (MSI), and (iii) provide graphical reporting. Up to now, tumor alterations from 356 patients have been identified using both our pipeline and the commercial one (TSO500). Regarding small variants, our pipeline implements an integrative approach of multiple callers and external databases to generate a consensus and more robust list of variants annotated with genetic, functional, and clinical information. Additionally, this pipeline identifies not only most variants reported by TSO500 (~98%), but also additional ones that could be relevant in the diagnostics process. In terms of TMB scores, a very strong correlation (R=0.95) with TSO500 is obtained. Regarding CN, TSO500 only reports CN values of 59 genes, while our pipeline can calculate CN scores for all the 523 genes, including informative visual plots that enable the interpretation of the results. Our study shows how the implementation of a customized pipeline provides more insight about the genomic alterations in clinical samples, and it can be updated on a continuous to incorporate further bioinformatics advances.

Data

Analysis of phenotypes using ontology and word embedding

Canh Nguyen-Duc (Department of Evolutionary Developmental Genetics, Georg-August-University Göttingen), Gregor Bucher (Department of Evolutionary Developmental Genetics, Georg-August-University Göttingen) and Jürgen Dönitz (Department of Medical Bioinformatics, Georg-August-University Göttingen).

Abstract:

To represent and organize phenotypic data, many databases such as Flybase and Mouse Genome Database are using ontologies. An ontology, readable by both human and machine, comprises terminologies, their attributes and relationships with other terminologies which all together describe a specific domain of knowledge. Here in iBeetle-Base, the database of phenotypes obtained by silencing Tribolium castaneum genes using RNA interference, we utilize ontology to annotate and perform phenotype-based analyses to reveal interesting patterns among Tribolium genes and across species. Each phenotype in iBeetle-Base consists of 1) the morphological entity (head, antena) and 2) the aspect and nature of the change of the entity (not present, size decreased). The former is best described by TrOn, the Tribolium-specific morphological ontology. Using the graph structure of TrOn, we compute distances between every pair of annotated phenotypes, cluster them to obtain a phenotypic summary for each gene and then compare the genes at phenotype level as opposed to the ubiquitous sequence-level comparison. The latter can also be represented by an ontology such as Phenotypic and Trait Ontology (PATO). However, relying on the fact that the latter part of our phenotypes is composed of common words, we use word2vec model to create phenotype embeddings which can reveal the semantic similarity of words without the human effort of mapping phenotypes to ontology terms. Our results are made available in iBeetle-Base as resources to support Tribolium community as well as any interested researcher.

Data

Analyzing socio-demographic biases in cellphone mobility data used for COVID-19 contact predictions

Conrad Halle (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Ferdous Nasri (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Jeremias Dötterl (Machine Learning Unit, Department of Engineering, NET CHECK GmbH, 10829 Berlin, Germany), Steven Schulz (Machine Learning Unit, Department of Engineering, NET CHECK GmbH, 10829 Berlin, Germany), Sten Rüdiger (Machine Learning Unit, Department of Engineering, NET CHECK GmbH, 10829 Berlin, Germany) and Bernhard Y Renard (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam).

Abstract:

Human mobility data generated by millions of mobile devices is a vital resource in the analysis of epidemic spread, and it has become even more important during the current COVID-19 pandemic. However, in many cases, it is unclear how representative these datasets are for the general population. More specifically, the behavior of different sociodemographic groups within the pandemic is not yet fully understood and cannot be inferred from data if the biases are not taken into account during data analysis.

To examine these biases, we analyzed a cellphone-based human mobility dataset along with its derivative estimation for contact numbers, which is used to estimate SARS-CoV-2 spread in Germany. We combined this dataset with sociodemographic data from the official German Census survey using an approximation of a home location for each device. The publicly available sociodemographic data has a spatial resolution of 100x100m. The combination of datasets made it possible to detect biases in the representation of certain age groups in the population. We achieved this by assigning each device a relative age distribution based on its home location, to then compare the average age among the devices to the age distribution of the general population. We applied a similar approach to analyze these biases also regarding the contact numbers within different age groups. It showed that people with an age between 40 and 70 are significantly overrepresented in the data, while the contact behavior of the age groups is changing over time, highlighting the potential need for a rebalanced dataset.

Data

Assessing SARS-CoV-2 evolution through the analysis of emerging mutations

Anastasios Mitsigkolas (Faculty of Science, Vrije Universiteit Amsterdam, The Netherlands), Nikolaos Pechlivanis (Institute of Applied Biosciences, Centre for Research and Technology, Hellas) and Fotis Psomopoulos (Institute of Applied Biosciences, Centre for Research and Technology, Hellas).

Abstract:

Intro:

Since the beginning of COVID-19 pandemic, the number of SARS-CoV-2 related studies including a phylogeny of the virus is constantly increasing. However, a lot of these studies have also revealed the difficulties of inferring a reliable phylogeny, especially given the highly similar sequences and the relatively low number of mutations evident in each sequence (https://dx.doi.org/10.1093/molbev/msaa314). Identifying the evolutionary history of the SARS-CoV-2 virus is thus considered challenging yet necessary, in order to assist the phylogenetic analysis process, keep track of the virus and its characteristic mutations, and ultimately find patterns of the emerging mutations. This study is offering a computational solution to the problem of detecting new patterns of co-occurring mutations beyond the strain-specific/strain-defining ones, in SARS-CoV-2 data, through the application of ML methods.

Methods:

Moving beyond the traditional phylogenetic approaches, we designed and implemented a graph representation learning method, relying on novel modeling of the sequences as binary vectors. This process ultimately creates a dendrogram of the involved mutations, that can potentially be used to identify trends in emerging variants. Our proposed method has been tested out in publicly available sequences and validated using the assigned Pango lineages.

Results:

Our method makes it possible to identify correlated mutations, the correlation of which might indicate an underlying “association” between the corresponding lineages. However, highly correlated mutations do not necessarily indicate causal relationships among emerging mutations. Thus, we conclude that the results are preliminary and further research is needed to address this topic.

Data

ASTERICS: A Tool for the ExploRation and Integration of omiCS data

Élise Maigné (Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France), Céline Noirot (Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France), Jérôme Mariette (Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France), Yaa Adu Kesewaah (Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France), Sébastien Déjean (IMT, UMR5219, Université de Toulouse, CNRS, UPS, 31062, Toulouse, France), Camille Guilmineau (Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France), Julien Henry (Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France), Arielle Krebs (Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France), Laurence Liaubet (GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet-Tolosan, France), Fanny Mathevet (Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France), Hyphen-Stat (Hyphen-stat, https://hyphen-stat.com/, Toulouse, France), Christine Gaspin (Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France) and Nathalie Vialaneix (Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France).

Abstract:

The rapid development of omics acquisition techniques has induced the production of a large volume of heterogeneous and multi-level omics datasets measured on the same individuals. Complex information of biological interest is obtained from so-called integration methods, which have been increasingly developed in the past few years. Some of these methods are already available in R packages (like mixOmics). However, the use of these packages still requires to learn a programming language and to have access to sufficient statistical knowledge to choose method parameters and interpret outputs. We present ASTERICS, a web application that aims at making complex exploratory and integration analysis workflows easily available to biologists. Data edition, exploration and integration menus organize the interface to perform data edition, missing value imputation, normalization, data exploration with interactive plots, numerical summaries, PCA, tests, clustering, self-organizing maps, and data integration with various methods. Analyses are adapted to the most standard omics datasets. ASTERICS is also designed to make the analysis flow understandable with a navigable workspace that displays uploaded or obtained datasets and performed analyses in a graph. Finally, it also comes with a documentation for beginners that helps interpret the results, choose proper options or the next analysis to perform. ASTERICS is based on Rserve, pyRserve, and flask. R package versions are controlled using renv. Frontend is developed in Vue.js and uses the CSS framework Bulma. It is available online at http://asterics.miat.inrae.fr/ and can be installed from source or using our docker deployment, both distributed under GPLv3.

Data

Benchmarking methods for the identification of mislabeled data in genomics

Lusine Nazaretyan (Berlin Institute of Health at Charité – Universitätsmedizin Berlin), Martin Kircher (Institute of Human Genetics, University Medical Center Schleswig-Holstein, University of Lübeck) and Ulf Leser (Institute for Computer Science, Humboldt-Universität zu Berlin).

Abstract:

Machine learning recently gained growing importance in biomedical research. To train reliable models, bioinformaticians need credible data, which is not always available. A particularly hard and widespread problem are mislabeled samples (Northcutt CG et al, 2021). For instance, prior disease diagnoses might be overturned due to research progress. Another common source of mislabeling are weakly defined labels, labels that change their meaning, or labels annotated by different groups following different guidelines or having different evidences at hand. In this regard, Harrison SM et al. 2017 found that around 17% of variants submitted to NCBI ClinVar have conflicting interpretations, such as being labeled as "benign" and as "likely pathogenic". Because mislabeling leads to deteriorating prediction quality, it is essential for scientists to be able to identify wrong labels efficiently and effectively.

Here, we benchmark various methods for the identification of mislabeled instances that can be applied to high-dimensional omics data. In addition to experiments on datasets with artificially introduced noise at controllable levels, we also report results on real-life genomic datasets with known mislabeling. We find that most of the methods perform well on datasets with a high amount of noise but fail to find noisy instances when the proportion of wrong labels is low. Furthermore, none of the methods excels over all others in isolation, while ensemble-based methods often outperform individual models. We provide all data sets and code to enable a better handling of mislabeling and to foster further research in this field.

Data

BEstimate: a python tool for in silico analysis of potential edits with CRISPR base editors

Cansu Dincer (Wellcome Sanger Institute), Matthew Coelho (Wellcome Sanger Institute) and Mathew Garnett (Wellcome Sanger Institute).

Abstract:

Genome editing techniques enable the precise manipulation of DNA sequences1. Base editors (BEs) can generate alterations on targeted nucleotides without needing donor DNA or causing double-strand breaks or indels2. Thus, they have the potential to scale-up functional analysis of genes, and to model genetic diseases by engineering disease-associated variants. In this study, we created BEstimate, a computational tool that systematically maps gRNA targetable sites across given genes for BEs. It provides a user-defined interface to find gRNAs for any BEs by adjusting protospacers, activity windows, PAM regions and nucleotide changes, as well as providing gRNA off-target information. BEstimate provides clinical, functional and structural effects of the potential edits by implementing Ensembl, Uniprot and Variant Effect Prediction Application Programming Interfaces. For given input, BEstimate annotates potential edits for their locations on the gene and protein sequences as well as whether they are clinically relevant, damaging or whether they disrupt post translational modification or protein interaction sites. The tool also provides gRNAs specific for mutated sequences to improve the efficiency and accuracy of editing, and to facilitate the programmed reversion of disease-associated variants. In summary, BEstimate is a flexible tool that enables the design of gRNA libraries for a broad range of base editing applications.

1. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science 346, 1258096 (2014).

2. Rees, H. A. & Liu, D. R. Base editing: precision chemistry on the genome and transcriptome of living cells. Nat. Rev. Genet. 19, 770–788 (2018).

Data

BLOOD LIPID PROFILE AS AN INDICATOR OF PSYCHIATRIC DISORDERS

Anna Tkachev (V. Zelman Center for Neurobiology and Brain Restoration, Skolkovo Institute of Science and Technology), Anna Morozova (Mental-health Clinic No. 1 named after N.A. Alexeev of Moscow Healthcare Department), Yana Zorkina (Mental-health Clinic No. 1 named after N.A. Alexeev of Moscow Healthcare Department), Alexander Reznik (Mental-health Clinic No. 1 named after N.A. Alexeev of Moscow Healthcare Department), Denis Andreyuk (Mental-health Clinic No. 1 named after N.A. Alexeev of Moscow Healthcare Department), Elena Stekolshchikova (V. Zelman Center for Neurobiology and Brain Restoration, Skolkovo Institute of Science and Technology), Nickolay Anikanov (V. Zelman Center for Neurobiology and Brain Restoration, Skolkovo Institute of Science and Technology), Georgiy Kostyuk (Mental-health Clinic No. 1 named after N.A. Alexeev of Moscow Healthcare Department) and Philipp Khaitovich (V. Zelman Center for Neurobiology and Brain Restoration, Skolkovo Institute of Science and Technology).

Abstract:

There is an increasing interest in the role lipids play in physiological processes, while their potential use as biomarkers is expanding, as well. Identifying reliable and reproducible blood plasma lipid alterations is particularly promising for psychiatric disorders, where, currently, no clinically useful diagnostic tests exist. Here, we use measurements on lipid abundance in blood plasma to predict disease status of patients with schizophrenia.

Lipids were assessed using an untargeted evaluation methodology, allowing us to assess most of the main lipid classes present in blood plasma, including glycerolipids, glycerophospholipids, sphingolipids. Using the produced lipidomics measurements, we defined a logistic regression diagnostic model.

A total of 119 blood plasma samples of psychiatric patients and healthy controls were collected. The proportion of correctly identified schizophrenia (SCZ) and first psychotic episode (FEP) patients was 0.82 and 0.84 (sensitivity), respectively, while the specificity was found to be 0.95. Interestingly, schizotypal (SCZtyp) individuals showed a high variability in prediction labels, as the distribution of predicted probabilities for these samples was quite different than the distributions for both SCZ/FEP and control. These results are in line with the classification of SCZtyp disorder as an intermediate schizophrenia-spectrum phenotype.

Our results indicate a robust signature of lipid abundance in blood plasma separating SCZ from control. They also showcase the potential applicability of the model for at-risk individuals with milder psychiatric conditions, such as SCZtyp individuals.

Data

Breast Cancer Detection in UKBiobank Data Using Mutation Based Gene Weighting and Deep Learning

Nuriye Özlem Özcan Şimşek (Boğaziçi Unversity), Fikret Gurgen (none) and Arzucan Ozgur (Bogazici University).

Abstract:

Breast cancer is the most common cancer type in the UK. About 1 in 8 women are diagnosed with breast cancer during their lifetime [1]. For diagnosis, mammography or biopsy might be needed. If detected at an early stage, it can be treated by surgery, chemotherapy, radiotherapy or a combination of these.

In our study, we propose a blood sample based system for the detection of breast cancer. It requires dna sequencing and variant calling by only using a blood sample. The method was published in 2019 [2]. The published study was applied on tumor samples. In this study, we prove the success of our method by applying on blood sample based data. Our new dataset is composed of VCF files from UKBiobank [3]. The case set is 6264 breast cancer patients and the control set is 6264 samples which are selected randomly among samples with no cancer diagnosis. The accuracy of our system on the new dataset is 97.76 +- 0.34 which is slightly higher than the published result. The f-score is 96.66 +- 0.51 which is the same as the published result.

The performance metrics show that our method achieves consistent performance with both tumor and blood samples. Detection of the disease with blood samples will eliminate the need of biopsy and improve early detection rate.

References

1. NHS website, https://www.nhs.uk/conditions/breast-cancer/

2. Şimşek et'al. Statistical representation models for mutation information within genomic data. BMC Bioinformatics. 2019;20:32

3. UKBiobank website, https://www.ukbiobank.ac.uk/

Data

CABiNet – Biclustering and joint cell-gene visualization of single-cell transcriptomics

Yan Zhao (Max Planck Institute for Molecular Genetics), Clemens Kohl (Max Planck Institute for Molecular Genetics) and Martin Vingron (Max Planck Institute for Molecular Genetics).

Abstract:

Clustering cells and the subsequent determination of marker genes for cell type annotation are routine tasks in single-cell RNA-seq data analysis. These tasks fundamentally try to solve a biclustering problem. We here propose "Correspondence Analysis for Biclustering on Networks" (CABiNet) as a novel approach to co-cluster cells and cell type-specific marker genes in a single step. CABiNet utilizes the geometrical properties of Correspondence Analysis to construct a Shared Nearest Neighbor graph of cells and genes and uses graph clustering algorithms to detect the biclusters. CABiNet also allows a joint visualization of the biclustering results by a non-linear embedding approach, the biUMAP.

We show that in the application to single-cell RNA-seq data, be it simulated or real, CABiNet proves to be a simple yet powerful algorithm not only for cell clustering but also for cell-gene biclustering. In comparison to other cell clustering algorithms, CABiNet obtains comparable clustering results regarding the cells while additionally determining cluster-specific genes. Compared to other biclustering algorithms, CABiNet delivers better biclustering results combined with fast runtime.

Data

Can we trust probabilities in deep drug activity models? A comparative calibration study.

Hannah Rosa Friesacher (ESAT-STADIUS, KU Leuven, 3001 Leuven, Belgium), Lewis Mervin H (Molecular AI, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK), Ola Engkvist (Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden), Yves Moreau (ESAT-STADIUS, KU Leuven, 3001 Leuven, Belgium) and Adam Arany (ESAT-STADIUS, KU Leuven, 3001 Leuven, Belgium).

Abstract:

Uncertainty estimation for classification tasks can provide valuable information in decision making processes. In drug discovery, well calibrated probabilities are useful to reflect the true uncertainty associated with prediction to allow well-informed decisions during the development of therapeutic agents by giving an estimate of the expected success rate. In our work, we compare a wide range of common probability calibration approaches, including Platt scaling, ensemble modeling and Bayesian methods, to assess their effect on the multitask deep-learning model SparseChem [1]. SparseChem enables the prediction of target-ligand interactions from sparse bioactivity data with high accuracy, but poor probability calibration. The results of our study revealed an unexpected underperformance of Platt scaling [2] and ensemble based techniques, which are commonly used methods for calibrating probabilities. Additionally, we compared the probability calibration of the conventional SparseChem model to Bayesian approaches using the Markov Chain Monte Carlo method Hamiltonian Monte Carlo (HMC) [3] as well as models trained on single tasks. Furthermore, the calibration error of models trained using different optimization methods (eg.: ADAM, SGD) were compared.

[1] Arany, A., et al. (2022). arXiv preprint arXiv:2203.04676.

[2] Platt, John. Advances in large margin classifiers 10.3 (1999): 61-74.

[3] Neal, Radford M. Bayesian learning for neural networks. Vol. 118. Springer, 2012

Data

ClustAssess: tools for assessing the robustness of single-cell clustering

Irina Mohorianu (Cambridge Stem Cell Institute), Arash Shahsavari (Cambridge Stem Cell Institute) and Andi Munteanu (Cambridge Stem Cell Institute).

Abstract:

The transition from bulk to single-cell analyses refocused the computational challenges for high-throughput sequencing data-processing. The core of single-cell pipelines is partitioning cells and assigning cell-identities; extensive consequences derive from this step; generating robust and reproducible outputs is essential. From benchmarking established single-cell pipelines, we observed that clustering results critically depend on algorithmic choices (e.g. method, parameters) and technical details (e.g. random seeds).

We present ClustAssess, a suite of tools for quantifying clustering robustness both within and across methods. The tools provide fine-grained information enabling (a) the detection of optimal number of clusters, (b) identification of regions of similarity (and divergence) across methods, (c) a data driven assessment of optimal parameter ranges. The aim is to assist practitioners in evaluating the robustness of cell-identity inference based on the partitioning, and provide information for choosing robust clustering methods and parameters.

We illustrate its use on three case studies: a single-cell dataset of in-vivo hematopoietic stem and progenitors (10x Genomics scRNA-seq), in-vitro endoderm differentiation (SMART-seq), and multimodal in-vivo peripheral blood (10x RNA+ATAC). The additional checks offer novel viewpoints on clustering stability, and provide a framework for consistent decision-making on pre-processing, method choice, and parameters for clustering.

Data

Concordant or not? Performance and concordance rates analysis of ten prediction algorithms on clinically relevant variants from the BRCA1 and BRCA2 genes.

Erda Qorri (Biological Research Centre, Institute of Genetics, HCEMM-BRC Mutagenesis and Carcinogenesis Research Group, Szeged), Bertalan Takács (Biological Research Centre, Institute of Genetics, HCEMM-BRC Mutagenesis and Carcinogenesis Research Group, Szeged), Alexandra Gráf (Biological Research Centre, Institute of Genetics, HCEMM-BRC Mutagenesis and Carcinogenesis Research Group, Szeged), Márton Z. Enyedi (Delta Bio 2000 Ltd., Szeged. HCEMM - Single Cell Omics Advanced Core Facility, Szeged, Hungary), Lajos Pintér (Delta Bio 2000 Ltd., Szeged), Ernő Kiss (Biological Research Centre, Institute of Genetics, HCEMM-BRC Mutagenesis and Carcinogenesis Research Group, Szeged) and Lajos Haracska (Biological Research Centre, Institute of Genetics, HCEMM-BRC Mutagenesis and Carcinogenesis Research Group, Szeged).

Abstract:

The widespread use of next-generation sequencing technologies in clinical diagnostics has resulted in the identification of thousands of novel genomic variants whose clinical significance is unknown. To address this issue, a variety of computational tools have been developed that facilitate the interpretation of these variants of uncertain significance. However, the vast range of these tools and the ambiguous guidelines often make it challenging for the lay user to choose the appropriate software. In this regard, systematic benchmarking is crucial for evaluating the efficiency of the prediction algorithms and selecting the best-performing ones.

By using two independent benchmarking datasets composed of missense variants from two routinely screened genes in clinical settings, BRCA1 and BRCA2, we evaluated the performance of ten widely used prediction algorithms. In addition, based on current guidelines for variant interpretation which encourage the use of computational analysis as supporting evidence only when multiple prediction algorithms are used and are in agreement with each other we analyzed the inter-rater agreement, and false concordance rates across different pairs of prediction algorithms.

Our results show that the performance of the prediction algorithms varies widely across the datasets. Furthermore, based on the Area Under the Curve (AUC) and Matthews Correlation Coefficient (MCC) values, the computational tools were categorized into two major categories: best-performing and poor-performing algorithms. Moreover, we also demonstrate that by equally weighing the output of multiple computational tools, the current guidelines allow for poor-performing algorithms to disrupt the congruence of the better-performing algorithms, thus rendering the computational evidence ineffective.

Data

Covid-19 Mutation Incidence Forecasting using Spatio-Temporal Graph Neural Networks

Larissa Hoffaeller (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, 14482 Potsdam), Athar Khodabakhsh (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, 14482 Potsdam) and Bernhard Renard (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, 14482 Potsdam).

Abstract:

In this work, we analyze daily incidence numbers of Covid-19 mutation data to comprehend the spreading patterns between cities over time. Since the start of the pandemic, Covid-19 test samples have been collected by various laboratories within a country. Systematically selected samples from infected individuals have been sequenced to obtain their mutation type. The variation of daily case numbers of a specific mutation in different cities can be modeled as a spatio-temporal network. The nodes represent the location of the laboratories where the samples were collected. In each timestep, edges are added between nodes that have cases to model the potential spread from one node to another. We remove a node's incident edges once their cases drop back to zero. When a location has no cases, we assume no potential spread could have happened. We consider the geographical distance between cities and information about commuting population behavior as edge features to represent the spatial dependencies. Graph Neural Networks are then employed to extract the topological characteristics of the spatio-temporal network. Recurrent Neural Networks take the resulting embeddings as input to learn the dynamic behavior. Future edges are predicted based on previously seen edges. Thus, we can infer which edges potentially have an impact on the next spreading step. This approach is evaluated on mutations of multiple variants to compare the outcomes in different phases of the pandemic. The experimental results forecast which regions might have infected other regions to track the spread of the Covid-19 virus over time.

Data

Creation of an UHPLC-ESI-QTOF-MS Library and Application on Seafoods

Sapna Sharma (Technical University of Munich), Corinna Dawid (Technical University of Munich) and Sebastian Dirndorfer (Technical University of Munich).

Abstract:

A collection of natural reference compounds with mass ranges from 50 to 1000 Da was analyzed by means of UHPLC-ESI-QTOF-MS using the information-dependent acquisition mode (IDA) in order to create in-house reference libraries. We have developed an automated pipeline that stores the key information such as the retention time, high resolution precursor m/z values, high resolution production ion spectrum (MS/MS) as well as meta data describing the molecular structure (InChI, InChIKey, SMILES) and the taxonomic classification. Key advantage of our tool is that it includes the reference compounds in their natural form as well as derivatized using the 3-NPH/EDC method enhancing the ionization and retention capability of otherwise hard to capture compounds using LC-ESI-MS. Obtained IDA data from reference compound measurements and their respective meta data were combined and transferred to a versatile applicable .msp universal library file format for each polarity and the derivatized reference compounds by developing a pipeline solely relying on free of charge open-source software (MS-DIAL, MS-FINDER, Python) capable of handling the raw data from different vendors of LC-MS systems. The performance of the created libraries was determined by applying them on several sample sets from ongoing projects from our lab. The acquisition was performed using the data-independent acquisition (DIA) mode SWATH (Sequential Window Acquisition of all Theoretical Mass Spectra) allowing a comprehensive acquisition of product-ion spectra of even low abundant compounds in complex matrices.

Data

Cross-Mapper for Structural Domains (CroMaSt): A workflow for domain family curation through cross-mapping of structural instances between protein domain databases

Hrishikesh Dhondge (Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France), Isaure Chauvot de Beauchêne (Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France) and Marie-Dominique Devignes (Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France).

Abstract:

Protein domains can be viewed as building blocks, essential for understanding structure-function relationships in proteins. However, each domain database classifies protein domains using its own methodology. Thus, in many cases, boundaries between different domains or families differ from one domain database to the other, raising the question of domain definition and enumeration. The answer to this question cannot be found in a single database. Rather, expert integration and curation of various databases are required to refine the contours of a domain of interest, in a domain-centric approach.

Here, we illustrate the role of 3-D structure in clarifying domain definition with the help of CroMaSt: "Cross-Mapper for Structural Domains", a fully automated workflow that classifies all structural instances of a given domain into 3 different categories (core, true and domain-like). CroMaSt is developed in Common Workflow Language (CWL) and takes advantage of 2 well-known and widely used domain databases, Pfam (sequence-based) and CATH (structure-based). It uses the domain definitions from Pfam and CATH and SIFTS resource for cross-mapping of structural instances from the above-mentioned sources. Structural alignments generated by Kpax allow to identify the false positive instances from each domain database.

We tested CroMaSt on the RNA Recognition Motif (RRM), the most prevalent and diverse RNA-binding domain. Starting from PF00076 and 3.30.70.330 domain families from Pfam and CATH respectively, our workflow identifies 882 core, 1120 true and 89 domain-like structural instances. The information generated by this method will play a crucial role in machine learning methods applied to domain-specific synthetic biology.

Data

Data and Computing Platform to facilitate NCER-PD (National Center of Excellence in Research in Parkinson's disease) project

Rajesh Rawal (University of Luxembourg), Carlos Vega (University of Luxembourg), Soumyabrata Ghosh (University of Luxembourg), Sascha Herzinger (University of Luxembourg), Kirsten Roomp (University of Luxembourg), Peter Banda (University of Luxembourg), Piotr Gawron (University of Luxembourg), Rejko Krüger (University of Luxembourg), Jens C. Schwamborn (University of Luxembourg) and Venkata P. Satagopam (University of Luxembourg).

Abstract:

NCER-PD research project focuses on improving the diagnosis and stratification of Parkinson's disease (PD) by combining clinical and molecular data to develop novel disease signatures.

The Data and Computing Platform (DCP) provides key infrastructure for the integration, curation and interrogation of anonymized clinical and experimental data. DCP manages multidimensional clinical data, bio-sample associated information, and multi-omics data. These different data flows are integrated through advanced data capture and transfer approaches. Clinical data is recorded at clinical-visit time, assuring integrity and standardization. To this end, REDCap1, a state-of-the-art clinical research data management system was implemented. Once collected, data becomes decoupled, with personal data safely stored in SMASCH2 and clinical information stored in REDCap. Thus, data is anonymized and sample-associated data is securely accessed directly at their storage location, the IBBL, with the LIMS of the biobank. For cohort study reporting, data exploration, and real-time monitoring of study progress, we developed Ada3, providing an analytics and discovery platform through integrated analytics. Ada was extended towards management of sample metadata and molecular omics data. Based on its graphical user interface, further extensions are available for integrating different workflows, pipelines, and machine learning approaches. In 2021, support for OpenID Connect4 was implemented. OMICS data-types were further supported by allowing join-function across different datasets, integrating clinical data, sample metadata and omics data, as scientific questions often require multi-modal data. In summary, DCP and an analytics dashboard platform improved the user experience of NCER-PD researchers and clinicians.

Data

Data integration analysis for profiling host-microbiome interactions in non-responsive Celiac disease patients.

Laura Judith Marcos-Zambrano (Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute. Madrid, Spain.), Blanca LaCruz (Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute. Madrid, Spain.), Viviana Loria-Kohen (Nutrition and Clinical Trials Unit, GENYAL Platform IMDEA-Food Institute, Madrid, Spain.), Ana Ramirez de Molina (Nutrition and Clinical Trials Unit, GENYAL Platform IMDEA-Food Institute, Madrid, Spain.) and Enrique Carrillo de Santa Pau (Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute. Madrid, Spain.).

Abstract:

Data integration enables understanding of the microbiome's role in health and disease from a more comprehensive perspective. We aim to study the association of microbiome taxonomic profiles with the persistence of symptoms in non-responsive Celiac Disease patients (NR-CeD) - a pathology characterised by the persistence of symptoms from CeD after one year on a gluten-free diet- from an integrative perspective.

We include 39 NR-CeD patients. Symptomatology data and GFD compliance were recorded through validated questionnaires. Markers associated with mucosal integrity and inflammation were measured, and the gut microbiome was analysed through shotgun metagenomics.

We selected the most informative clinical variables with Multiple factor analysis, then performed a Gaussian mixture model for clustering over the Canberra distance of the selected variables, and found two clusters of patients. Cluster 1: "Low-grade symptoms" (n=25), characterised by lower symptoms and inflammatory markers. Cluster 2: "High-grade symptoms" (n=14), characterised by higher punctuation in the patient-reported symptoms questionnaires, elevated inflammatory markers, and higher intestinal permeability. We constructed co-occurrence networks through sparse inverse covariance estimation and model selection to perform the microbial community structure analysis. We found differences according to each cluster in the predominant keystone taxa, network properties and the microbial metabolic pathways present. We found that the microbiome of Cluster 1 patients was enriched in SCFA producers bacteria and the eukaryotic Blastocystis spp. Also, contain metabolic pathways related to amino acid biosynthesis and nitrogen metabolism. In contrast, Cluster 2 patients were characterised by a microbiome enriched in gram-negative and proteolytic bacteria.

Data

Data Management in Galaxy

Beatriz Serrano-Solano (Forschungszentrum Jülich / University of Freiburg) and Bjoern Gruening (University of Freiburg).

Abstract:

The increasing amounts of data generated by scientific research poses the challenge of providing an adequate infrastructure and tools that facilitate FAIR (Findable, Accessible, Interoperable and Reusable) data access, manipulation, analysis and visualization. Often, the burden of managing the metadata associated with the original data and the analysis lies with the researchers.

The open source Galaxy platform is well-known for supplying tools and workflows for reproducible and transparent data analysis across scientific disciplines. However, Galaxy is more than a workflow manager: it provides scientists with access to reference data, databases (ENA, UniProt, NCBI, PDB, Ensembl…), external repositories (FTP, SFTP, Dropbox…), data sources through standard APIs (TRS, DRS from GA4GH); enriches the metadata during the analysis to finally enable mechanisms to export (S3, ENA…) and share the results of the analysis.

There are three large Galaxy instances (US, Europe and Australia) used by hundreds of thousands of researchers worldwide and that are using PBs of data. Galaxy handles the metadata transparently, releasing scientists from the burden and making it less prone to human errors. These features can be used in various ways depending on the user profile, from scientists without technical background that can simply use a web browser; to more computationally-skilled ones that are willing to perform the analysis programmatically through the Galaxy API.

This poster will describe how the Galaxy platform assists researchers from diverse technical backgrounds and scientific domains across the whole data life cycle: data access, processing, analysis, preservation, sharing and reusability.

Data

Dereplication with DEREPer: a tool for High Throughput Metabolomics with LC-MS/MS

Simone Zorzan (LIST), Kjell Sergeant (LIST) and Sophie Charton (LIST).

Abstract:

The untargeted knowledge on metabolites in samples would provide essential information. However, standardization, throughput, and depth of analysis for metabolites is low mainly because metabolite identification is slow, driven by manual interpretation of mass spectra. Although targeted metabolite analyses allow for higher throughput and improved standardization, they do not give an as exhaustive view on sample composition as untargeted metabolite analyses. Dereplication is the use of previous data to identify compounds would speed up this process, thereby liberating time to identify “new” compounds. The objective of DEREPer is to provide a workflow for the dereplication of metabolite identification in LC/MS-MS.

This software implements a database for compounds memorization. The database allows storing compound details, such as name, formula, m/z, protocol, project, organism, references, peaks list etc.) and to store data for the matches of an already known compound across experiments. A graphical user interface allows to input and edit a compound or a match to an existing one. Fragmentation patterns can be visualized. When a new compound has the same formula as an existing one, the user can thus verify if the two compounds are the same or not. Data can be exported in binary format, or in the standard .msp text format, for compatibility with external software.

DEREPer overcomes the “identification bottleneck”, and metabolite identification will be quicker. More and more diverse metabolites will be identified, making metabolomics a high-throughput approach.

Data

Designing a Cloud and HPC Based Model&Simulation platform to Investigate Diseases Mechanisms

Maria Paola Ferri (INB - Barcelona Supercomputing Center), Laia Codó Tarraubella (INB - Barcelona Supercomputing Center), Josep Lluis Gelpi Buchaca (Universitat de Barcelona, INB - Barcelona Supercomputing Center), Dimitrios Lialios (Barcelona Supercomputing Center) and Francesco Gualdi (Instituto de investigaciones médicas Hospital del Mar).

Abstract:

The development of an automated and specialized platform can represent the best hybrid technology with perks on both healthcare data management and computational environments: though rendering automatic not only the database, but prediction and simulation models in a user-friendly integrated system, may facilitate a difficult diagnosis and forward therapy, especially considering the various forces at play in a multi-omics prognosis. Based on the European Open Science Cloud (EOSC) vision, and within the HORIZON MSCA Disc4All project, a platform for IVD (Intervertebral DIsc Degeneration) Models & Simulations (M&S) tools would furnish an easy-to-use and easy-to-understand environment, exposed on a front-end, to guarantee reproducibility, accessibility for experts and non-experts. The ultimate purpose of this infrastructure, deployed on the biological, bioinformatics algorithms, image analysis and ML/AI models and tools, is to integrate and interpolate primary patient data and achieve a coherent attribution of a MS (multi-factorial musculoskeletal) phenotype. To get a homogeneous configuration between all the tools, all the tools and softwares would be adjusted to a similar framework, the BioBBs (BioExcel Building Blocks). Python-based, with their mutual unique syntax, they will offer a layer of interoperability between the wrapped tools, to facilitate the building process of a biomolecular/bioinformatics workflow, and also offer portability for their instantiation in a variety of environments, such as Galaxy, Virtual Environments and HPC.

BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows. Pau Andrio et Al. Nature Scientific Data,Volume 6, Issue 1, p.169, (2019).

Data

DiASPora: A Natural Language Processing platform for automated extraction of microbial phenotype information from scientific articles

Arindam Halder (ZB MED), Ziyad Ziyad (ZB MED) and Konrad U Förstner (ZB MED).

Abstract:

BacDive ((https://bacdive.dsmz.de) is one of the leading databases for structured and manually curated data about bacterial and archaeal strains encompassing information about phenotypic traits like shape, metabolic substrates, growth conditions, and more. The large-scale analysis and integration of microbial biodiversity information have been impeded by heterogeneity and fragmentation of data sources in conjunction with a lack of tools to extract information in a scalable manner accurately. The advances made in Natural Language Processing (NLP) and Deep-Learning (DL) have enabled the development of tools to efficiently develop models for fast and efficient information extraction. This works aims at automating the process of information extraction about bacterial phenotypes and integrating them into BacDive We developed a pipeline with the following major components: (1) Collection of relevant scientific articles and text mining; (2) extraction of phenotypic information as defined in BacDive; (3) Development of a web-based text annotation tool and use the expert annotations to create a feedback loop to improve the underlying DL model for information extraction; and (4) Creating information extraction models based on training data generated from text annotation.

Data

Digital Heart (DHART) Research: A multi-omics resource portal for the cardiac community

Etienne Boileau (University Hospital Heidelberg), Christoph Dieterich (University Hospital Heidelberg), Harald Wilhelmi (University Hospital Heidelberg) and Enio Gjerga (University Hospital Heidelberg).

Abstract:

Integrative omics holds a promise to revolutionize medical research. With recent advances in high-throughput biology, novel methods of analysis such as machine learning are used to pinpoint molecular mechanisms associated with disease, to identify novel markers, and to drive fundamental and clinical research. While multi-omics has become the workhorse of biomedical research, data accessibility and usability still remains a critical issue. Extracting knowledge from these complex data has become a challenge for the life scientist with little expertise in bioinformatics. These challenges have not yet been addressed by the cardiac community.

Here, we present the Digital HeART (DHART) Research portal. The DHART portal provides a platform for simultaneous exploration of multiple datasets and data modalities, empowering scientists to interrogate data in public and private domains.

The portal brings together a large resource of cardiac data to allow a deeper insight into the molecular mechanisms of the cardiovascular system.

By providing the tools for data sharing, data mining, hypothesis generation, and knowledge discovery, the DHART portal provides an online platform for multi-omics analytics and visualization, offering the cardiac community the potential to leapfrog unilateral benchside research.

This work builds on the gEAR framework [1] and on our expertise in multi-omics and database management [2,3], and is currently underpinned by large infrastructures to enable long-term efficient, secure data management, adhering to FAIR principles.

[1] Nat Methods 18, 843–844 (2021)

[2] J Mol Cell Cardiol 150, 23-3 (2021)

[3] Nucleic Acids Res 43, (Database issue):D160-7 (2015)

Data

digitalDLSorteR: An R package to deconvolute bulk RNA-Seq from scRNA-Seq data

Diego Mañanes Cayero (Centro Nacional de Investigaciones Cardiovasculares), Carlos Torroja (Centro Nacional de Investigaciones Cardiovasculares), Carlos Relaño Ruperez (Centro Nacional de Investigaciones Cardiovasculares), Inés Rivero García (Centro Nacional de Investigaciones Cardiovasculares) and Fátima Sánchez-Cabo (Centro Nacional de Investigaciones Cardiovasculares).

Abstract:

Despite recent advances in scRNA-Seq, deconvolution of bulk RNA-Seq remains a key issue in scenarios in which the analysis at the single cell level is too expensive or experimentally challenging. Several methods have been hence developed to tackle this issue, but most of them rely on a prior selection of cell type markers and disregard the biological context of the samples, e.g. tissue or tumor type. In this work, we present digitalDLSorteR, a new R package that allows building context-specific deconvolution models based on Deep Neural Networks using scRNA-Seq data as input. These models are able to make accurate estimates of the cell composition of bulk RNA-Seq samples from the same context using the advances provided by Deep Learning and the rich information provided by single-cell RNA-Seq data.

Data

Disentangling spatial and non-spatial transcriptomic signals: A modified variational auto encoder model

Loïc Chadoutaud (Institut Curie, PSL Research University, Paris, France), Andrei Zinovyev (Institut Curie, PSL Research University, Paris, France) and Emmanuel Barillot (Institut Curie, PSL Research University, Paris, France).

Abstract:

Spatial Transcriptomic technologies have enabled researchers to study gene expression whilst not losing information about tissue organization. Methods based on generative modeling, such as Variational Auto Encoder and its many variants, have proven to be a powerful tool for analyzing scRNA-seq data. However, these approaches developed for single cell data cannot be directly used for this kind of data as they do not allow the integration of spatial information with gene expression patterns.

Here we plan to develop a model based on a variational auto encoder to take into account this new type of information. Depending on the level of prior information available, we proposed to make use of either pathologist annotations when available or simply spatial coordinates. Compared to classical approaches, we aim at obtaining two latent spaces that each highlight a different important feature of the data. To do this, we proposed to borrow techniques from the domain adaptation field or from the research in disentanglement representation learning.

With this approach, we can make the latent space used in the generative process more interpretable by separating out the main factors of variability. We hypothesize that when spatial annotations are not available, this method can identify spatial domains, which we can compare with the underlying image. It could also be used to group genes according to the variability factor they most closely match. Overall, we hope it will enable new biological discoveries by taking advantage of the spatial organization of tissues that has been missing until now.

Data

Efficient and rapid genome profiling of SARS-CoV-2 by covSonar

Alice Wittig (Robert Koch-Institut), Kunaphas Kongkitimanon (Robert Koch-Institut) and Stephan Fuchs (Robert Koch-Institut).

Abstract:

The SARS-CoV-2 pandemic, along with increasingly affordable sequencing technologies, has led to unprecedented sequence coverage. At the peak of the BA.2 wave in March 2022, the Robert Koch Institute received over 25,000 genome transmissions per week via the German Electronic Sequence Data Hub (DESH, https://zenodo.org/record/6759033). Challenging tasks, which are no longer manually possible given the vast amount of data, are the reliable derivation of genomic profiles, the efficient linkage of genome data and metadata, and the fast filtering, accessing, and display of requested mutation profiles depending on their metadata.

Here, we introduce covSonar (https://github.com/rki-mf1/covsonar), a lightweight and SQLite-supported data management system for derivation of genomic profiles based on optimal global alignments. covSonar stores mutation profiles at amino acid and nucleotide level, supports multiple reference and segmented genomes, and imports custom metadata. As of June, the covSonar database generated from DESH sequences and the NC_045512.2 reference contains 927,995 SARS-CoV-2 genome sequences, 45,363,192 mutations and 4,692,602 indels. The data import is fast and efficient through a non-redundant sequence caching with automated genome profile tests and parallelization of alignments that enables daily data updates. covSonar has an optimized design for efficiently functional application and fewer consumption resources, and a normalized SQLite database schema with a class structure for best query performance. Moreover, it comes with an user-friendly query framework and common output formats such as VCF and CSV that allow quick data queries without knowledge of SQLite.

Data

Evaluation of the bioinformatics tumour-control approach in a next-sequencing panel in pediatric leukemia

Beatriz Ruz-Caracuel (Translational Research in Pediatric Oncology, Hematopoietic Stem Cell Transplantation and Cell Therapy, IdiPaz), Carlos Rodríguez-Antolín (Experimental Therapies and Novel Biomarkers in Cancer. IdiPAZ.), Carmen Rodríguez Jiménez (Genetics of Metabolic Diseases Laboratory, Hospital Universitario La Paz, IdiPAZ), Sonia Rodríguez Novoa (Genetics of Metabolic Diseases Laboratory, Hospital Universitario La Paz, IdiPAZ), Victoria Eugenia Fernández Montaño (Structural and Functional Genomics, Department of Genetics, Hospital Universitario La Paz), Victoria Gómez del Pozo (Structural and Functional Genomics, Department of Genetics, Hospital Universitario La Paz), Inmaculada Ibáñez (Experimental Therapies and Novel Biomarkers in Cancer. IdiPAZ.), Javier de Castro (Experimental Therapies and Novel Biomarkers in Cancer. IdiPAZ.), Adela Escudero López (Translational Research in Pediatric Oncology, Hematopoietic Stem Cell Transplantation and Cell Therapy, IdiPaz) and Antonio Pérez-Martínez (Pediatric Hemato-Oncology Department, La Paz University Hospital).

Abstract:

Leukaemias are a common disease in the pediatric population where the tumour content of the samples analyzed is usually low. Through high throughput sequencing, we have designed a validation cohort to evaluate the performance of 7 tumour-only and 8 tumour-control approaches in a DNA capture panel with pediatric oncology genes. The cohort consists of samples obtained as a result of mixing two NIST samples (HG001 and HG002) at 1, 4, 10 and 25% frequencies to simulate low-frequency variants and samples with low tumour content.

We evaluated the performance of both approaches in a context of targeted gene panels in which we sequenced the “tumour” samples with average mean depths in the range 200-1000X and the “control” samples in the range 60-300X.

With the current methodology and in targeted gene panel context, it seems more cost-effective to sequence different patients at greater depth than to sequence tumour and control samples in parallel. The first approach allows to characterize tumours in depth, whereas the second one does not remove all the “noise variants”. Moreover, in some cases it is not able to recover somatic variants as them if they have an allelic frequency that the algorithm considers as germline variants.

Data

Event segmentation in highly modified Nanopore direct RNA sequencing

Wiep van der Toorn (Systems Medicine of Infectious Disease (P5), Robert Koch Institute), Patrick Bohn (HIRI, Helmholtz Centre for Infection Research), Liuwei Wang (Systems Medicine of Infectious Disease (P5), Robert Koch Institute, International Max-Planck Research School BAC), Redmond Smyth (HIRI, Helmholtz Centre for Infection Research, Faculty of Medicine, University of Würzburg) and Max von Kleist (Systems Medicine of Infectious Disease (P5), Robert Koch Institute).

Abstract:

Structural probing methods are used to study the structure of RNA by introducing modifications at unpaired residues. Modification patterns in the sequenced RNA inform structural models, therefore a high modification density is favorable. Nanopore direct RNA sequencing (DRS) allows for sequencing full length RNA molecules, which has the potential to greatly improve structural probing analysis, as well as the detection of endogenous RNA modifications. Currently, however, the analysis of highly modified DRS data is challenging. Analysis of DRS data is done by segmenting the signal in 5mer events with a 'base-caller' model. Available base-callers are trained on unmodified RNA and are not equipped to handle highly modified signals which may have different 5mer signal distributions. Post-processing the erroneous results (e.g. with Oxford Nanopore's Tombo 'resquiggling') causes an accumulation of errors in downstream analyses. Correct segmentation of signals of highly modified reads is fundamental to capitalize on the promise of DRS for structural probing, as well as, for example, to prepare datasets to train base-callers for modified signals. In this study, we explore methods to accurately detect 5mer change-points in highly modified reads. We propose to use the inherent stochasticity in dwell time per 5mer to iteratively improve a fast Bayesian segmentation using Dynamic Time Warping Barycenter Averaging. We show that the obtained segmentation can be used to detect modifications in highly modified reads and can thus strongly improve analysis of direct RNA sequencing for detecting RNA modifications.

Data

Exploiting Pretrained Biochemical Language Models for Targeted Drug Design

Gökçe Uludoğan (Bogazici University), Elif Olmez (Roche AG), Nilgün Karalı (Istanbul University), Kutlu Ö. Ülgen (Bogazici University) and Arzucan Ozgur (Bogazici University).

Abstract:

The development of novel compounds targeting proteins of interest is one of the most important tasks in the pharmaceutical industry. Deep generative models have been applied to targeted molecular design and have shown promising results. Recently, target-specific molecule generation has been viewed as a translation between the protein language and the chemical language. However, such a model is limited by the availability of interacting protein-ligand pairs. On the other hand, large amounts of unlabeled protein sequences and chemical compounds are available and have been used to train language models that learn useful representations. In this study, we propose exploiting pretrained biochemical language models to initialize (i.e. warm start) targeted molecule generation models. We investigate two warm start strategies: (i) a one-stage strategy where the initialized model is trained on targeted molecule generation (ii) a two-stage strategy containing a pre-finetuning on molecular generation followed by target specific training. We also compare two decoding strategies to generate compounds: beam search and sampling. The results show that the warm-started models perform better than a baseline model trained from scratch. The two proposed warm-start strategies achieve similar results to each other with respect to widely used metrics from benchmarks. However, docking evaluation of the generated compounds for a number of novel proteins suggests that the one-stage strategy generalizes better than the two-stage strategy. Additionally, we observe that beam search outperforms sampling in both docking evaluation and benchmark metrics for assessing compound quality.

Data

Exploring Chemical Diversity with the Chemical Checker

Martino Bertoni (Institute for Research in Biomedicine (IRB)) and Patrick Aloy (Institute for Research in Biomedicine (IRB)).

Abstract:

Screening large and diverse sets of compounds is central to any virtual screening (VS) effort. Selecting a subset of these purchasable compound libraries is necessary to optimize any High Throughput Screening (HTS) experiment and can impact significantly costs and outcomes.

Classical structural and physicochemical descriptors have been widely used to this extent. However, thanks to the wealth of annotations, small molecules and drugs can be described also taking into account their biological activity. Recently, we integrated the major chemogenomics and drug databases in a single resource named the Chemical Checker (CC), which is the largest collection of small-molecule bioactivity signatures available to date. We trained a collection of deep neural networks able to infer bioactivity signatures for any compound of interest, even when little or no experimental information is available for them. Our signaturizers relate to bioactivities of 25 different types (including target profiles, cellular response and clinical outcomes) and can be used as drop-in replacements for chemical descriptors in day-to-day chemoinformatics tasks. Here, we push this approach to its limits applying it on very large datasets of diverse molecules. We perform comparative analyses and present selection strategies for dissecting purchasable pre-plated or focused libraries with the aim of contextualizing their chemical matter content under the different facets that affect the biological activity of molecules.

Data

Exploring the development of innate immune responses to infection in early life: A call for data.

Mary McCabe (Wellcome-Wolfson Institute for Experimental Medicine, Belfast, United Kingdom.), Dr Guillermo Lopez Campos (Wellcome-Wolfson Institute for Experimental Medicine, Belfast, United Kingdom.), Dr Helen Groves (The Royal Belfast Hospital for Sick Children, Belfast, United Kingdom) and Professor Ultan Power (Wellcome-Wolfson Institute for Experimental Medicine, Belfast, United Kingdom.).

Abstract:

Chronological age is a major risk factor for respiratory viral infection. Children <2 years old experience more respiratory viral infections, which are often more severe and result in hospitalisations compared to older children. However, the mechanisms behind respiratory viral hypersusceptibility in early life remain unclear. The respiratory epithelium is the primary site of infection for most respiratory viruses, with innate immune responses to infection likely key determinants of disease severity. We hypothesised, therefore, that innate immune responses to infection within the respiratory epithelium develop with chronological age, independent of the infecting virus. To address this, we assessed the possibility of using in silico approaches based on the identification and reanalysis of published datasets.

A search query was designed incorporating three concepts: 1) respiratory epithelial cells, 2) early life and 3) the infecting virus. Gene expression data of seven common paediatric respiratory viruses and SARS-CoV-2 were queried within the NCBI GEO and EMBL-EBI ArrayExpress databases. To determine differential expression with chronological age the R package limma-voom was used (Ritchie et al., 2015).

One dataset was identified that met the inclusion criteria: GSE117827. Statistically significant differential expression was only present with infection (controls vs infected).

However, no statistically significant differential expression was detected when comparing responses with chronological age.

Despite respiratory viral infection constituting a major cause of childhood morbidity and mortality, there is a lack of respiratory expression data to explore innate immune responses to infection in early life. Future data generation and availability are needed to address this issue.

Data

Extending the potential of plasmiR in the cancer microRNA biomarker landscape

Marios Miliotis (Hellenic Pasteur Institute/DIANA-Lab, Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly), Dimitris Grigoriadis (Hellenic Pasteur Institute/DIANA-Lab, Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly), Spyros Tastsoglou (Hellenic Pasteur Institute/DIANA-Lab, Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly), Nikos Perdikopanis (Hellenic Pasteur Institute/DIANA-Lab, Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly), Athanasios Alexiou (Hellenic Pasteur Institute/DIANA-Lab, Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly) and Artemis Hatzigeorgiou (Hellenic Pasteur Institute/DIANA-Lab, Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly).

Abstract:

plasmiR (www.microrna.gr/plasmir) is a manually curated collection of circulating microRNA (miRNA) biomarkers with experimental support. miRNAs are steady-state RNAs, amply detected in circulation. Their differential abundance in pathophysiological states, as well as their association with disease outcomes and treatment responses, are widely studied, offering the unique potential of yielding accurate, minimally invasive biomarkers. Relevant information on miRNA biomarkers is inconveniently dispersed across articles and supplements, underlining the need for a comprehensive and systematically annotated database which, in contrast to existing resources, specifically focuses on highlighting the biomarker validation choices, the employed experimental and statistical methods, along with extensive cohort details. plasmiR is user-friendly, providing interconnections with reference miRNA/disease resources and encouraging smart queries, cross-disease contrasts and hypothesis-free explorations.

Here, we attempt to dive into the circulating miRNA/cancer associations catered in plasmiR and inspect the direction and effect size of dysregulation in the corresponding disease tissue, by utilizing publicly available data from the International Cancer Genome Consortium. Variation events identified in the associated miRNA regulatory elements (e.g., binding sites, precursor-genes, promoters), potentially disrupting their fine-tuning mechanisms, are also studied, gaining further insight into the underlying mechanisms that take place in the disease tissue. This valuable extra resource will permit direct comparison of the abundance of cancer-associated miRNAs between circulating and tissue states quantified under the same cancer type, as well as integration with important genomic dysregulation events, further assisting studies related to the roles and origins of circulating miRNAs under distinct pathological conditions.

Data

FAIR by Design - FAIRification Strategy in Large Data Centric Projects

Soumyabrata Ghosh (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg), Irina-Afrodita Balaur (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg), Basile Rommes (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg), Kavita Rege (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg), Hanna Ćwiek-Kupczyńska (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg), Wei Gu (Luxemb

Abstract:

We are involved in multiple large-scale data centric consortia including Innovative Medicines Initiative (IMI)-BIOMAP which require a FAIR strategy (making data Findable, Accessible, Interoperable, Reusable) for efficient data management and analytics. While Biomap has a specific disease area related to Atopic Dermatitis and Psoriasis, the wide spectrum of data, originated from various clinical cohorts, and the diversity of analytical requirements are challenging. To address such challenges from both technical and data management perspectives, we designed a “FAIR by design” strategy in the consortium. The strategy has been implemented at different layers of the data platform architecture namely data capture, curation, visualisation and access. The data platform, which is hosted at LCSB’s data and computing infrastructure, was built using open-source tools following FAIR data standards and workflows recommended by scientific communities including ELIXIR and FAIRplus. We also focused on reproducible research and developed customisable data engineering components that are being reused in other translational projects. Here, we will present the planning and implementation of the IMI-Biomap FAIR data architecture and we will discuss the initial challenges and the ways for solving them, in addition to the lessons learnt. We conclude that the FAIR data architecture presented here can be employed or easily customised for use in other data centric projects handling FAIRification of heterogeneous large-scale biomedical datasets.

Data

Far tail approximation of non-standard test statistic distribution

Krzysztof Mnich (University of Bialystok), Wojciech Lesiński (University of Bialystok) and Witold Rudnicki (University of Bialystok).

Abstract:

Apart from usual statistical tests, non-standard techniques are often used to identify informative variables in biological data analysis.

These include various estimates of mutual information, various measures of multivariate interactions, analysis of importance of the variable in a machine learning model or influence on the results of simulations.

In the most cases, however, the null-hypothesis distribution of the score of a variable is not known.

To find out, whether the result is statistically significant, the distribution is usually compared with one obtained for random variables.

The "brute force" method consists in comparison between the score of the variable under scrutiny with top scores of random ones.

It requires several dozens times more synthetic variables than the size of the original data set to obtain a reasonable precision with FWER correction, which may be very expensive computationally.

Instead, we propose to utilise the observation, that the far tail of null distribution is often close to an exponential function of the score.

Fitting the exponential tail to the empirical distribution proves to be an efficient approach only for a narrow class of distributions.

Therefore, we introduce a two-parameter approximation, which can be fitted to a broader class of distributions, from normal to power one.

The method was tested on artificial data with known distribution, and then used for analysis of real-world biological data.

It proved to be much more computationally efficient than the commonly used "brute force" approach at the same precision.

Data

Forest-Guided Clustering - Explainability for Random Forest Models

Lisa Barros de Andrade E Sousa (Helmholtz AI), Dominik Thalmeier (Helmholtz AI), Helena Pelin (Helmholtz AI) and Marie Piraud (Helmholtz AI).

Abstract:

Complex supervised machine learning methods, like Random Forest (RF) models, are often considered to be 'Black Boxes'. Such models can make highly accurate predictions but their complexity hinders us to understand the decision-making process for certain predictions. To deploy such models to the real world, it is indispensable to not only make accurate predictions but also to understand the logic behind those predictions. Only by understanding the model's decision-making process, we can ensure that the model produces valuable insights. Standard explainability methods like permutation feature importance are commonly used to pinpoint the individual contribution of features to the model performance. However, such methods assume feature independence and hence, might miss the role of correlated features in the model's decision-making process. In addition, the provided output makes it almost impossible to uncover feature interactions, an important aspect considering the non-linear nature of RF models. We addressed those problems by developing the Forest-Guided Clustering algorithm, which computes feature importance based on subgroups of instances that follow similar decision paths within the RF model, thus focusing on pattern-driven rather than performance-driven importance. The importance of each feature can be analyzed on a global scale, giving an overview on features that drive the underlaying patterns in the data, but also on a local scale, measuring the relevance of each feature for a specific subgroup. The pattern-driven importance metric of our method avoids the misleading interpretation of correlated features, allows the detection of feature interactions and gives a sense for the generalizability of identified patterns.

Data

From climate defined ecological niches to microbiome diversity and intra-community synergies

Dagmara Błaszczyk (Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland), Witold Wydmański (Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland), Krzysztof Mnich (Computational Centre, University of Białystok, Białystok, Poland), Valentyn Bezshapkin (Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland), Michał B. Kowalski (Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland), Alina Frolova (Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland), Witold Rudnicki (Computational Centre, University of Białystok, Białystok, Poland) and Paweł P. Łabaj (Małopolska Centre of Biotechnology of Jagiellonian University).

Abstract:

Microbiota research is becoming increasingly focused on the exposome factors and relation to metadata. It allows finding microorganisms that are specific to ecological niches. However, most of the research examines microorganisms found in a sample as separate features, which does not give the whole picture of synergies between microorganisms and their influences on processes carried out in ecological niches. Our project aims to discover the synergies between microorganisms and examine whether those dependencies influence the classification of samples in one of the established Polish microclimate clusters.

In our study, we use 240 soil samples collected from different locations in Poland which have been sequenced with extreme depth of over 100M paired-end reads (Whole Metagenome Sequencing). The sampling locations have been selected based on climate characteristics supported by over 10 years of weather conditions parameters history and represent three different Polish microclimate clusters.

We use MDFS (MultiDimensional Feature Selection) (Mnich & Rudnicki, 2020), which is based on Mutual-information theory, to reveal synergies between microorganisms in corresponding climate niches. This further allows us to investigate how exploiting microbial synergies impacts the classification of samples into specific climate clusters.

The first results indeed confirm the existence of microclimate-related and local-specific microbial communities, which is in line with earlier studies of MetaSUB Consortium (Danko et al., 2021) on a global scale. Just with Poland being very homogeneous from a climate and biome perspective, we investigate whether microbial synergies might be the key to studying the diversity of microorganisms between closely related ecological niches.

Data

Gene co-occurrence analysis for prediction of unknown gene function

Bartosz Baranowski (Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland) and Krzysztof Pawłowski (University of Texas Southwestern Medical Center, Dallas, TX, USA).

Abstract:

Despite improvement of wet-lab techniques and bioinformatics tools, still many genes remain functionally uncharacterized. Knowledge of biochemical and signaling pathways that involve uncharacterized genes in microorganisms could be key to understanding microbial biology and mechanisms of infectious diseases. It can be expected that genes co-occurring across different genomes will share similar biological functions more likely than random pairs of genes. This feature has been used in a few bioinformatic tools for functional relationship prediction eg. STRING COGNAT or G-NEST.

In our approach, we search for significantly co-occurring proteins using four measures: the Fisher's test, Jaccard index, mutual information and scalar product.

Having in mind uneven sampling of the microbial world by genome sequencing projects (e.g. hundreds of different Escherichia coli strains are available publicly) we propose a novel algorithm that allows collapsing the co-occurrence relationships at different taxonomic levels, strains or species. Current version of the tool provides three datasets: the proteomes of Legionella pneumophila, Escherichia coli and Homo sapiens. In addition, a rigorous statistical assessment with provision for multiple testing is also available.

The use of the proposed algorithm may find application in co-occurrence analysis of uncharacterized effector proteins from L. pneumophila to understand their role in infections. In case of E. coli and H. sapiens proteome databases, co-occurrence analysis may be useful to provide data to help understand the function of unknown genes.

Data

GenOptics: An intuitive platform of visual analytics for integrative analysis of large-scale multi-omics data

Konstantinos A. Kyritsis (Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece), Nikolaos Pechlivanis (Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece), Angeliki Magklara (Laboratory of Clinical Chemistry, Faculty of Medicine, University of Ioannina, Ioannina Greece), Anastasia Kougioumtzi (Laboratory of Clinical Chemistry, Faculty of Medicine, University of Ioannina, Ioannina Greece), Panagioti

Abstract:

During the past two decades, computational analysis has become paramount for biological research. Advancements in high-throughput methods and computational tools resulted in the generation of large amounts of data from different omics fields (multi-omics), such as genomics, epigenomics, transcriptomics, and metabolomics. This plethora of large scale and diverse omics data is being driven by the understanding that a single-omic type does not provide adequate information and integrative analysis of multi-omics data is optimal to gain sufficiently meaningful insights into the actual biological mechanisms. Although various open-source tools have been developed for this purpose, multi-omics data integration and analysis are still beset by a number of problems, including software compatibility, complex parameter selection and creation of functional pipelines with multiple steps of analyses. In this work we present GenOptics, a novel visual analytics platform that aims to facilitate the integration and subsequent analysis of diverse multi-omics datasets as well as meta-data (e.g., clinical data), through a fully interactive environment. The platform comprises of two separate parts. The first incorporates asynchronous analyses of Next-Generation Sequencing raw data, including RNA-, Whole exome-, and ChIP-Seq, using workflows implemented with the Common Workflow Language (doi:10.6084/m9.figshare.3115156.v2) and Docker containers to automate software installation and confer cross-platform portability. The second part constitutes the analytical platform itself, designed to facilitate the execution of robust bioinformatics analyses by life scientists with minimal or no knowledge of programming. GenOptics constitutes an open source (https://genoptics.github.io/), computational biology platform for novel pattern and biomarker detection.

Data

GWAS Central: a resource for the discovery and comparison of summary-level genome-wide association study data

Tim Beck (University of Leicester), Tom Shorter (University of Leicester), Thomas Rowlands (University of Leicester) and Anthony Brookes (University of Leicester).

Abstract:

The GWAS Central (https://www.gwascentral.org/) resource [1] is an ELIXIR-UK Node Data Service that provides integrative access and visualisation of a uniquely extensive collection of genome-wide association study (GWAS) data, while ensuring safe open access to prevent research participant identification. GWAS Central is the world's most comprehensive openly accessible repository of summary-level GWAS association information, providing over 72.5 million P-values for 5000 studies investigating over 1700 unique phenotypes. GWAS data sets are received as direct submissions from authors and consortia in addition to being gathered from various public sources. Phenotype descriptions are standardised with Medical Subject Headings and the Human Phenotype Ontology to enable the precise identification of genomic variants associated with diseases, phenotypes and traits of interest.

GWAS can be used to support the clinical relevance of mouse genetic studies, so we have extended our human data integration techniques to incorporate model organism genotypic and phenotypic data. Additionally, natural language processing tools have been developed to extract GWAS data and metadata from the biomedical literature to enable scalable curation workflows for importing studies. GWAS Central data are discoverable from the perspective of genetic markers, genes, genome regions or phenotypes, via graphical visualisations, detailed downloadable data reports, a BioMart interface and a Global Alliance for Genomics and Health (GA4GH) Beacon API endpoint.

References

1. Beck, T, et al. (2020) GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. Nucleic acids research. 48(D1):D933–D940. https://doi.org/10.1093/nar/gkz895

Data

IDEAS: Integrative and Differential Expression and Alternative Splicing Analysis

Leonie Pohl (LMU Munich), Armin Hadziahmetovic (LMU Munich), Alexandra Schubö (LMU Munich) and Ralf Zimmer (LMU Munich).

Abstract:

High-throughput data sets with multiple conditions are difficult to interpret: Modern multi-dimensional genome-wide data sets allow for a multitude of pairwise differential analyses in order to detect and understand relevant differences for the respective research question. E.g., a setup including measurements of responses to virus infection of several viruses, over several cell-lines, over several time points post infection, potentially even including various sequencing techniques, all allow for the identification of differences of those viruses from a vast number of differential comparisons. Thus, it is often not clear how to combine the various evidences to answer research questions of interest. Any straight-forward method of merging and/or intersecting result lists does not even come close to a truly multi-dimensional analysis of the data set at hand. In addition, the integration of different data sources is a challenge which cannot easily be dealt with by standard tools. Various conditions need to be compared to detect relevant differences in order to understand biological regulation and network mechanisms.

The IDEAS approach provides an integrated differential analysis of both gene expression and alternative splicing across conditions and helps us combine them over specific contexts. Interesting candidates for viral differences may be investigated within the evolutionary context to quantify and visualize differential transcript regulation accounting for the identified difference. Moreover, many additional aspects can be considered: such as viral miRNAs and their binding sites in particular, are analyzed with IDEAS to provide a possible mechanistic explanation of the observed differences in transcript usage.

Data

Identifying Viral miRNAs with Text Mining at the example of SARS-CoV-2 (VIM-TM)

Markus Joppich (LMU Munich), Armin Hadziahmetovic (LMU Munich), Alexandra Schubö (LMU Munich) and Ralf Zimmer (LMU Munich).

Abstract:

miRNAs are known to modulate complex human diseases, like atherosclerosis or ALS. Viral miRNAs also play a key role in the interplay between virus and host, e.g., miRNAs encoded in Epstein-Barr- or Herpes-viruses and, recently, SARS-CoV-2.

Viral miRNAs and their interaction with the host are described in scientific literature, e.g. in expert-compiled reviews, but systematic and comprehensive collections are missing. While databases of viral miRNAs exist e.g. VIRmiRNA or ViralmiR, these are limited to some 44 viruses, outdated for most viruses, or simply not available for new viruses such as SARS-CoV-2.

We present a framework for finding and extracting literature-described viral miRNAs and their host interactions. We benchmark our method on literature describing SARS-CoV-2-related viral miRNAs, for which a manually curated gold standard was prepared based on the comprehensive LitCovid collection. Checking the viral genome for the found miRNA sequences ensures to only report bona-fide miRNA candidates. Finally, we provide an overview of viral miRNA literature and the sequences published (validated or predicted) therein as an easily accessible, comprehensive resource.

Relevant abstracts and full texts are found by using co-occurrences of viral species and miRNA-related terms. For such identified texts, supplemental data is automatically downloaded. The texts, and its supplemental data, are then automatically searched for miRNA sequences, which are checked by aligning them to their respective genome.

In this poster we describe our viral miRNA detection and extraction strategy, and discuss advantages and disadvantages of the described workflow as well as its performance as derived from the LitCovid benchmark.

Data

Imaging Data: Adding a new type of annotations to 3DBionotes-WS

Carolina Simón Guerrero (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)), Jose Ramon Macias (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)), Erney Ramírez-Aportela (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)), Jose Luis Vilas Prieto (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)), Marta Martinez Gonzalez (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)), Carlos Oscar Sanchez Sorzano (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)) and Jose Maria Carazo (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)).

Abstract:

3DBionotes-WS is a web application designed to automatically annotate biochemical and biomedical information onto structural models in a fully interactive 3D graphical environment. Current sources of information include domain families, genomic variations associated with diseases, immune epitope sites, short linear motifs, disordered regions and post-translational modifications that can be explored interactively at the sequence and structural level. 3DBionotes-WS is always in continuous expansion and looking for new data types to integrate.

The imaging data field has experienced very fast growth in content and complexity. The size of biological image datasets is quite large, which makes it hard to submit, manage and publish data. Metadata defining imaging protocols, biological systems or the processing outputs sometimes are not properly included in publications, making it difficult to analyse datasets. For that reason, in the last few years, it has become crucial to create an infrastructure for image data standardisation, deposition, sharing and, in the end, better data availability.

A few initiatives have arisen to fulfil this need, being the Image Data Resource (IDR) & the BioImage Archive (BIA) some of the more relevant. Among all the useful data provided by these repositories, we have focused on the results from High-Content Screening experiments for SARS-CoV-2 and a well-defined collection of compounds with potential clinical use. 3DBionotes-WS users will be able to search for COVID-19-related macromolecular structures bound to a ligand already tested with the corresponding link to the experimental data source for further reference.

Data

Impact of preprocessing in data integration of single-cell RNA-seq data

Youngjun Park (UNIVERSITÄTSMEDIZIN GÖTTINGEN) and Anne-Christin Hauschild (UNIVERSITÄTSMEDIZIN GÖTTINGEN).

Abstract:

Recent advances in single-cell RNA (scRNA) sequencing have opened the possibility to study tissues down to the level of cellular populations. Subsequently, this enabled various scRNA studies that reported novel or previously undetected subpopulations and their functions. However, the heterogeneity in single-cell sequencing data makes it unfeasible to adequately integrate multiple datasets generated from different studies. This heterogeneity originates from various sources of noise due to technological limitations. Thus, particular procedures are required to adjust such effects prior to further integrative analysis. Subsequently, over the last years, numerous single-cell data analysis tools have been introduced de-noising methods implementing various data transformation methods. Here, we investigated 22 of the most recent single-cell studies and found that many analyses procedures employed various data transformation and preprocessing steps without further reasoning. This fact is particularly alarming since these read-count transformations can alter data distribution and affect downstream cell clustering results. This study aims to investigate the effects of the various data transformation on three different public data scenarios and evaluate these using popular dimensionality reduction and clustering analysis. Furthermore, we discuss implication on the use of transfer learning for batch correction and de-noising. In summary, our benchmark work shows that a large portion of batch-effects and noise can be mitigated by simple but well chosen data transformations and suggest that such analysis should be the baseline for all studies and a proper comparison between batch effect correction methods.

Data

IMPaCT-Data Biomedical Cloud: An initial iteration for a federated virtual computing environment in the context of Precision Medicine in Spain.

María Chavero-Díez (Barcelona Supercomputing center, Spanish National Bioinformatics Institute), Jose María Fernández (Barcelona Supercomputing center, Spanish National Bioinformatics Institute), Laia Codó (Barcelona Supercomputing center, Spanish National Bioinformatics Institute), Lidia Lopez (Barcelona Supercomputing center, Spanish National Bioinformatics Institute), Salvador Capella-Gutierrez (Barcelona Supercomputing center, Spanish National Bioinformatics Institute), Josep Lluis Ge

Abstract:

The Spanish Precision Medicine Infrastructure associated with Science and Technology (IMPaCT), aims to lay the foundations for impulsing precision medicine within the Spanish National Health System. IMPaCT revolves around three main pillars: predictive medicine, data science and genomic medicine.

As part of the Data Science program, the IMPaCT-Data Biomedical Cloud is being established for providing a scalable and flexible analysis environment, enabling the integration, management and analysis of structured clinical, richly described genomic and annotated medical imaging data available within IMPaCT.

The infrastructure is designed as a federated cloud system resultant from the interconnection of an increasing number of research organizations functioning under coordinated policies. The infrastructure components include a Life-Sciences-ID-compliant federated authentication system, a data infrastructure to manage public datasets and prospectively sensitive data (leveraging EGA technologies), and a distributed set of computational resources offering platforms like Galaxy, among others. IMPaCT-Data Biomedical Cloud will benefit from services like bio.tools, BioContainers, WorkflowHub and OpenEBench for proper software management and evaluation.

IMPaCT-Data Biomedical Cloud main objective is to provide a platform as a service that enables the management of resources, and exchange of controlled datasets in their lawful limits, delivering services for data applications and analysis for research and clinical personnel with any kind of experience with information technologies.

Data

Implementing best practices for setting a bioinformatics core facility

Pau Marc Muñoz Torres (Vall d'Hebron Institute of Oncology), Merce Alemany-Chavarria (Vall d'Hebron Institute of Oncology) and Lara Nonell (Vall d'Hebron Institute of Oncology).

Abstract:

The recently created bioinformatics unit at XXXX has followed the implementation model learnt during the first EMBO Practical Course edition: Research to service: Planning and running a bioinformatics core facility. The mission of the facility is to support VHIO research groups from the experimental design to data analysis and final publication, providing research groups with state-of-the-art computational resources for the analysis of cancer-related omics data.

Integrated by three members and some temporary master students, we define a service as a collaboration with the internal researchers, free of charge but with a clear workflow. The service starts by meeting with the researcher to identify bioinformatics needs, define clear objectives and set the timing. We deliver our results together with a report and refine the analysis whenever needed. The service ends by uploading the data and code to repositories, following FAIR principles.

Our computational procedures are based on open-source software and are at the same time developed in a safe and reproducible environment. We have settled a scalable computational cluster infrastructure, with Slurm, Docker and Nextflow, to fulfill our computational needs as well as those of the bioinformaticians integrated in the groups. Besides, we have defined internal programming standards, store our code in GitHub and use mainly nf-core pipelines. In addition, we mentor and coordinate an internal bioinformatics network with the objective of sharing knowledge and optimizing resources. To that end, we mentor several bioinformaticians and organize a monthly Journal club to discuss interesting articles related to computational biology.

Data

Indication expansion via geometric scattering-based knowledge graph embedding

Dhananjay Bhaskar (Yale University), Sergio Picart-Armada (Boehringer Ingelheim Pharma GmbH & Co. KG) and Smita Krishnaswamy (Yale University).

Abstract:

The success rate of first-in-class treatments in clinical trials is low, with many promising candidates failing to clear efficacy and safety requirements. Alternatively, target-centric repurposing of safe compounds can enable fast-track approval of novel treatments for indications with unmet clinical need. This approach can be formulated as a link prediction task on a knowledge graph (KG) that encodes relationships between genes, tissues, and indications. Previous studies have shown that a KG combining both structured biological data and literature sources is better suited for predicting links via node embedding-based methods [1].

Here we employ learnable geometric scattering [2] to embed the KG. Node embeddings obtained by encoding the scattering transform of the KG through an autoencoder achieve state-of-the-art performance on target-indication link prediction. This technique uses lazy random walks to learn wavelet filters for multiscale graph representation. We use the manifold geometry preserving method, PHATE [3], to visualize the embeddings and gain further insight into the model performance. Finally, we investigate the sensitivity of the link predictions to the node embeddings and the input graph signals, to increase confidence and interoperability of our results.

This work was funded by a Yale – Boehringer Ingelheim Biomedical Data Science Fellowship.

[1] Gurbuz et al., “Knowledge Graphs for Indication Expansion: An Explainable Target-Disease Prediction Method,” Front Genet. 2022; 13: 814093.

[2] Tong et al., “Data-Driven Learning of Geometric Scattering Modules for GNNs,” IEEE MLSP 2021; pp. 1-6.

[3] Moon et al., “Visualizing structure and transitions in high-dimensional biological data,” Nat. Biotechnol 2019; 37, pp. 1482-1492.

Data

InSoLiTo: The research software graph-based network from OpenEBench

Sergi Aguiló-Castillo (Spanish National Bioinformatics Institute (INB/ELIXIR-ES), Barcelona Supercomputing Center (BSC)), José M. Fernández (Spanish National Bioinformatics Institute (INB/ELIXIR-ES), Barcelona Supercomputing Center (BSC)), Josep Ll. Gelpí (Spanish National Bioinformatics Institute (INB/ELIXIR-ES), Barcelona Supercomputing Center (BSC)) and Salvador Capella-Gutierrez (Spanish National Bioinformatics Institute (INB/ELIXIR-ES), Barcelona Supercomputing Center (BSC)).

Abstract:

InSoLiTo is a graph-based network of the co-usage of research software from the tools collection hosted by OpenEBench. Co-usage is understood as being cited in the same scientific publications.

The initial aim of the project is to identify how bioinformatics tools relate to each other allowing to potentially infer analytical workflows commonly used in the literature. These results can be then used by research communities organizing scientific benchmarks on popular pipelines in their domains. Interestingly, this analysis can also be extended to include widely used databases and other relevant resources as part of those workflows.

Data is freely accessible from a data portal. In the webpage, the user can interact with the graph network by filtering any bioinformatics' tool stored in OpenEBench and see which other tools works with. Also, users can filter per year to see the evolution of the usage of other software in relation to the selected tool, and see which bioinformatics' community it belongs to. Moreover, the inter- and intra-relationships between different branches of bioinformatics regarding its tools co-usage can be examined.

This addition of OpenEBench will bring a new perspective of benchmarking and monitoring research software that will complement the existing capabilities of the platform.

Data

Integration of multi-omics data using graph neural networks to identify and contextualize biomarker genes for psychiatric disorders

Svitlana Oleshko (Helmholtz Center Munich / Boehringer Ingelheim Pharma GmbH & Co. KG), Francesc Fernández-Albert (Boehringer Ingelheim Pharma GmbH & Co. KG), Matthias Heinig (Helmholtz Center Munich), Sergio Picart-Armada (Boehringer Ingelheim Pharma GmbH & Co. KG) and Annalisa Marsico (Helmholtz Center Munich).

Abstract:

Mental health today is a burden on global level with various psychiatric disorders leading to a lower quality of life, a significant number of deaths and a higher pressure on healthcare systems. As developing more effective treatments for the most common mental disorders is of high priority for international public health, substantial research is nowadays dedicated to this issue. This includes studies where emphasis is put on biological and mechanistic understanding of the brain. For example, the BrainSeq initiative provides various omics data generated from brain tissue and already presents some insights into the molecular mechanisms of gene regulation for psychiatric disorders, in particular - for schizophrenia. In order to achieve a higher level of interpretability, we aim to develop a method based on a graph neural network which leverages prior knowledge of gene-gene associations in the form of protein-protein interaction or co-expression network and omics data generated from brain tissue. We apply graph attention network model to integrate transcriptomic and DNA methylation data with gene-gene associations to identify candidate genes related to schizophrenia. We assess the value of including both tissue-agnostic and brain-specific gene-gene associations as an additional data source and the advantages of graph neural networks as opposed to other deep learning architectures that have been used on this data so far. Together with the interpretation framework we have developed for model explanation, the model allows for patient-level disease state prediction on top of gene-level disease state prediction which can be framed in the context of precision psychiatry.

Data

Integrative network analysis interweaves the missing links in cardiomyopathy diseasome

Pankaj Chauhan (NCBS-TIFR) and Ramanathan Sowdhamini (NCBS-TIFR).

Abstract:

Cardiomyopathies are genetic diseases showing an abnormal heart phenotype that contribute to approximately 36% of heart failures in the patients. Cardiomyopathies are known to share overlapping genetic and phenotypic features with other diseases. Several studies have pointed out that several drugs, including anti-cancer, antiretroviral and antipsychotic, pose the risk of cardiotoxicity and drug-induced cardiomyopathies. Genome-wide association (GWAS) and candidate gene analysis approaches have reported many common genetic variants that might be one of the plausible reasons for toxicity or induced cardiomyopathy. Hence, a fundamental knowledge of the cardiomyopathies and the molecular players involved is critical for developing novel approaches for its prevention and treatment. The emergence of network science has aided in understanding complex systems like protein-protein interactions and disease-disease associations. In this pursuit, we constructed the cardiomyopathy diseasome network to systematically inquire about the genetic interplay of the cardiomyopathy with other diseases and uncover the molecular players underlying these associations. In addition, we predicted a range of modifier genes for cardiomyopathies through the integration of protein-protein interactions datasets. Through integrative systems analysis of heart-specific mouse knockout data and disease tissue-specific transcriptomic data, strengthen predicted prominent modifiers.

Data

Introducing the National Research Data Infrastructure for the Research of Microbiota (NFDI4Microbiota)

Adrian Fritz (Helmholtz-Centre for Infection Research), Anke Becker (Philipps-Universität Marburg), Peer Bork (European Molecular Biology Laboratory), Thomas Clavel (RWTH Aachen University), Ulisses Nunes da Rocha (Helmholtz Centre for Environmental Research), Konrad U. Förstner (ZB MED - Information Centre for Life Sciences), Alexander Goesmann (Justus-Liebig-University Giessen), Barbara Götz (ZB MED - Information Centre for Life Sciences), Manja Marz (Friedrich Schiller University Jena), Jörg Overmann (German Collection of Microorganisms and Cell Cultures), Carmen Paulmann (Helmholtz Centre for Infection Research), Kristin Sauerland (Helmholtz Centre for Infection Research), Alexander Sczyrba (Bielefeld University), Jens Stoye (Bielefeld University) and Alice C. McHardy (Helmholtz Centre for Infection Research).

Abstract:

Microbes play an important role in human and environmental health. As well as helping to address global health threats such as antimicrobial resistance and viral pandemics, microbial research has a key role to play in areas such as agriculture, waste management, water treatment, ecosystems remediation, and the diagnosis, treatment and prevention of various diseases. Driven by technological advances that allow e.g. high-throughput molecular characterization of microbial species and communities, microbiological research helps to address these global health threats. The recent advances result in the generation of large data sets, yet the use and re-use of this data so far has failed to exploit its potential. NFDI4Microbiota has started its activity in October 2021 together with several other NFDI (National Research Data Infrastructure) consortia in Germany. It consists of ten well-established partner institutions and is supported by five professional societies and more than 50 participants. It aims to facilitate the digital transformation of the microbiological community and thus the mission of the consortium is to support the microbiology community with access to data, analysis services and (meta)data standards. Besides this, training as well as community engaging activities are offered. Furthermore, a cloud-based system will be created that will make the storage, integration and analysis of microbial data - especially omics data - consistent, reproducible, and accessible. Thereby, NFDI4Microbiota will promote the FAIR (Findable, Accessible, Interoperable and Re-usable) principles and Open Science. Through the dual emphasis on education and services, NFDI4Microbiota will ensure that microbial research in Germany is synergistic and efficient.

Data

It's FLAN time! Summing feature-wise latent representations for interpretability

An-Phi Nguyen (ETH Zurich, IBM Research Zurich), Stefania Vasilaki (IBM Research Europe (Zurich Lab)), Maria Rodriguez Martinez (IBM Research Europe (Zurich Lab)) and Alice Driessen (IBM Research).

Abstract:

Interpretability has become a necessary feature for machine learning models deployed in critical scenarios, e.g. legal system, healthcare. In these situations, algorithmic decisions may have (potentially negative) long-lasting effects on the end-user affected by the decision.

While deep learning models achieve impressive results, they often function as a black-box.

Inspired by linear models, we propose a novel class of structurally-constrained deep neural networks, which we call FLAN (Feature-wise Latent Additive Networks). Crucially, FLANs process each input feature separately, computing for each of them a representation in a common latent space. These feature-wise latent representations are then simply summed, and the aggregated representation is used for the prediction. These feature-wise representations allow a user to estimate the effect of each individual feature independently from the others, similarly to the way linear models are interpreted.

We demonstrate FLAN on a series of benchmark datasets in different biological domains. Our experiments show that FLAN achieves good performances even in complex datasets (e.g. TCR-epitope binding prediction), despite the structural constraint we imposed. On the other hand, this constraint enables us to interpret FLAN by deciphering its decision process, as well as obtaining biological insights (e.g. by identifying the marker genes of different cell populations). In supplementary experiments, we show similar performances also on non-biological datasets.

Data

Latent Variable Random Forest for Enhanced Feature Importance

Melpomeni Kasapi (Imperial College London), Kexin Xu (Imperial College London), James S. Ware (MRC London Institute of Medical Sciences, Imperial College London), Declan P. O'Regan (MRC London Institute of Medical Sciences, Imperial College London), Timothy M.D. Ebbels (Imperial College London) and Joram M. Posma (Imperial College London).

Abstract:

Random Forest (RF) classifiers are often used in biological data contexts due to their interpretability and dimensionality scaling. They are algorithms that can deal with a large number of variables, achieve reasonable prediction scores, and yield highly interpretable feature importance values. As such, they are appropriate models for feature selection and further dimensionality reduction (DR) in integrated datasets. A few techniques have been proposed for enhancing feature importance calculations in datasets with highly correlated variables, i.e. permutation feature importance. Addressing correlation relationships in high dimensional datasets, is imperative for reducing the number of variables that are assigned high importance, and hence making the DR most efficient. Here, we propose a novel method that derives latent variables based on distance characteristics of each feature and aims to incorporate the correlation factor in the splitting step. A user-specified distance matrix is calculated to select the k closest points to the variable of interest (VOI), which form the VOI neighborhood. The first left singular vector of the VOI neighborhood is calculated for each bagging-selected VOI and is used to determine the best split using Gini Impurity. The proposed Latent-Variable RF (LVRF) yields non-inferior prediction accuracies to traditional RFs (HNMR example dataset mean accuracy ± SD: 0.624±0.035 LVRF, 0.630±0.035 RF), while enhancing the feature importance interpretability. Unlike traditional RFs, LVRF is unaffected by single 'important' noisy features (false positives), as it considers the local neighborhood. LVRF therefore highlights neighborhoods of features-- reflecting real signals-- that collectively impact the predictive ability of the model.

Data

LCRBert: word embedding to support detection of articles containing low-complexity regions

Sylwia Szymańska (Department of Computer Networks and Systems, Silesian University of Technology) and Aleksandra Gruca (Department of Computer Networks and Systems, Silesian University of Technology).

Abstract:

Low complexity regions (LCRs) are characterized by low amino acid diversity in protein sequences. Recent studies suggest that LCRs may play important functional roles in cells, among which we can highlight DNA/RNA or metal binding. They may also take part in regulation of cellular processes. The increased awareness of the importance of LCRs resulted in a growing number of scientific publications investigating their functional features, however our knowledge on LCR functions is still limited and mostly unstructured.

We compare different embedding creation methods based on BERT-like transformer language models such as token/sentence/document level and dimensionality reduction. Created embeddings are then used as input to the state-of-the-art classification machine learning algorithms that are trained to recognize papers describing functions of specific LCRs. To create embeddings and to train the classifiers we use a manually annotated database of LCR-positive and LCR-negative set of titles and abstract representing articles describing functions of LCRs enriched with leucine, proline and glycine. The performance of the classifiers trained using embeddings based on BERT-like transformer language models is then compared to the performance of the models trained using classical NLP feature extraction techniques based-on the bag-of-words approach.

Data

Light-weight alignment enables fast and accurate metagenome and metatranscriptome quantification

Shuba Varshini Alampalli (Helmholtz-Institut für RNA-basierte Infektionsforschung (HIRI), 97080 Würzburg, Germany), Eva Weiss (Institut für molekulare Infektionsbiologie (IMIB)) and Lars Barquist (Helmholtz-Institut für RNA-basierte Infektionsforschung (HIRI); Faculty of Medicine, University of Würzburg).

Abstract:

Quantification of metagenomics and metatranscriptomics data from complex samples is challenging and is affected by both the choice of quantification method and reference sequence databases. In addition, many quantification approaches are extremely resource intensive, particularly in light of continuously expanding reference databases. Here, we investigate the applicability of computationally-efficient light-weight aligners to the meta'omics. We show that these tools perform similarly to state-of-the-art metagenomic classifiers, with reduced resource requirements. Further, we show that light-weight aligners can be used with marker gene databases as reliable filter for full metagenome quantification. Finally, we illustrate the advantages of light-weight alignment against pangenome databases for the quantification of metatranscriptomics using a realistic simulation study, and show that database composition is critical to maximizing read recovery in a metatranscriptomic study of murine infection with Salmonella Typhimurium. Based on these findings, we have developed a flexible pipeline for meta'omics quantification using light-weight alignment in Nextflow.

Data

Low Complexity Regions in kinases

Joanna Ziemska-Legięcka (Institute of Biochemistry and Biophysics, Polish Academy of Sciences), Aleksandra Gruca (Silesian University of Technology) and Marcin Grynberg (Institute of Biochemistry and Biophysics, Polish Academy of Sciences).

Abstract:

Low Complexity Regions (LCRs) in proteins are fragments with high accumulation of similar amino acids. These regions can be composed from one to few types of amino acid. LCRs with repeats of one amino acid are called homopolymers. LCRs with more types of amino acids can contain repeats of short patterns or can lack regularity. LCRs can be responsible for important functions in proteins. For example, the subunit alpha of casein kinase II contains a K-rich region which improves functions of that protein by stabilizing the structure of the protein.The other known function of LCR is protein STE20/SPS1-related proline-alanine-rich kinase the low complexity region called PAPA box, which simplifies binding to actin. Functions of LCRs are poorly annotated and not many of them are known.

LCRs are more common in kinases than in all sequences from the SwissProt database. LCRs occur in at least 71% of protein kinases. 58 % of SwissProt proteins contain kinase domains. Therefore, in order to find functions of LCRs in kinases we analyze them. We collected annotations of LCRs of kinase domains from LCRAnnotationsDB, a database of LCR annotations. 15% of LCRs are located in the ATP binding site and 3% are present in protein binding sites. In human kinase domains, almost 70% of LCRs are located between 5 and 6 beta-strands and D and E helices. According to IUpred3 results, 53% of LCRs in kinases are in disordered regions.

Data

Machine learning for extraction of biochemical reactions from the scientific literature

Blanca Cabrera Gil (SIB Swiss Institute of Bioinformatics), Anne Morgat (SIB Swiss Institute of Bioinformatics), Venkatesh Muthukrishnan (SIB Swiss Institute of Bioinformatics), Elisabeth Coudert (SIB Swiss Institute of Bioinformatics), Kristian Axelsen (SIB Swiss Institute of Bioinformatics), Nicole Redaschi (SIB Swiss Institute of Bioinformatics) and Alan Bridge (SIB Swiss Institute of Bioinformatics).

Abstract:

Rhea (www.rhea-db.org) is an expert curated knowledgebase of biochemical reactions built on the chemical ontology ChEBI (www.ebi.ac.uk/chebi), the reference vocabulary for enzyme and transporter annotation in UniProtKB (www.uniprot.org) and an ELIXIR Core Data Resource. Rhea currently describes over 14,000 unique reactions and provides annotations for over 23 million proteins in UniProtKB in forms that are FAIR – but most knowledge of enzymes remains locked in literature and is inaccessible to researchers. Machine learning methods provide a powerful tool to address this problem. Here we present work designed to accelerate the expert curation of Rhea by extracting putative enzymatic reactions automatically from the literature using language models trained on large organic chemistry and enzymatic reaction datasets from Rhea and MetaNetX (www.metanetx.org).

Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, CMU, 1 Michel-Servet, 1211 Geneva 4, Switzerland

Data

Machine learning identifies weak prognostic signal in colorectal polyp's transcriptome

Russell Hung (Canon Medical Research Europe), Simon Fisher (Canon Medical Research Europe), Ditte Andersen (BioClavis), Gerard Lynch (University of Glasgow), Noori Maka (NHS Greater Glasgow and Clyde), Jennifer Hay (University of Glasgow), Jakub Jawny (University of Glasgow), William Sloan (NHS Greater Glasgow and Clyde), Stephen McSorley (NHS Greater Glasgow and Clyde), Joanne Edwards (University of Glasgow) and Ian Poole (Canon Medical Research Europe).

Abstract:

Metachronous polyps refer to the occurrence of secondary polyps at a different location after the removal of primary ones during index colonoscopies. The Integrated Technologies for Improved Polyp Surveillance (INCISE) project aims to develop a comprehensive metachronous risks prediction tool to improve colorectal cancer screening program.

As part of the project, this study is concerned with identifying metachronous polyp risk factors through the analysis of the INCISE cohort index polyp transcriptome (n=1789). Potential genes of interest associated with colorectal cancers were identified from a GWAS Catalogue, PGS Catalogue and Protein Atlas databases. The dataset was split into training and holdout validation set. The training set was used for model development and parameter tuning. The holdout validation set was used solely for evaluation.

We use binary classification (logistic regression, support vector machine and random forest) and time-to-event (autoencoder survival model) approaches to model the risks of future polyp occurrence with gene expression data. The usefulness of incorporating time-to-event was evaluated.

Model performance in the holdout validation set indicates that gene expression did not show significant association with metachronous risk, regardless of the approach taken (ROC-AUC: binary – 0.52, time-to-event – 0.54). Overall, a signal has not yet been identified in the index polyp transcriptome. However, further analysis with other methodologies is underway.

Data

Microbial co-occurrence network reveals climate and geographic patterns for soil diversity on the planet

Nikolaos Pechlivanis (Institute of Applied Biosciences, Centre for Research and Technology, Hellas), George Karakatsoulis (Institute of Applied Biosciences, Centre for Research and Technology, Hellas), Stefanos Sgardelis (Department of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki), Ilias Kappas (Department of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki) and Fotis Psomopoulos (Institute of Applied Biosciences, Centre for Research and Technology, Hellas).

Abstract:

Soil microbiota play an integral role in the shaping of the overall biodiversity of our planet. Yet systematic investigations on how microbial communities are affected by geographic and climatic factors, especially with the ongoing climate change, are still limited. Previous studies (10.1038/ismej.2015.261) have tried to explore the microbial diversity in specific geographic areas or were focused more on the underlying microbial interactions rather than the structure changes (10.1186/s40168-020-00857-2). Here, we explore the effects of key climatic factors and geographic patterns on the soil diversity and microbial interactions across the globe. To this end, we have used data from the Earth Microbiome Project (EMP, 10.1038/nature24621), as it offers a massive collection of planetary-scale microbiome datasets. In conjunction with the Köppen-Geiger climate classification (10.1038/sdata.2018.214) and other bioclimatic variables from the WorldClim database (10.1002/joc.5086), the above dataset assembly can be used to identify important variations in soil diversity. Initial comparisons between different climate classifications revealed statistically significant differences amongst them, highlighting the role that temperature and precipitation play in microbiota presence. In addition, a microbial co-occurrence network was built to capture soil microbial interactions using the SpiecEasi workflow (10.1371/journal.pcbi.1004226). The resulting network was used to identify hub species in each climatic environment and revealed significant topological shifts on network level for different climatic regions. Our study of soil diversity in different climate environments contributes to a better understanding of how climatic factors affect its distribution on Earth.

Data

Mitigating failure modes of molecular optimization using expert constraints and ensembles

Adam Arany (ESAT/STADIUS, University of Leuven), Jaak Simm (ESAT/STADIUS, University of Leuven), Natalia Dyubankova (Janssen Pharmaceutica NV.), Jonas Verhoeven (Janssen Pharmaceutica NV.), Martijn Oldenhof (ESAT/STADIUS, University of Leuven) and Yves Moreau (ESAT/STADIUS, University of Leuven).

Abstract:

There is a growing number of machine learning work addressing the problem of generating molecular structures with favorable properties. Usually the generated structures cannot be evaluated by laboratory experiments due to the prohibitive cost of de novo chemical synthesis. To avoid this constraint often a black box machine learning model is used as a quality function. It is observed that this setup results in the generation of structures with adversarial behavior [1]. In the present work we analyze the effect of a two-fold mitigation strategy. On the one hand, we make the quality model more robust using ensembles. On the other hand we restrict the generator to create only chemically meaningful candidates. To do so we use a Monte Carlo Tree Search[2] based planner, to generate candidates based on chemical reactions, commercially available building blocks and initial scaffolds validated by chemist experts. We used inhibitory activity on JAK2 kinase from ChEMBL[3] and toxicity data from Tox21[4] as training data for the quality function. We applied the strategy to evaluate robustness of the generator as in Renz et al.[1], and found that our method significantly improves the quality of generated molecules. Additionally, the method directly outlines a chemical synthesis plan.

[1] Renz, Philipp, et al., Drug Discovery Today: Technologies 32 (2019): 55-63.

[2] Segler, Marwin HS, Mike Preuss, and Mark P. Waller. Nature 555.7698 (2018): 604-610.

[3] Mendez, David, et al. Nucleic acids research 47.D1 (2019): D930-D940.

[4] Richard, Ann M., et al. 34.2 (2020): 189-216.

Data

MTLSurv: Predicting Breast Cancer Patients' Survival with Multimodal Deep Neural Networks and Modality-Specific Transfer Learning

Sören Richard Stahlschmidt (Systems Biology Research Center, University of Skövde), Göran Falkman (Skövde Artificial Intelligence Lab, University of Skövde), Benjamin Ulfenborg (Systems Biology Research Center, University of Skövde) and Jane Synnergren (Systems Biology Research Center, University of Skövde).

Abstract:

Cancer prognosis is of great importance to physicians and patients since it may inform treatment and life decisions. However, for such complex diseases, accurate predictions of a patient's survival are often challenging as it is determined by a variety of underlying molecular changes, the cancer's spread, and other factors. Analyzing multi-omics datasets with machine learning methods have been shown to hold the potential to predict survival accurately. Particularly multimodal deep neural networks (DNN) can model the complex, often nonlinear relationships between different molecular features and survival. The different architectures and the hierarchical representations learned by DNNs allow flexible fusion of different modalities. To avoid overfitting, these algorithms require large sample sizes of paired data, which are frequently not available. However, data sharing efforts in the biomedical research community have resulted in a vast collection of uni- and bimodal publicly available datasets. We therefore investigate whether pre-training subnetworks of intermediate and late fusion DNNs with such source datasets can improve the accuracy of survival prediction on multimodal target datasets. We term this approach modality-specific transfer learning for multimodal survival networks (MTLSurv). Here, we present initial results based on The Cancer Genome Atlas – BRCA (TCGA-BRCA) data as a multimodal target dataset and a variety of unimodal source datasets. By testing MTLSurv on this case, we investigate whether it is possible to utilize existing datasets to improve survival prediction by enabling the training of larger multimodal DNNs for cancer prognosis.

Data

Multi-omics lung cancer subtyping by machine learning informs biomarker-guided drug development

Sven-Eric Schelhorn (Merck Healthcare KGaA).

Abstract:

Lung cancer is among the most prevalent oncologic diseases worldwide and consists of multiple histologically and molecularly defined subtypes that differ in both prognoses and treatment responses to modern TKI, IO, and DDR treatments. Automatically inferring these subtypes based on genomics and imaging data provides significant advantages to drug development efforts by allowing to precision-target new therapies to well characterized, sensitive patient populations.

We developed an automatic machine learning pipeline to infer 15 lung cancer histological and molecular subtypes derived from literature for both non-small cell lung cancer (NSCLC) and small-cell lung cancer (SCLC) based on multi-omics training data from patient cohorts with extensive real-world data (n=2,176), cancer cell lines (n=162), and patient-derived xenograft (PDX) models (n=285). Besides comparing multiple machine learning approaches, dealing with class imbalance in a multi-class classification setting, and using nested resampling strategies for estimating the test error on unseen data we further validate our predictions on an hold-out set of PDX models annotated by both experts histologists and an orthogonal, deep-learning based histopathology AI method.

Our inferred, hierarchical consensus lung cancer subtypes on disease, histology, and molecular subtype levels provide state-of-the-art accuracy with per-patient prediction confidences, correspond to known molecular consensus subtypes derived from literature. In addition, they are correlated with actionable clinical covariates from real-world data such as overall and progression-free survival, oncogene status, and treatment responses to a wide range of lines of therapy. The model is currently brought into production for automatic subtype annotation within a large pharmaceutical organization.

Data

Namco: A microbiome explorer

Monica Matchado (Chair of Experimental Bioinformatics, Technical University of Munich), Alexander Dietrich (Chair of Experimental Bioinformatics, Technical University of Munich), Maximilian Zwiebel (Chair of Experimental Bioinformatics, Technical University of Munich), Benjamin Ölke (Chair of Experimental Bioinformatics, Technical University of Munich), Michael Lauber (Chair of Experimental Bioinformatics, Technical University of Munich), Ilias Lagkouvardos (ZIEL - Institute for Food & Health,Technical University of Munich), Beate Brandl (ZIEL - Institute for Food & Health,Technical University of Munich), Thomas Skurk (ZIEL - Institute for Food & Health,Technical University of Munich), Jan Baumbach (Institute for Computational Systems Biology, University of Hamburg), Hans Hauner (Institute of Nutritional Medicine, TUM School of Medicine, Technical University of Munich), Dirk Haller (ZIEL - Institute for Food & Health,Technical University of Munich), Sandra Reitmeier (ZIEL - Institute for Food & Health,Technical University of Munich) and Markus List (Chair of Experimental Bioinformatics, Technical University of Munich).

Abstract:

Background:16S rRNA gene profiling is currently the most widely used technique in microbiome research and allows for studying microbial diversity, taxonomic profiling, phylogenetics, functional and network analysis. While a plethora of tools have been developed for the analysis of 16S rRNA gene data, only few platforms offer a user-friendly interface and none comprehensively cover the whole analysis pipeline from raw data processing down to complex analysis. Results: We introduce Namco, an R shiny application that offers a streamlined interface and serves as a one-stop solution for microbiome analysis. We demonstrate Namco's capabilities by studying the association between rich fibre diet and the gut microbiota composition. Namco helped to prove the hypothesis that butyrate-producing bacteria are prompted by fibre-enriched intervention. Conclusion: Namco provides a broad range of features from raw data processing and basic statistics down to machine learning and network analysis, thus covering complex data analysis tasks that are not comprehensively covered elsewhere. Namco is freely available at https://exbio.wzw.tum.de/namco/.

Data

nf-core as the standard for BovReg reference pipelines

Jose Espinosa-Carrasco (Centre for Genomic Regulation), Björn E. Langer (Centre for Genomic Regulation), Philip A. Ewels (Seqera Labs), Harshil Patel (Seqera Labs), Peter Harrison (European Molecular Biology Laboratory-European Bioinformatics Institute) and Cedric Notredame (Centre for Genomic Regulation).

Abstract:

Research genomic consortia often produce large quantities of raw data that are processed into new annotations using a combination of computational methods known as pipelines. For a variety of reasons that mostly involve environment dependencies (third-party pipelines, operating system and computational setup), the pipelines are often difficult to re-use thus leading to irreproducible analyses. The BovReg project aims at both functionally annotating the cattle genome and producing a set of bioinformatics reference pipelines that will become a long-lasting resource for the community. To fulfill the latter goal, BovReg reference pipelines should adhere to current bioinformatics best practices. Hence, we adopted nf-core as our computational standard. nf-core is a collection of curated Nextflow pipelines defined around a precise standard for pipeline development (https://nf-co.re/). It relies on an active community that collects bioinformatics pipelines implemented with the Nextflow workflow manager. These pipelines follow best computational practices which guarantee reproducibility, portability, interoperability and a unified minimal functionality. Fortunately, some nf-core pipelines already cover certain BovReg analysis aspects and can be used directly. For others, the need for additional novel functionality was identified within BovReg and implemented as required. Finally, in the few cases where a pipeline doesn't currently exist on nf-core, BovReg partners are implementing novel, nf-core compliant pipelines that could eventually become part of the nf-core collection. Notably, this approach was also adopted by other EuroFAANG projects (https://eurofaang.eu/), such as GENE-SWitCH and AQUA-FAANG. As a result, a working group has been created that includes developers from the different consortia and nf-core members.

Data

Nightingale, visualizing biology on the web following standards

Aurélien Luciani (EMBL-EBI), Gustavo Salazar (EMBL-EBI), Daniel Rice (EMBL-EBI), Swaathi Kandasaamy (EMBL-EBI) and Maria Martin (EMBL-EBI).

Abstract:

Nightingale is a library of standard reusable visualization components for the web, aimed at representing data related to proteins with its main concept being that of tracks representing the sequence horizontally. Its composable architecture is meant to be able to combine as many tracks as needed vertically, representing protein sequence features like domains, variants, structures, interactions, binding or active sites, and facilitate their visual comparison to discover relationships across annotations at similar positions. It relies on standard web components which means it is not an isolated widget in the page but rather can be integrated with the rest of the HTML and is interoperable with other standard components within the page, either native elements, or from other libraries using the same standard concepts, regardless of the framework used to build the site.

It started as an effort from UniProt, InterPro, and PDBe, in order to build one set of components that would be reused within these resources' websites - like the full-fledged ProtVista visualization - but also have the flexibility to extend their specific needs. It now contains 20+ elements and has been adopted and customized by many data resources and tools.

In developing this library, we focused on interoperability, interaction with other libraries, and simplicity of use. As a permanent work in progress, we would like to engage with other visualization efforts in order to discover and promote best practices around the use of web standards in libraries and promoting the use of web components by other biological visualizations authors

Data

Omics data integration with Correlation guided Network Integration (CoNI)

José Manuel Monroy Kuhn (Computational Discovery Research Unit, Helmholtz Zentrum München, Neuherberg, Germany), Sonja C. Schriever (Research Unit Neurobiology of Diabetes, Helmholtz Zentrum München, Neuherberg, Germany), Viktorian Miok (Computational Discovery Research Unit, Helmholtz Zentrum München, Neuherberg, Germany), Andreas Peter (Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Center Munich at the University of Tübingen), Martin Heni (Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Center Munich at the University of Tübingen), Paul T. Pfluger (Research Unit Neurobiology of Diabetes, Helmholtz Zentrum München, Neuherberg, Germany) and Dominik Lutter (Computational Discovery Research Unit, Helmholtz Zentrum München, Neuherberg, Germany).

Abstract:

In the omics era, high throughput methods promise to unlock an unprecedented understanding of biological systems. However, despite this development, there is a great need for tools to analyze and integrate the complex data generated by these new methods. To this end, we developed the method Correlation Guided Network Integration (CoNI) and implemented it as an R package.

CoNI is a fully unsupervised correlation-based method for integrating two numerical omics datasets as a complex hypergraph-like network. The output allows for multiple network representations. CoNI can identify relevant biological confounders enriched for a specific subnetwork of correlated vertices. In addition, the output networks from different conditions can be compared, and a followed-up analysis can be applied to these networks.

We applied CoNI on metabolomics and transcriptomics data from murine livers under standard chow or high-fat diet. As a result, we could pinpoint eleven genes with a potential regulatory effect on liver metabolism. Furthermore, five of them were validated in humans by their expression correlation with diabetes-related traits such as overweight, hepatic fat content, and insulin resistance (HOMA-IR). Additionally, we ran CoNI with simulated data and compare the results to those obtained with other unsupervised integration methods. CoNI showed overall good performance and unique results not obtained by the other methods.

CoNI is a versatile data-driven R package for integrating two numerical omics datasets from the same samples. Our method can identify priority candidates of biological importance or compare network structures dependent on different conditions.

Data

Panomicon, allowing heterogeneous multi-omics analysis on the web

Rodolfo Allendes (National Institutes of Medical Innovation, Health and Nutrition), Johan T. Nystroem-Persson (Lifematics Inc / JNP Solutions), Yuji Kosugi (Lifematics Inc), Kenji Mizuguchi (Osaka University / National Institutes of Biomedical Innovation, Health and Nutrition) and Yayoi Natsume-Kitatani (National Institutes of Biomedical Innovation, Health and Nutrition).

Abstract:

The last decade has seen a continuous increase in the number of multi-omics-based studies [1]. Based on various technological developments and the corresponding increase in data availability, the combination of transcriptomics, genomics, proteomics, and so on, has led to important new insights in many areas of biology and medicine.

However, in order to properly support this type of analysis, software capable of handling the particular difficulties associated with multi-omics datasets is crucial. In an effort to address such challenges, we previously developed Panomicon [2], a web-based, interactive analysis environment for multi-omics data. From this initial publication, and through several iterations, Panomicon has evolved into a solution that provides the tools required for the storage and handling of heterogeneous omics data, together with functionality aimed at its analysis.

Panomicon comprises both a back-end and a front-end. While the back-end is implemented in scala and is optimized to handle the database and data requests, the front-end is an Angular application that focuses on displaying data and performing lightweight analysis.

Panomicon is free to use under an MIT license scheme. Users can choose between uploading their data to our server (data is not necessarily publicly available to others); or downloading the source code and hosting their own deployment of the tool.

[1] Sonia Tarazona et al. Undisclosed, unmet and neglected challenges in multi-omics studies. Nature Computational Science, 2021.

[2] Allendes Osorio RS et al. Panomicon: A web-based environment for interactive, visual analysis of multi-omics data. Heliyon. 2020 Aug 19;6(8)

Data

PascalX: Detecting shared genes and pathways between pairs of GWAS traits

Daniel Krefl (University of Lausanne) and Sven Bergmann (University of Lausanne).

Abstract:

Genome-wide association studies (GWAS) are well established for identifying links between genotypes and phenotypes in terms of individual SNP effect sizes. Aggregating such effects at the level of genes or pathways map potential signals to biologically relevant entities. Different traits may share functional genes and pathways, but we lack efficient tools for uncovering such shared mechanisms directly from GWAS summary statistics.

Here we present a rigorous, yet fast method for detecting genes and pathways with consistent association signals for two traits, facilitating local fine-coarsed cross-GWAS analyses. To this end we devised a new significance test for covariance of data-points not drawn independently, but with known inter-sample covariance structure. We show that the distribution of its test statistic is a linear combination of chi-squared distributions enabling us to calculate the corresponding cumulative distribution function efficiently. Our software PascalX uses this framework to test for dependence between SNP-wise effect sizes of two GWAS at the level of genes. The resulting gene signals can be further aggregated to the pathway level. We demonstrate the utility of our method by uncovering potential genetic links between severity of COVID-19 and specific other traits entirely from GWAS summary statistics.

PascalX will be useful for cross-GWAS and similar analyses involving data-points with known auto-correlation structures.

Data

Pilot: Making Research Data FAIR

Christina Yung (Indoc Research), Moyez Dharsee (Indoc Research), Fan Dong (Indoc Research), Kenneth Evans (Indoc Research), Susan Evans (Indoc Research), Tom Gee (Indoc Research), Mojib Javadi (Indoc Research), Stephane Pollentier (Indoc Research) and Shahab Shahnazari (Indoc Research).

Abstract:

Pilot (https://github.com/PilotDataPlatform) is a comprehensive data platform solution for secure management, analysis and sharing of medical research data. It is an open-source product developed by Indoc Research, a not-for-profit company in Canada and its not-for-profit subsidiary in Germany. Pilot consists of an intuitive Web Portal, a protected "Green Room" zone for initial data landing and preprocessing, automated ingestion of data from diverse health data sources (e.g.. electronic medical records, radiological images, genomic sequences), data warehouses for structured data, a metadata repository and knowledge graph for representation and query of semantic information, and extensive capture of data provenance and data lineage. By having specialized data zones such as the Green Room, sensitive information can be securely isolated for pseudonymization and transformation before being made available to research teams.

Pilot has a web-based Analytics Workspace for processing, analyzing, and visualizing datasets, including integration with business intelligence tools, Jupyter notebooks, and high-performance computing (HPC) resources. Pilot is developed with privacy by design to provide a secure, reliable, scalable and fault-tolerant infrastructure. It is a portable solution that has been deployed to different virtualized computing platforms both on premise and commercial clouds. Pilot has been deployed as a GDPR-ready Virtual Research Environment (VRE) at the Charité Hospital in Berlin. Pilot's FAIR (Findable, Accessible, Interoperable, Reusable) capabilities and architecture are being further expanded in the Health Data Cloud (https://healthdatacloud.eu/) through integration with the Human Brain Project's EBRAINS services to support management of sensitive data across a federated network of data centres.

Data

Predicting Gene Dependencies using Machine Learning from Proteomic Data

Robert Nkwo (Barts Cancer Institute, Queen Mary University of London, United Kingdom.), Shirin Khorsandi (King's College London, United Kingdom.) and Pedro Cutillas (Barts Cancer Institute, Queen Mary University of London, United Kingdom. The Alan Turing Institute.).

Abstract:

The mapping of essential genes in human cancer cells via CRISPR/Cas9 knockout or RNAi silencing, to reveal gene dependencies, offers an attractive opportunity for identifying potential drug targets in cancer. However, modelling gene dependencies using data at the protein level has not been systematically explored. Therefore, here, we aimed to identify relationships between protein expression and essential cancer genes and subsequently building a machine learning (ML) algorithm to predict gene dependency in tumours without the need for extensive functional screening. To this end, we employed statistical learning methods to predict CRISPR/Cas9 derived gene dependencies for 1355 genes using proteomics data. For predicting gene dependencies, we built Random Forest (RF) regression models which takes proteomic data as features and DepMap 21Q4 CRISPR/Cas9 gene dependency as labels. RF models were trained using 5-fold cross validation on 85% of label-free proteomics data (PRIDE identifier PXD013455) including 88 solid tumour cell lines across 11 different cancer types and 7020 proteins. We evaluated model performance on predicting gene dependencies using an 85% training set, 15% unseen test set, and an independent validation set consisting of 18 cell lines (Gerdes et.al). The overall performance for the regression models across all data sets showed the following mean square error (MSE) between measured CRISPR/Cas9 gene dependencies and those predicted using proteomic data: train MSE = 0.036, test MSE = 0.055, validation MSE = 0.057. Predicting gene dependency in primary tumours with our ML models could help elucidate patient specific drug targets, thereby assisting management in the clinic.

Data

PROMPT: Toward PRecisiOn Medicine for the Prediction of Treatment response in major depressive disorder through stratification of combined clinical and -omics

Júlia Perera-Bel (Hospital del Mar Medical Research Institute), Alessandra Minelli (University of Brescia; IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli), Johannes Zang (University of Münster), Britta Kelch (University of Münster), María Martínez de Lagrán (Center for Genomic Regulation (CRG)), Mara Dierssen (Center for Genomic Regulation (CRG)), Bernardo Carpiniello (University of Cagliari), Massimo Gennarelli (University of Brescia; IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli), Filip Rybakowski (Poznan University of Medical Sciences), Marie-Claude Potier (Paris Brain Institute ICM, Salpetriere Hospital), Ferran Sanz (Hospital del Mar Medical Research Institute) and Bernhard T Baune (University of Münster).

Abstract:

Major depressive disorder (MDD) is the most common psychiatric disease worldwide, with huge socio-economic impacts. Pharmacotherapy represents the first-line therapeutic choice, but about 30% of patients are treatment resistant (TRD). TRD is associated with specific clinical and biological features; however, taken individually, these signatures have limited power in response prediction. The project's aim is the development of a predictive algorithm that integrates different omics and clinical layers for the early detection of non-responder patients, more prone to later develop TRD. Recruited patients will undergo a comprehensive clinical assessment and molecular profiling (genomic, transcriptomic and miRNomic). An algorithm integrating all these data will be developed to predict response to therapy and evaluated in a prospective cohort. We have already recruited 192 patients with MDD, including 104 TRD/88 responders. This cohort is composed of 70% of females, equally represented in both groups. BMI and age are associated with TRD, as well as psychiatric comorbidities (e.g. anxiety and personality disorders). We have identified differentially expressed genes between the two groups. Though still preliminary, we observe a downregulation of immune related pathways in TRD patients. Once the project will be completed, it will perform a multi-omic integration with which we expect to gain a comprehensive understanding of the molecular mechanisms driving TRD including differences between groups of patients (e.g. sex specific differences). In the translational perspective, this project will provide a new predictive tool for future use in the clinical practice, enabling a better prevention and management of MDD treatment resistance.

Data

RDMkit: the ELIXIR Research Data Management Kit

Laura Portell Silva (Barcelona Supercomputing Center) and Munazah Andrabi (University of Manchester).

Abstract:

Tailored guidance for management of research data is increasing in importance throughout the lifecycle of data-driven investigations. On the one hand, funders or host institutions demand research data management (RDM) by asking researchers to develop and implement data management plans for projects. On the other hand, research infrastructures provide a plethora of RDM support in the form of tools, policies, standards and guidelines. Researchers often find themselves lost in the middle, overwhelmed in their task to take advantage of the available support and to meet the funder demands for RDM.

To address these issues, RDM experts in the ELIXIR community have combined forces, through ELIXIR-CONVERGE, to build a toolkit called the RDMkit. The RDMkit is an online guide containing good data management practices applicable to research projects from the beginning to the end. RDMkit provides different entry points based e.g. on different roles, and highlights the different aspects of making life-science data Findable, Accessible, Interoperable and Reusable (FAIR). The RDMkit integrates with other ELIXIR tools and resources such as FAIRCookbook, FAIRsharing, bio.tools, Data Stewardship Wizard and TeSS training portal enabling the user to deliver a seamless data management plan and form a RDM knowledge commons for the ELIXIR community.

Data

Recommendations on training VAEs on TCGA transcriptome data

Mostafa Eltager (Delft University of Technology), Tamim Abdelaal (Leiden University Medical Center), Mohammed Charrout (Delft University of Technology), Ahmed Mahfouz (Leiden University Medical Center), Marcel Reinders (Delft University of Technology) and Stavros Makrodimitris (Delft University of Technology).

Abstract:

Deep generative models, such as variational autoencoders (VAE), have gained increasing attention in computational biology due to their ability to capture complex data manifolds and make predictions for new drugs in complex diseases like cancer. However, these models are difficult to train as they are sensitive to the enormous number of parameters. To get a better understanding of the importance of the different hyperparameters, we examined six different VAE models when trained on TCGA transcriptomics data. We studied the effect of the size of the latent space, learning rate, optimizer and initialization on the quality of subsequent clustering of the TCGA samples when evaluated in comparison to the known cancer types. We found that beta-TCVAE and DIP-VAE are the top performing VAE models on average, and derive recommendations on how select the different hyper-parameters settings. Next, we examined whether the learned latent spaces capture biologically relevant information. Hereto, we correlated the different representations with various data features such as age, metastasis date, immune infiltration, and mutation signatures. Results show that all models learn a latent space that encodes most of the factors tested, but not always in a disentangled way, even for models specifically designed for disentanglement.

Data

Rhea, a FAIR resource of expert curated biochemical and transport reaction data.

Parit Bansal (SIB), Anne Morgat (SIB, Swiss Institute of Bioinformatics), Kristian Axelsen (SIB), Venkatesh Muthukrishnan (SIB), Elisabeth Coudert (SIB Swiss Institute of Bioinformatics), Lucila Aimo (SIB), Nevila Hyka-Nouspikel (SIB Swiss Institute of Bioinformatics), Elisabeth Gasteiger (SIB Swiss Institute of Bioinformatics), Arnaud Kerhornou (SIB), Teresa Neto (SIB Swiss Institute of Bioinformatics), Monica Pozzato (SIB), Marie-Claude Blatter (SIB), Nicole Redaschi (SIB Swiss Institute of Bioinformatics) and Alan Bridge (SIB Swiss Institute of Bioinformatics).

Abstract:

Rhea (www.rhea-db.org) is a FAIR resource of expert curated biochemical and transport reactions described using the ChEBI ontology of small molecules (www.ebi.ac.uk/chebi/). Rhea is the reference vocabulary for enzyme annotation in UniProtKB (www.uniprot.org) and provides reference reaction data for the Gene Ontology, Reactome, MetaboLights, and a host of other knowledge and data resources.

In this poster, we will describe progress in Rhea curation efforts and demonstrate how to access Rhea data via our website, API, and SPARQL endpoint (https://sparql.rhea-db.org/sparql) to query and download in a range of formats for biologists, chemists and semantic web users.

Data

ROLE OF RNA-SEQ PREPROCESSING STEPS IN MOLECULAR SUBTYPE CLASSIFICATION IN MUSCLE-INVASIVE BLADDER CANCER

Ariadna Acedo-Terrades (Hospital del Mar Medical Research Institute(IMIM), Research Programme of Biomedical Informatics(GRIB); Barcelona, Spain), Júlia Perera-Bel (Hospital del Mar Medical Research Institute(IMIM), Research Programme of Biomedical Informatics(GRIB); Barcelona, Spain), Joaquim Bellmunt (Hospital del Mar Medical Research Institute(IMIM);Barcelona,Spain,Division of Hematology and Oncology,BIDMC; Boston,USA) and Lara Nonell (Vall d'Hebron Institute of Oncology (VHIO); Barcelona, Spain).

Abstract:

Background: Muscle-invasive bladder cancer (MIBC) is a disease that is characterized by genomic instability and a high mutation rate. Therefore, transcriptome profiling has been used to classify MIBC into six molecular subtypes. Transcriptomic profiling can be obtained using RNA-Seq in which different methods (ie. alignment, quantification) can be used to obtain the final table of counts. Our hypothesis is these steps affect the resulting table of counts and, hence, might also influence the classification into molecular subtypes.

Objective: Assess the role of RNASeq preprocessing steps in the molecular subtype classification of bladder cancer samples

Materials and Methods: A study of different preprocessing methodologies was conducted by comparing: STAR and Hisat2 tools for the alignment step, and featureCounts, HTSeq, StringTie and RSEM for the quantification step. Regarding the normalization step, the methodologies that were used are: TPM, log2TPM, TMM, rawData and log2rawData. We applied the pipelines to 3 MIBC datasets from the GEO database. Accuracy was evaluated using the classification label obtained from the consensusMIBC classifier.

Results: Our preliminary results demonstrate that STAR was the best aligner, producing the highest accuracy and being the most stable classification results. Regarding quantifiers, almost all of them have high accuracy. Finally, TMM and log2TPM were the normalization methods that showed a high accuracy across almost all the methods used for alignment and quantification steps.

Conclusions: According to our results, we propose STAR+featureCounts+TMM as the most accurate pipeline to generate a counts table to use for downstream molecular subtype classification.

Data

scDAVIS: Single-cell Data Analysis and VISualization

Carlos Torroja (Centro Nacional de Investigaciones Cardiovasculares), Daniel Jimenez Carretero (Centro Nacional de Investigaciones Cardiovasculares), Jon E. Sicilia (Centro Nacional de Investigaciones Cardiovasculares), Juan L. Onieva Zafra (Centro Nacional de Investigaciones Cardiovasculares), Jorge G. Garcia Gomez (Centro Nacional de Investigaciones Cardiovasculares), Celia Centeno Tundidor (Centro Nacional de Investigaciones Cardiovasculares) and Fatima Sanchez-Cabo (Centro Nacional de Investigaciones Cardiovasculares).

Abstract:

The possibility of studying phenotypical and molecular characteristics of cells in a high-throughput manner is changing our understanding of biological systems. However, the analysis of single-cell omics data requires the use of computational tools that need to be tailored and correctly interpreted. This fine-tuning is a time-consuming process that requires expert curation to extract relevant information out of the data. Also, different high-throughput techniques are often combined, ranging from cytometry to imaging and transcriptomics.

scDAVIS (https://bioinfo.cnic.es/scdavis/) is a web-based tool, open to the whole community, for the analysis and visualization of single-cell omics data of different types, including scRNA-Seq, imaging data, and cytometry. Importantly, scDAVIS is also a repository of publicly available single-cell experiments, allowing further exploration and reuse of data, in line with the compliance of FAIR principles. Also, it can serve as a repository to share new data with the community and to ease integration.

scDAVIS has the following modules: (1) Analysis of raw data, including steps for feature selection, transformation/scaling, dimensionality reduction (PCA/tSNE/UMAP), and clustering from any of the modalities previously specified. (2) Upload previously processed data. (3) Load published analysis (currently, 64 public datasets), searchable through a keyword-based search engine. (4) Download analysis, exporting data for future reanalysis. (5) QC-Stats with basic information about cells and features/genes profiled (6) Plots for visualization of results: Dim-Plots, Violin-Plots, Bar-Plots, Heatmaps, Scatter-Plots and Dot-Plots. Finally, scDAVIS incorporates interactive tools for results representation, statistical contrasts, data filtering, and manual annotation/correction.

Data

Short read de novo transcriptome assembly by means of ant colony optimization

Karl Johan Westrin (KTH CBH), Olof Emanuelsson (KTH Royal Institute of Technology) and Henric Zazzi (PDC KTH).

Abstract:

Several methods have been proposed to solve the NP-hard problem of sequence assembly, both for genomic and transcriptomic data. Most of these methods, however, are either computationally inefficient or performs less than optimal on several datasets.

We are developing a novel transcriptome assembler, which makes use of the meta-heuristic ant colony optimization to extract the transcripts from a compacted de Bruijn graph. This is implemented with the memory cheap graph framework Bifrost and uses the hash counter from Jellyfish to obtain sequencing depth. We hope that this assembler will be on par with state-of-art assemblers, but more time- and memory efficient. Further studies on more datasets will be made.

Data

Single-cell transcriptomic analyses reveal distinct B cell subsets and their class-switch recombination dynamics

Joseph Ng (King's College London), Alexander Stewart (University of Surrey), Deborah Dunn-Walters (University of Surrey) and Franca Fraternali (King's College London).

Abstract:

B cells are effective antigen-presenting cells which are capable to differentiate into antibody-producing plasma cells. An important biological process in humoral response is class-switch recombination (CSR), where B cells change the isotype of their receptors in order to partake in different downstream signalling events. The characterisation of B cells and CSR traditionally relies heavily on labour-intensive techniques, e.g. investigating surface protein expression and molecular cloning to identify CSR events. We show here that single-cell RNA sequencing data can be used to offer deep insights into characterising B cell states and their CSR dynamics. By profiling B cells from peripheral blood of healthy individuals, we identify distinct transcriptomic states which reflect their surface protein characterisation (Stewart, Ng et al Front Immunol 2021, 12:602539), and apply these data as cell atlas to understand how B cells deviate from these "ground-states" during immune challenges. We also hypothesise that these data contain "sterile" immunoglobulin transcripts which prime CSR events, but are conventionally disregarded in raw data processing workflows due to the omission of such transcripts in reference transcriptome annotations. By explicitly enumerating these transcripts, we can distinguish B cell subsets by their CSR states, in terms of the current isotype of their B cell receptors as well as the isotypes to which they are primed to switch. These analyses will allow us to predict the dynamics of CSR by means of mathematical modelling, and investigate how CSR interacts with clonal expansion and the overall transcriptional dynamics of B cells to shape its developmental pathway.

Data

Software Observatory

Eva Martin del Pico (Barcelona Supercomputer Center (BSC)), Salvador Capella-Gutiérrez (Barcelona Supercomputing Center (BSC)) and Josep Ll Gelpi (Dept. Bioquimica i Biologia Molecular. Univ. Barcelona).

Abstract:

The Software Observatory aims to be an instrument for the systematic observation and diagnosis of the quality of research software in the Life Sciences. The ultimate goal is promoting the adoption of software development best practices, to which the Software Observatory can contribute through the identification of trends in the way research software is being developed. This can help detect needs and design strategies to be adopted at individual and community levels.

The Software Observatory is a constitutive part of OpenEBench, an initiative developed within ELIXIR, that aims to provide a permanent platform to support benchmarking in Life Sciences. The Software Observatory complements the scientific evaluation of research software providing quality motoring of over 40,000 tools. Features analyzed are those directly related to FAIR for Software Principles and indicators. These include version handles, license usage and journal publication patterns among others. Moreover, an overview of FAIR for software scores broken down by principles and identifiers is also available. Software metadata and metrics used for the analysis are collected from registries and further enriched using code repositories. Other metrics such as publication citations and site accessibility are generated purposely by OpenEBench.

Data

SQANTI3: how to curate a Long Read-defined transcriptome in 3 steps.

Francisco J. Pardo-Palacios (I2SysBio), Ángeles Arzalluz Luque (I2SysBio) and Ana Conesa (I2SysBio).

Abstract:

Long Read (LR) Sequencing has changed the way we build new transcriptomes. Nowadays, it's feasible to obtain a sample-specific reference transcriptome without the issues associated with short read assemblies. However, as a consequence of the high sequencing error rate, many transcript models are incorrectly defined.

Here we present SQANTI3, the newest version of the already widely accepted SQANTI tool. SQANTI3 performs not only quality control analysis, but it also includes a complete pipeline to curate a transcriptome. It will make the most of all the descriptive attributes associated to a detected isoform to take an informed decision about if it should be included or excluded from the final transcriptome. To do so, SQANTI3 follows three steps:

1. Quality Control: Compare the initial transcriptome with a reference annotation, gathering around each isoform as much evidence as possible and classifying in up to nine structural categories.

2. Filtering: Remove possible sequencing artifacts. SQANTI3 allows two approaches: rules filtering and Machine-Learning filtering. Rules filtering is based on defining a set of characteristics that a reliable isoform must fulfill to accept it. ML-filter will use a random forest algorithm to build a classifier that will distinguish between a likely true isoform and a false one.

3. Rescue: Those isoforms classified as artifacts will be re-evaluated to bring back their possible missing source from the reference annotation. This avoids the loss of complete genes because all the isoforms detected were not good enough to pass the filters.

Data

SQANTISIM: a simulator of controlled novelty and degradation of transcripts sequenced by long-reads

Jorge Mestre-Tomás (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia Spain), Francisco J. Pardo-Palacios (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia Spain) and Ana Conesa (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia Spain).

Abstract:

Computational methods developed for isoform identification and discovery detect many new transcript models even in well-annotated organisms. However, long-read transcriptomics technologies and reconstruction algorithms are not error-free and one of the biggest questions in the field is how accurate the different methods are to identify both annotated and novel transcript models. While simulated data is an effective strategy to benchmark novel methods, existing simulators for long-read RNA-seq data do not support the simulation of novel transcripts. This limits our ability for evaluating the capacity of lrRNA-seq to confidently detect the novelty associated with these new technologies.

Here, we present SQANTISIM, a lrRNA-seq simulation engine that simulates Nanopore and PacBio long reads with precise control of transcript novelty based on SQANTI3 structural categories. SQANTISIM implements the following steps:

1. Classify all the transcripts annotated in the reference GTF for their potential SQANTI3 structural category when compared to other transcripts.

2. Remove a user-defined number of transcripts classified as novel from the annotation to produce a “reduced” annotation.

3. Simulate reads using either NanoSim or IsoSeqSim based on the complete reference annotation. The evaluated transcriptome reconstruction algorithm should use the simulated data, together with the “reduced” reference to predict transcript models.

4. SQANTI3 is run with the reduced GTF as reference and the reconstructed long read-defined transcriptome. The deleted transcripts should be identified as novel, and any novel transcript not simulated will be false calls.

Four popular lrRNA-seq reconstruction methods were benchmarked to demonstrate the effectiveness of this tool.

Data

Standardized analysis of complex RNAseq experiments using Snakemake

Christian M. Heyer (Biomedical Infor., Fac. of App. Inf. and Med. Fac., U. of Augsburg; DKFZ Grad. School; Faculty of Biosci. Heidelberg U.), Ashik Ahmed Abdul Pari (Div. of Vascular Onco. and Metast., (DKFZ-ZMBH); Dep. of Vasc. Biol. and Tumor Angio. Fac. of Biosci., Heidelberg U.), Hellmut G. Augstin (Dep. of Vasc. Oncology (ECAS), Med. Fac. Mannheim; Heidelberg U., and (DKFZ-ZMBH Alliance) Heidelberg, Germany) and Matthias Schlesner (Biomedical Informatics, Data Mining and Data Analytics, Faculty of Applied Infor. and Med. Faculty, Uni. of Augsburg).

Abstract:

Advances in high-throughput RNA sequencing have allowed the analysis of experimental setups covering two or more conditions and analyzing more than one variable in an experiment. While processing, alignment and quality control is identical to experiments analyzing two conditions, downstream analysis steps to characterize changes between conditions increase in complexity. Nevertheless, core downstream analysis steps, such as differential gene expression and enrichment analysis, are shared and can be run in a standardized manner.

Building off the snakemake workflows repository, this workflow provides a standardized downstream analysis workflow for RNAseq quantified gene expression data, covering differential expression analysis and various enrichment analysis for datasets with multiple conditions. By using snakemake as its workflow management system, the generation of HTML reports covering QC, exploratory- and enrichment analyses were standardized. Besides differential gene expression analysis with DESeq2 and gene set enrichment analysis, transcription factor activity and signaling pathway activities can be inferred using PROGENy and Dorothea, respectively. Together these analyses assist in inferring changes in gene regulatory networks and help shed light on changes in intracellular signaling.

Here, this workflow has been run a dataset of cell-sorted endothelial cells from apelin-treated young (eight weeks old) and aged mice (15-month-old), illustrating how it robustly supports the analysis of experiments covering more than two experimental conditions.

Further work is planned in supporting the analysis of transcript-level expression data, workflow dockerization and integration testing.

Data

Systematic benchmarking and error evaluation of basecallers for Nanopore Direct RNA-seq

Wang Liu-Wei (Systems Medicine of Infectious Disease (P5), Robert Koch Institute), Patrick Bohn (Helmholtz Centre for Infection Research (HIRI)), Wiep van der Toorn (Systems Medicine of Infectious Disease (P5), Robert Koch Institute), Redmond Smyth (Helmholtz Centre for Infection Research (HIRI)) and Max von Kleist (Systems Medicine of Infectious Disease (P5), Robert Koch Institute).

Abstract:

Traditionally, high-throughput RNA sequencing protocols have relied on reverse transcription and amplification, which introduce various errors and biases that confound downstream analysis. Direct RNA sequencing (dRNA-seq) on the Oxford Nanopore platforms has been developed in recent years with the potential to cover full-length transcripts. Moreover, dRNA-seq data contain decipherable information of RNA base modifications in the form of signal alterations. This makes the platform valuable for studying endogenous RNA modifications that are epigenetic (e.g. methylation), as well as exogenous modifications from chemical probing experiments that encode structural information of RNA.

Similar to long-read DNA sequencing, the electrical signal output of a Nanopore sequencer needs to be translated to RNA sequences with machine learning algorithms, a process known as “basecalling”. While the basecalling of dRNA-seq is known to be error-prone, the performance and error characteristics of various basecallers have not yet been extensively benchmarked, which limits the interpretation of downstream analysis, including RNA modification detection.

In this study, we comprehensively benchmarked the performance of existing basecallers on diverse dRNA-seq datasets. We found that the basecalling errors are ubiquitous and show systematic biases towards specific sequence contexts/motifs. Furthermore, we discuss and investigate where the error characteristics of dRNA-seq originate, and how they can be accounted for and used to improve existing pipelines for RNA modification detection. These results strongly suggest the importance of error normalization/correction in dRNA-seq data analysis.

Data

Temporal changes in microbiome composition after FMT in subject with Clostridioides difficile infection: a network perspective

Marco Cappellato (Department of Information Engineering, University of Padova, Padova, Italy), Massimo Bellato (Department of Information Engineering, University of Padova, Padova, Italy), Giacomo Baruzzo (Department of Information Engineering, University of Padova, Padova, Italy), Sonia Facchin (Department of Surgery, Oncological and Gastroenterological Sciences, University of Padova, Padova, Italy), Luisa Barzon (Department of Molecular Medicine, University of Padova, Padova, Italy), Valeria Besutti (Azienda Ospedale Padova, Padova, Italy), Paola Brun (Department of Molecular Medicine, University of Padova, Padova, Italy), Simone Del Favero (Department of Information Engineering, University of Padova, Padova, Italy), Luca Schenato (Department of Information Engineering, University of Padova, Padova, Italy), Edoardo Vincenzo Savarino (Department of Surgery, Oncological and Gastroenterological Sciences, University of Padova, Padova, Italy), Ignazio Castagliuolo (Department of Molecular Medicine, University of Padova, Padova, Italy) and Barbara Di Camillo (Department of Information Engineering, University of Padova, Padova, Italy).

Abstract:

Clostridioides difficile is a spore-forming, anaerobic Gram-positive bacterium, physiologically present in the gut microbiota and historically considered one of the main nosocomial pathogens associated with antibiotic resistance. Over the past two decades, the incidence of C. difficile-associated disease (CDAD) has grown significantly in countries with higher health standards, along with the severity of cases, relapses, and mortality. CDAD follows disruption of the indigenous gut microbiota by antibiotics, which promotes C. difficile growth, leading to intestinal damage and colitis. Fecal microbiota transplantation (FMT) is emerging as the most effective treatment for recurrent CDAD. In this work, we decipher the network of influence among microorganisms and the evolution of the microbiota toward a diseased or back to a healthy state in CDAD by exploiting 16S rDNA-seq data. Stools samples are collected from healthy subjects and CDAD patients at different time points (t). Then, after rDNA-seq technique, a bioinformatics pipeline is used to pre-process data, obtaining a taxonomic profile for each subject. We infer microbial interaction network from abundance data exploiting a reverse engineering approach. Finally, a complete evaluation of occurrence patterns, network partitions and differential network over time are performed. We detect possible targets for therapeutic intervention, i.e. bacterial species to be up/down regulated to maintain or return to a healthy state. In the future, our project will provide a set of validated biomarkers to be correlated with (and validated by) clinical endpoints.

Data

Text mining resources for extracting genetic and phenotypic data from scientific publication full-texts and tables

Thomas Rowlands (University of Leicester), Janet Pinero (University Pompeu Fabra), Pablo Accuosto (University Pompeu Fabra), Tom Shorter (University of Leicester), Joram Posma (Imperial College London), Laura Furlong (University Pompeu Fabra) and Tim Beck (University of Leicester).

Abstract:

Genome-wide association studies (GWAS) provide a deeper understanding of disease aetiology by detecting associations between genetic variants and disease traits in population samples. Comprehensive curated online databases such as GWAS Central (https://www.gwascentral.org/) and DisGeNET (https://www.disgenet.org/) enable the convenient visualisation and interrogation of GWAS findings. These, and similar data resources, are reliant on the efforts of biocurators to interpret and import information published in the scientific literature. Text mining has been shown to facilitate accelerated biocuration and has a role in enabling the scalable extraction of human genotype and phenotype entities from the scientific literature. We are consolidating, developing, and using text mining tools and resources to support the automated extraction of human genotype-phenotype associations from publications. These new capabilities will be implemented to support data curation activities in the GWAS Central and DisGeNET databases.

We are developing a reusable text mining workflow that will identify genetic and phenotypic entities and relations from scientific literature full-texts, tables and supplementary materials. The Auto-CORPus text standardisation tool [1] is applied to diverse sources of biomedical text to convert them to the standardised BioC computer interpretable format that is processed by the workflow. Additionally, we are building a semi-automatically annotated full-text GWAS corpus that will be used to evaluate the performance of the text mining workflow and will be available to the text mining community to benchmark future methodologies.

1.Beck, T, et al.(2022)Auto-CORPus: A Natural Language Processing Tool for Standardizing and Reusing Biomedical Literature.Front Digit Health. 4:788124. https://doi.org/10.3389/fdgth.2022.788124

Data

The Wisdom of the Crowd: comparing individual and aggregated solutions for metagenomics-based inflammatory bowel disease diagnostics in the scope of the sbv IMPROVER MEDIC challenge

Lusine Khachatryan (Philip Morris International R&D), Carine Poussin (Philip Morris International R&D), Yang Xiang (Philip Morris International R&D), Adrian Stan (Philip Morris International R&D), James Battey (Philip Morris International R&D), Giuseppe Lo Sasso (Philip Morris International R&D), Stephanie Boue (Philip Morris International R&D), Nicolas Sierro (Philip Morris International R&D), Nikolai Ivanov (Philip Morris International R&D) and Julia Hoeng (Philip Morris International R&D).

Abstract:

A growing number of reports showing changes in gut microbiota in subjects with inflammatory bowel disease (IBD) indicate the benefit of exploiting metagenomics for noninvasive disease diagnostics. To investigate the diagnostic potential of metagenomics data to discriminate patients with IBD from non-IBD subjects, we organized the crowdsourced sbv IMPROVER Metagenomics Diagnosis for Inflammatory Bowel Disease Challenge (MEDIC), which was open to the worldwide scientific community from September 2019 to March 2020.

Overall, 81 anonymized submissions, each exploiting a unique combination of algorithms for inter-cohort metagenomics-based IBD diagnostics, were received, scored, and analyzed. These submissions were used to explore the phenomenon known as “the wisdom of the crowd,” wherein the collective knowledge of a community is greater than the knowledge of any individual. We investigated the quality of IBD diagnostics using random aggregations of independent individual predictions, predictions based on the same metagenomics data type, and predictions across two different data types.

Aggregated predictions were more robust to random noise and achieved on average better diagnostic quality than individual predictions. These findings demonstrate the necessity of exploiting different analysis pipelines for metagenomics data processing and may be useful for the design of a strategy for non-invasive IBD diagnostics.

Data

tolDB: A framework to integrate Tolerogenic Dendritic Cells data using linked knowledge graph and NLP

Ayesha Sahar (Newcastle University), Phillip Lord (Newcastle University), Catharien Hilkens (Newcastle University), Michael Hughes (SciBite) and Lee Harland (SciBite).

Abstract:

Tolerogenic dendritic cells (tolDCs) are immunoregulatory antigen-presenting cells and their use as an immunotherapeutic tool is becoming increasingly promising. Data integration of knowledge about tolDCs could be hugely beneficial for the immunoregulatory field in several ways. However, the field suffers from lack of data reporting and standardisation. Consequently, the reported data is sparse and heterogeneous to repurpose or reuse. Given that, published articles are the main source of information present in the field.

This paper provides an empirical approach to extract information and build knowledge graph for tolDCs from published articles. A corpus was built based on directly related keywords from PubMed. Then, SciBite's Semantic Analysis Platform was used to extract the important entities such as genes, diseases, cell markers, drugs etc. Structured information available in the form of web resources was also integrated with these extracted entities, which includes pathways, gene-gene, gene-disease and gene-drug associations. After integration, a knowledge graph was built to visualise the key nodes, information and links between the literature and entities.

As a result of this, 1.2m entities were extracted. Top reported 38% entities are genes. In addition to providing specific sets of entities to the tolDC field, the resulting graph can be used for numerous useful operations such as filtering the papers based on the co-occurrence of one or more entities, visualising the pathways or gene associations among others. This shows the potential of leveraging NLP and knowledge graph technologies to integrate data for tolDC and driving future research.

Data

Towards a knowledge graph for pre-/probiotics and microbiota-gut-brain axis diseases

Ting Liu (Vrije Universiteit Amsterdam), Gongjin Lan (Southern University of Science and Technology), K. Anton Feenstra (Vrije Universiteit Amsterdam), Zhisheng Huang (Vrije Universiteit Amsterdam) and Jaap Heringa (Vrije Universiteit Amsterdam).

Abstract:

Scientific publications present biological relationships but are structured for human reading, making it difficult to use this resource for semantic integration and querying. Existing databases, on the other hand, are well structured for automated analysis, but do not contain comprehensive biological knowledge. We devised an approach for constructing comprehensive knowledge graphs from these two types of resources and applied it to investigate relationships between pre-/probiotics and microbiota-gut-brain axis diseases. To this end, we created (i) a knowledge base, dubbed ppstatement, containing manually curated detailed annotations, and (ii) a knowledge base, called ppconcept, containing automatically annotated concepts. The resulting Pre-/Probiotics Knowledge Graph (PPKG), combining these two knowledge bases and three further public databases (i.e. MeSH, UMLS and SNOMED CT). To validate the performance of PPKG and to demonstrate the added value of integrating two knowledge bases, we created four biological query cases. The query cases demonstrate that we can retrieve co-occurring concepts of interest, and also that combining the two knowledge bases leads to more comprehensive query results than utilizing them separately. The PPKG enables users to pose research queries such as ``which pre-/probiotics combinations may benefit depression?'', potentially leading to novel biological insights.

Data

Towards a more inductive world: A review on graph embedding methods for drug repurposing approaches

Jesús De la Fuente (Tecnun University of Navarra), Guillermo Serrano (Center for Applied Medical Research (CIMA) University of Navarra), Uxía Veleiro (Center for Applied Medical Research (CIMA) University of Navarra), Mikel Casals (Tecnun University of Navarra), Oier Azurmendi-Senar (Center for Applied Medical Research (CIMA) University of Navarra), Antonio Pineda-Lucena (Center for Applied Medical Research (CIMA) University of Navarra), Idoia Ochoa (Tecnun University of Navarra), Silve Vicent (Center for Applied Medical Research (CIMA) University of Navarra), Olivier Gevaert (Stanford Center for Biomedical Informatics Research (BMIR), Stanford University) and Mikel Hernaez (Center for Applied Medical Research (CIMA) University of Navarra).

Abstract:

Motivation. In-silico prediction of drug-target interaction (DTI) facilitates drug repurposing tasks by reducing experimental costs. In this context, several machine learning methods have been proposed to predict DTIs considering complex heterogeneous graphs. These techniques, which generate embeddings from graph nodes, can be classified as inductive or transductive. While the former store information during training to generate embeddings for unseen nodes, the latter only generate embeddings for nodes seen during training leading to data leakage at the evaluation stage. Thus, it is of utmost importance to propose fair and generalizable evaluation guidelines. Besides, current golden standard datasets are small and outdated.

Approach. To address these issues, first, we proposed a novel subsampling approach as DTI data is imbalanced since non-positive labels are far more abundant than positive ones. Since evolution preserves protein structure more than sequence itself, we identified as plausible drug-target pairs those with a potential structural interaction and extracted them from the training set. Second, we proposed novel splitting approaches: we designed three different split modes to evaluate possible data leakages in transductive methods. To assess the proposed methodology, we selected four models of each type (inductive and transductive) and curated the latest versions of the required databases. Additionally, we augmented the DTI network for those methods that need complementary information, e.g., disease nodes. Finally, we present a novel and simpler GNN model that outperforms complex transductive methods.

Conclusion. Collectively, this review sheds light on the evaluation of DTI models and updated benchmark datasets.

Data

Towards interpretable machine learning applications in human microbiome via information theory-guided feature selection

Valentyn Bezshapkin (Małopolska Center of Biotechnology, Jagiellonian University, Gronostajowa 7A, Kraków, Poland), Witold Wydmański (Małopolska Center of Biotechnology, Jagiellonian University, Gronostajowa 7A, Kraków, Poland), Krzysztof Mnich (Computational Center, University of Białystok, Ciołkowskiego 1M, Białystok, Poland), Michał Kowalski (Małopolska Center of Biotechnology, Jagiellonian University, Gronostajowa 7A, Kraków, Poland), Dagmara Błaszczyk (Małopolska Center of Biotechnology, Jagiellonian University, Gronostajowa 7A, Kraków, Poland), Tomasz Kościółek (Małopolska Center of Biotechnology, Jagiellonian University, Gronostajowa 7A, Kraków, Poland), Witold Rudnicki (Computational Center, University of Białystok, Ciołkowskiego 1M, Białystok, Poland) and Paweł Łabaj (Małopolska Center of Biotechnology, Jagiellonian University, Gronostajowa 7A, Kraków, Poland).

Abstract:

The microbiome is often studied from a "vertical" perspective: a researcher observes perturbations in microbiome composition and associates it with the disease. However, the microbiome has its internal structure: microbes do interact with each other with the help of proteins, metabolites or horizontal gene transfer. Therefore, they should be viewed as a tightly interconnected community.

Mutual information (MI) is an information theory-based measure of dependence between two variables. It can be generalized into multiple dimensions, accounting for nonlinear and synergistic interaction between variables (Mnich and Rudnicki, 2020). We apply this method to analyse the link between microbiome data and different health conditions in the American Gut Project data (McDonald et al., 2018). We then use the results of the feature selection filter for classification using machine learning algorithms. Addition of important variables from 2D analysis improved classification in 2/5 analysed diseases (in diabetes and IBD). The procedure did not hurt classification performance, even though >60% of variables were discarded in each case. Additionally, we observed an increase in the stability of model weights. The correlation of model weights with MI increased as well with 1D and 2D analysis (0.45 vs. 0.78 vs. 0.86 respectively), thus contributing to the interpretability of the model.

To sum up, the use of MI in microbiome data helps to improve the interpretability of the machine learning models accounting for interaction effects between different bacteria. In further research, we would like to identify bacteria participating in such interactions and explore underlying biological mechanisms.

Data

Transmorph: A computational framework for dataset integration

Aziz Fouché (Institut Curie Paris), Loïc Chadoutaud (Insitut Curie Paris) and Andrei Zinovyev (Institut Curie Paris).

Abstract:

Dataset integration in single-cell describes the process of merging two or more datasets in a joint representation, so that cells of similar type end up close from one another. It can be used to produce joint embeddings of batches across different sources or technologies, to identify rare cell subpopulations or to correct counts of all datasets with respect to a reference one. Substantial efforts have been carried out over the last years to tackle this problem, and many approaches have been reported. We propose a computational data integration framework providing a tools to conceive, benchmark and run data integration models. We demonstrate this framework can be used to develop competitive data integration pipelines both in terms of time and performance, while providing great interpretability and modularity. This framework is implemented in a well documented, open source python package.

Data

tRNAstudio: facilitating the study of human tRNA-seq datasets

Marina Murillo Recio (IRB Barcelona - Institute for Research in Biomedicine), Adrian Gabriel Torres (IRB Barcelona - Institute for Research in Biomedicine) and Lluís Ribas de Pouplana (IRB Barcelona - Institute for Research in Biomedicine).

Abstract:

Transfer RNAs (tRNAs) are non-coding RNAs that bring amino acids to the ribosome and are thus essential for protein synthesis.

Alterations in tRNAs and in the enzymes responsible for tRNA biogenesis, modification and processing are related to complex diseases such as cancer, diabetes and neurological dysfunctions. To date, their molecular role in disease is currently poorly understood.

High-throughput sequencing of transfer RNAs (tRNA-Seq) is a powerful approach to characterize the cellular tRNA pool. However, the presence of tRNA modifications, the sequence similarity between different tRNAs, and the large number of tRNAs encoded in the genome impair the interpretation of results and downstream analysis of tRNA-Seq datasets. Therefore, processing tRNA-seq datasets require strong bioinformatics skills that are frequently not available in experimental laboratories.

To overcome this challenge, we present tRNAstudio, a user-friendly automated pipeline designed to analyze tRNA-Seq datasets that has been packaged into a graphical user interface (GUI). tRNAstudio GUI can be implemented locally upon running a few simple bash commands. The output obtained includes files with extensive graphical representations and an interactive html report to help interpret the data. Users can extract information on tRNA gene expression, post-transcriptional tRNA modification levels and tRNA processing.

This work brings bioinformatics closer to experimental laboratories and will help in expanding the knowledge on the role of tRNA biology in physiological and pathological scenarios.

Data

Using Statistical and Machine Learning Models to Identify Features Contributing to Class Switch Recombination

Lutecia Servius (King's College London), Joseph Chi-Fung Ng (King's College London), Davide Pigoli (King's College London) and Franca Fraternali (King's College London).

Abstract:

Class Switch Recombination (CSR) is a biological process where antibodies change isotope to adapt their function. The mechanism of CSR is not well understood but high throughput sampling of antibody sequences from human samples offers an opportunity to build data driven models to understand CSR determinants.

This work aims to identify the features of the antibody repertoire that contribute to CSR using models such as logistic regression (LR), random forest (RF), support vector machine (SVM) and the generalised logistic mixed model (GLMM) on a published antibody repertoire dataset of donors challenged by COVID-19, Ebola and Respiratory Syncytial Virus (Stewart et al. 2022 doi:10.3389/fimmu.2022.807104).

We observe a non-random prediction accuracy of CSR likelihood, suggesting that features in the antibody repertoire contain information contributing to CSR. A second experiment was performed by randomly splitting the donors to form training test sets. The inclusion of new donor data in the test set diminished the accuracy of all three models by ~20%. Analysing the RF hyperparameters indicates maximum tree depth, the default for many software packages, results in a model that performs well in identifying donor-specific trends but is not open to generalisation.

A GLMM was trained to detect the universal features that contribute to the probability of CSR. These analyses suggest that features of the antibody repertoire, many related to the variable region, contain signals which predict the likelihood of CSR; these features can be further investigated with other data types (e.g. structural models) to understand their implications on antibody function.

Data

VIROMEdash: Global Virome Sequence Metadata Visualizer

Eyyüb Ünlü (Istanbul University, Turkey) and Mohammad A. Khan (Perdana University, Malaysia / Bezmialem Vakif University, Turkey).

Abstract:

The global viral virome can be defined as the total collection of viruses in nature. It is said that there are more viruses than stars in the universe. However, public sequence databases report only a negligible fraction of the global virome. Nonetheless, over the years, major advances in the "Omics" technologies, including high-throughput sequencing and high-performance computing, have led to the generation and warehousing of large amounts of data and have greatly improved the understanding of viruses. There are several public repositories, ranging from primary, secondary to specialist, that warehouse viral sequence data. The Virus Variation Resource (NCBI Virus) is a newly-designed, secondary viral sequence data resource hosted by the National Center for Biotechnology Information, which provides a comprehensive collection of viral sequence data focusing on integration, standardization and harmonization of metadata across records. This enables systematic interrogation of the database across up to 23 standard metadata dimensions for important insights that provide a better evaluation of the current landscape of the global virome. Herein, we present VIROMEdash, a visualizer for NCBI Virus that mines the existing records and presents insights over five major metadata dimensions (taxonomy class, host-organism, geography, collection date, and Baltimore class). The portal is dynamic and visuals are interactive with options for download. Showcased are mined results of NCBI Virus data, up-to-date as of May 2022 and, VIROMEdash also allows users to upload their own dataset or an accession list from NCBI Nucleotide/Protein/Virus database. VIROMEdash is publicly available at https://viromedash.herokuapp.com/.

Data

WorkflowHub: a FAIR registry for workflows

Carole Goble (The University of Manchester), Finn Bacall (The University of Manchester), Stian Soiland-Reyes (The University of Manchester), Stuart Owen (The University of Manchester), Alan Williams (The University of Manchester), Ignacio Eguinoa (VIB-UGent Center for Plant Systems Biology), Bert Droesbeke (VIB-UGent Center for Plant Systems Biology), Hervé Ménager (Institut Pasteur), Laura Rodríguez Navas (Barcelona Supercomputing Center), José Maria Fernández González (Barcelona Supercomputing Center), Salvador Capella-Gutierrez (Barcelona Supercomputing Center), Michael R. Crusoe (Vrije Universiteit Amsterdam), Björn Grüning (University of Freiburg), Simone Leo (CRS4), Luca Pireddu (CRS4), Johan Gustafsson (Australian BioCommons), Phil Ewels (Seqera Labs) and Frederik Coppens (VIB-UGent Center for Plant Systems Biology).

Abstract:

The WorkflowHub (workflowhub.eu) is a FAIR workflow registry sponsored by the European RI Cluster EOSC-Life and the European Research Infrastructure ELIXIR. It is workflow management system agnostic: workflows may remain in their native repositories in their native forms. As workflows are multi-component objects, including example and test data, they are packaged, registered, downloaded and exchanged as workflow centric Research Objects using the RO-Crate specification, making the Hub an implementation of the FAIR Digital Object principles. A schema.org based Bioschemas profile describes the metadata about a workflow and encouraged use of the Common Workflow Language provides a canonical description of the workflow itself. Workflow management systems such as Galaxy, Nextflow, and snakemake are working with the Hub to seamlessly and automatically support object packaging, registration and exchange. The WorkflowHub provides features such as community spaces, collections, versioning and snapshots. Ensuring credit for the diversity of contributors to a workflow is an important focus. It supports community registry standards and services such as GA4GH TRS and LS Login, and integrates with the LifeMonitor workflow testing service. To date WorkflowHub contains 260 workflows from 98 teams.

Genes

A gene expression signature for survival and risk prediction of Breast Cancer that improves the signatures used in clinical genomic platforms

Santiago Bueno-Fortes (Department of Statistics, University of Salamanca (USAL), Cancer Research Center (IBMCC, CSIC/USAL) and IBSAL), Alberto Berral-Gonzalez (Cancer Research Center (IBMCC, CSIC/USAL), CSIC / University of Salamanca, and IBSAL), Natalia Alonso-Moreda (Department of Statistics, University of Salamanca (USAL), Cancer Research Center (IBMCC, CSIC/USAL) and IBSAL), Jose M Sanchez-Santos (Department of Statistics, University of Salamanca (USAL), Cancer Research Center (IBMCC, CSIC/USAL) and IBSAL), Manuel Martin-Merino (Computer Science School, Universidad Pontificia de Salamanca (UPSA)) and Javier De Las Rivas (Cancer Research Center (IBMCC, CSIC/USAL), CSIC / University of Salamanca, and IBSAL).

Abstract:

Modern genomic technologies allow us to perform genome-wide analyses to find gene markers associated with the risk and survival in cancer patients. Accurate risk prediction and patient stratification based on robust gene signatures is a critical path forward in personalized treatments and precision medicine. Several authors have proposed gene signatures to assign risk in breast cancer patients and some of them have been implemented as commercial platforms in the clinic, such as Oncotype and Prosigna. However, these platforms are black boxes in which the influence of selected survival markers is unclear and the risk scores provided have no clear biological interpretation. In addition, they can not be related to the standard clinicopathological tumor markers obtained by immunohistochemistry (IHC), which guide clinical and therapeutic decisions in breast cancer. Results: We have developed a framework to discover a robust list of gene expression markers associated with survival data that can be biologically interpreted in terms of the 3 main biomolecular factors (ER, PR and HER2) that define clinical outcome in breast cancer. To test and ensure the reproducibility of the results, we first compiled and integrated two independent datasets with a large number of tumor samples (1,024 and 879) that include full genome-wide expression profiles and survival data. Using these two cohorts, we have obtained a robust subset of gene survival markers that correlate well with the major IHC markers used in breast cancer. The geneset of survival markers that we identify (including 34 genes) significantly improves the risk prediction provided by the genesets included in the commercial platforms: Oncotype (16 genes) and Prosigna (49 genes). Furthermore, some of the genes we identified have recently been proposed in the literature as new prognostic markers and may deserve more attention in current clinical trials to improve breast cancer risk prediction.

Genes

A knowledge graph approach for interpretable prediction of pathogenic genetic interactions

Alexandre Renaux (Interuniversity Institute of Bioinformatics in Brussels, Université Libre de Bruxelles-Vrije Universiteit Brussel), Chloé Terwagne (Interuniversity Institute of Bioinformatics in Brussels, Université Libre de Bruxelles), Michael Cochez (Department of Computer Science, Vrije Universiteit Amsterdam), Ilaria Tiddi (Department of Computer Science, Vrije Universiteit Amsterdam), Ann Nowé (Artificial Intelligence Lab, Vrije Universiteit Brussel) and Tom Lenaerts (Interuniversity Institute of Bioinformatics in Brussels, Université Libre de Bruxelles-Vrije Universiteit Brussel).

Abstract:

An increasing number of clinical studies are reporting patterns of oligogenic inheritance in genetic diseases. Despite the advent of methods able to predict the pathogenicity of variant combinations, the underlying biological mechanisms remain unknown since these models offer limited interpretability. To advance towards a better understanding of oligogenic disease aetiology, we developed a new interpretable predictive method based on a knowledge graph. This heterogenous network integrates curated oligogenic combinations together with multiple biological networks and biomedical ontologies. Our approach successfully captures association rules solely based on multi-hop relationships between genes. It combines them as a decision set model which can predict the pathogenicity of new gene pairs. These predictions come with explanations, obtained by querying the knowledge graph, which highlight relevant paths. The benchmarking of this model in a cross-validation setting achieves a ROC AUC of 0.81 and could consistently recall 13/22 independant gene pairs from recently published digenic combinations. The analysis of the rule-based paths highlights relevant contributors to the disease and demonstrates the ability of this approach to generate knowledge-based hypotheses to investigate new disease mechanisms.

Genes

A multi cohort analysis workflow for RNA-seq data

Xinhui Wang (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg), Soumyabrata Ghosh (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg) and Venkata Satagopam (Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg).

Abstract:

A large volume of multi-omics data generated in recent years is publicly available in biorepositories like GEO. While these resources are highly valuable for translational researchers, the variety of data types, used experimental methods, and paucity of clinical annotation makes re-mining of the data a challenging task. To address the challenge in multi-cohort RNA-seq analysis, we built a workflow for analysing publicly available data from the Ulcerative Colitis studies using standard measurement platforms on different cell lines. Here we will present the workflow which started with the automated downloading of selected GEO datasets from blood and skin biopsies. Next, an R pipeline was built to perform the curation, quality check, and merging, followed by batch effect correction, and normalization. Then differential expression analysis was done on each merged dataset followed by enrichment analysis. The overlapping differentially expressed genes and related pathways were reported across different cell lines and tissues. The method was validated by similar findings reported in previously published analyses. In addition, novel features were also identified with this approach

Genes

A New Method for Discovering Drivers Made of Epistatic Gene Pairs in Cancer Tumors

Jairo Rocha (University of the Balearic Islands), Jaume Sastre Tomas (UIB), Victor Asensio-Landa (Instituto de Investigación Sanitaria de las Islas Baleares (Idisba)), Emilia Amengual-Cladera (Hospital Universitario Son Espases- IdISBa), Jessica Hernandez-Rodriguez (Health Research Institute of the Balearic Islands), Damià Heine-Suñer (Health Research Institute of the Balearic Islands) and Emidio Capriotti (University of Bologna).

Abstract:

We describe a new method for epistatic analysis that explicitly finds pairs of genes whose contingency tables show significant values in the cases where both genes are mutated in the tumor tissue but not both are mutated in normal tissue, one of which could have a germline mutation, and the other, a somatic mutation. With this method, both the normal and tumor tissues of pairs of genes are taken into account but neither gene by itself is related to the tumor in a significant way.

In a contingency table for two genes with mutation values "no", "somatic" and "germline", we let fix the cells ("somatic","somatic"), ("germline","somatic") and ("somatic","germline"). This is a new approach as other studies do not use the germline mutations. All other six cells are let free and the minimum dependency statistic is found, so that only the three cells are relevant.

Survival analysis is used to filter down the pairs to clinically related ones. To test the new algorithm, we applied it to the Colon Adenocarcinoma (COAD) and Lung Adenocarcinome (LUAD) subjects available at The Cancer Genome Atlas.

We have found gene pairs whose epistatic mutation appears significant in tumor tissue as opposed to normal tissue in the COAD and LUAD cases.

We believe that the method and, in particular, the gene pairs can be used to look for the possibility of new biological and medical insights into cancer, in general, and colon and lung cancer, in particular.

Genes

A spatial and single-cell transcriptomics approach to investigate scleroderma in human fibroblasts.

Till Baar (University of Cologne), Ann-Helen Rosendahl (University Hospital Cologne), Katrin Schönborn (University Hospital Cologne), Niklas Kleinenkuhnen (University of Cologne), Beate Eckes (University Hospital Cologne), Pia Moinzadeh (University Hospital Cologne), Thomas Krieg (University Hospital Cologne) and Achim Tresch (University of Cologne).

Abstract:

Systemic sclerosis or scleroderma is a rare condition under which fibroblasts secrete excessive amounts of extracellular matrix (ECM), which leads to tissue stiffening. Systemic scleroderma can compromise not only the skin but internal organs, as well, while localised scleroderma, also known as morphea, affects the skin exclusively. Although the underlying cause of scleroderma is unknown and no cure exists, treatment can ease its symptoms and slow its progression.

We performed single-cell RNA sequencing on patient samples and fibroblasts in culture, as well as spatial transcriptomics on scleroderma samples exhibiting distinct areas in different stages of lesion development. This allows us to investigate changing gene expression patterns in distinct histological regions. We used graph-based clustering to partition the cells and spatial autocorrelation to detect gene expression patterns with regional variability.

The integration of spatial transcriptomics and single-cell RNA sequencing data holds the promise to boost our analysis. While spatial transcriptomics preserves the information regarding the histological origin of the cells, its throughput and sequence coverage per spot is limited. In turn, single-cell sequencing cannot resolve cells spatially but offers much higher cell counts. By matching single-cell profiles with the regional expression characteristics, we increase the statistical power to detect expression patterns pertaining to non-physiological, scleroderma-associated processes.

Genes

A state-of-the-art and easy-to-use Python framework for plant phenotype prediction

Florian Haselbeck (TUM Campus Straubing for Biotechnology and Sustainability, Weihenstephan-Triesdorf University of Applied Sciences), Maura John (TUM Campus Straubing for Biotechnology and Sustainability, Weihenstephan-Triesdorf University of Applied Sciences) and Dominik Grimm (TUM Campus Straubing for Biotechnology and Sustainability, Weihenstephan-Triesdorf University of Applied Sciences).

Abstract:

Facing a growing world population and the climate crisis, plant breeding is a key technology for food security by developing more productive and robust crops. In this context, phenotype prediction based on genomic data has the potential to save costs and to accelerate the breeding process.

Currently, no easy-to-use Python API is available to enable the rigorous training, comparison and analysis of phenotype predictions for a variety of different models. For this purpose, we developed an open-source Python framework including multiple state-of-the-art prediction models to enable phenotype prediction in an easy-to-use way. Besides common genomic selection approaches, such as best linear unbiased prediction (BLUP) and models from the Bayesian alphabet, our framework includes eight machine learning methods. These range from classical models, such as regularized linear regression over ensemble learners, e.g. XGBoost, to deep learning-based architectures, such as Convolutional Neural Networks (CNN). To enable automatic hyperparameter optimization, we leverage state-of-the-art and efficient Bayesian optimization techniques. In addition, our framework is designed to allow an easy and straightforward integration of further prediction models. We demonstrated the usefulness of our framework in a case study in the model organism Arabidopsis thaliana, using both a variety of synthetic and real-world phenotypes.

Genes

acorde unravels functionally interpretable networks of isoform co-usage from single cell data

Ángeles Arzalluz-Luque (Institute for Integrative Systems Biology (UV-CSIC)), Sonia Tarazona (Universitat Politècnica de València) and Ana Conesa (Institute for Integrative Systems Biology (UV-CSIC)).

Abstract:

scRNA-seq is a powerful tool to study cell identity and cell fate. However, methods that infer isoform networks at the single cell level are lacking due to the inherent limitations of the data to provide reliable co-expression estimates for transcript variants. This limits our understanding of the contribution of alternative splicing to cell biology. The acorde pipeline successfully addresses this question while substantially contributing to improve single-cell computational analysis. First, our study introduces a sound strategy for isoform definition and quantification that leverages single-cell Illumina and bulk long reads, overcoming the low depth-per-cell of available single-cell long read technologies. Next, we develop and validate percentile correlations, an innovative approach that unlocks co-expression analysis in single-cell data, overcoming noise and sparsity constraints. Subsequent clustering and network-based analysis allows the detection of co-expressed isoform modules and provides a multi-group definition of Differential Isoform Usage (DIU) and co-Usage (coDIU). Our study additionally couples these analyses with extensive isoform-level functional annotation. When applied to a mouse neural single cell dataset, acorde identified functionally related cell-type specific co-expressed isoforms that confer cell-type identity. Furthermore, our percentile correlation strategy has the potential to enable co-expression analysis for other single-cell data modalities and we are currently extending this approach to the integration of multi-modal single-cell data.

Genes

Allele-specific copy-number based deconvolution of bulk tumour RNA sequencing data from the TRACERx study

Carla Castignani (Francis Crick Institute), Jonas Demeulemeester (Department of Human Genetics, KU Leuven, Belgium), Elizabeth Larose Cadieux (Francis Crick Institute), Robert E. Hynds (Cancer Institute, University College London (UCL)), Stefan C. Dentro (European Molecular Biology Laboratory, European Bioinformatics Institute), Tracerx Consortium (Francis Crick Institute), Charles Swanton (Francis Crick Institute) and Peter Van Loo (Francis Crick Institute / The University of Texas MD Anderson Cancer Center).

Abstract:

Analyses of bulk RNA sequencing data are central to most large-scale tumour sequencing studies. Bulk expression data represent population averages, and its interpretation is confounded by both normal cell contamination and copy number alterations. Although several computational methods to deconvolve tumour and normal expression profiles from bulk RNA-seq have been developed in the last years, these often rely on a set of cell-type-specific reference signatures and ignore the effect of copy number changes.

To address these issues, we have developed a method that formalizes the relationship between allele-specific copy number, expression and sample purity to deconvolve the expression profiles of tumour and normal cells from bulk RNA-seq data in an unbiased manner. Our method was applied to matched whole-exome and RNA-seq data produced by the TRACERx consortium, including a total of 414 primary tumor samples from 140 non-small-cell lung cancers patients.

Here, we were able to directly deconvolve a median of ~2,000 genes per sample and indirectly infer expression profiles of ~10,000 genes. The accuracy of the deconvolution was validated using in-silico mixtures of patient-derived tumour and normal cells. Our method revealed a strong and constitutive genome-wide overexpression in cancer cells compared to admixed normal cells. We were also able to improve the molecular subtype classification of tumour samples and identify genes recurrently expressed in tumour or normal cells.

Overall, this tool has potential applications in studies that include matched expression and copy number data and can provide new insights into the functional characterization, taxonomy of cancer and tumor evolution.

Genes

Analysis of differential cellular communication from single cell RNA-seq data with scSeqComm

Giulia Cesaro (University of Padova), Giacomo Baruzzo (University of Padova) and Barbara Di Camillo (University of Padova).

Abstract:

Recent advances in single-cell RNA-sequencing have enabled the investigation of cell-cell communication which plays a crucial role in regulating cellular activities. Differences and dysfunctionalities in cross-talk between cells and cellular responses may be associated with different experimental conditions, such as a pathological state or the exposure to different treatments.

Several computational models have been developed to infer intercellular crosstalk between groups of cells mediated by ligand-receptor interactions. However, only few tools investigate also the downstream intracellular signaling cascade that is activated, along with the triggered cell response, and none of the existing method is designed to identify differences between cellular interactions across different experimental conditions.

Here, we present an extension of scSeqComm (Baruzzo et al., Bioinformatics, 2022), a bioinformatic tool that enables the inference and analysis of differential cellular communication across different experimental conditions from scRNA-seq data. scSeqComm can identify both intercellular and intracellular signaling that are altered between the same groups of cells across different conditions, and provides a functional characterization of the downstream differential cellular response, by means of transcriptional regulation of target genes. The tool also provides a rich data visualization, including interactive plots and support for time-series data.

To appreciate the various faced of the method, we applied scSeqComm to a large (~250K nuclei) publicly available single-nucleus dataset of amyotrophic lateral sclerosis and pathological normal donors. We identified a deficit of interactions involving inhibitory and excitatory neurons with respect to the control, suggesting a dysfunctionality in cellular communication as response to neurodegeneration.

Genes

Analysis of transposable elements in R and Bioconductor with atena

Beatriz Calvo-Serra (Department of Medicine and Life Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain) and Robert Castelo (Department of Medicine and Life Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain).

Abstract:

Transposable elements (TEs) are DNA sequences that can mobilize within the genome either through a DNA or an RNA intermediate. Their insertions have resulted in a complex distribution of repeated elements occupying approximately half of the human genome. These elements, particularly endogenous retroviruses (ERVs), participate in physiological processes and have been involved in the development of some human diseases. They may exert their function through transcription, hence RNA sequencing can be used to detect their expression. However, due to their repetitive nature, reads sequenced from TE RNA transcripts usually map to multiple genomic loci (i.e. multi-mapping reads) and are consequently discarded in standard RNA sequencing data processing pipelines. For this reason, TE analysis software exists, such as ERVmap, Telescope and TEtranscripts. These software packages, developed outside the R and Bioconductor ecosystem, do interact with it for the purpose of differential expression analyses. To facilitate expression quantification of TEs and its integration with other Bioconductor software, we have developed atena (https://bioconductor.org/packages/atena), an open source software package for the analysis of TE expression in R available at Bioconductor. The atena package is a faster and accurate implementation of these three methods, with a quick, flexible and straightforward access and processing of RepeatMasker UCSC TE annotations. In summary, atena facilitates the integration of TE annotation and expression quantification with a wide range of differential expression and functional analyses pipelines available in Bioconductor.

Genes

APPRIS: selecting functionally important isoforms

Daniel Cerdán-Vélez (Spanish National Cancer Research Center (CNIO)), Jose Manuel Rodríguez (CNIC), Fernando Pozo (Spanish National Cancer Research Center (CNIO)), Tomás Di Domenico (Spanish National Cancer Research Center (CNIO)), Jesús Vázquez (CNIC) and Michael Tress (Spanish National Cancer Research Center (CNIO)).

Abstract:

The APPRIS database houses annotations for protein isoforms for a range of species. APPRIS selects principal isoforms for coding genes based on the preservation of protein structure and function, and on cross-species conservation. We have shown that most coding genes produce a single main protein isoform and that APPRIS principal isoforms agree with the main cellular isoform over more than 95% of genes.

Human genetic variation shows that exons that produce APPRIS principal isoforms are under purifying selection, while alternative exons are generally evolving neutrally. In addition, the distribution of clinical variants strongly supports the biological relevance of APPRIS principal isoforms. Recent research shows that more than 99.85% of validated PubMed-supported pathogenic mutations map to APPRIS principal transcripts.

APPRIS annotations now cover 11 model organisms, with the addition of three new species, Rhesus monkey, Cow and Chicken, and we have incorporated reliable models from the EMBL-EBI AlphaFold database into the protein structural information used by APPRIS.

However, the most significant change in APPRIS is in the way the database determines principal isoforms. We have incorporated a new method to determine functional importance for individual isoforms, TRIFID. TRIFID can separate alternative exons that are under purifying selection from those that are not. TRIFID provides a functional annotation score for every annotated protein isoform in APPRIS, and APPRIS uses this score to choose principal isoforms when APPRIS core methods are unable to make a decision.

Genes

BamQuery: A new proteogenomic tool to explore the Immunopeptidome

Maria Virginia Ruiz Cuevas (Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal), Marie-Pierre Hardy (Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal), Anca Apavaloaei (Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal), Sebastien Lemieux (Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal), Claude Perreault (Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal) and Gregory Ehx (Laboratory of Hematology, GIGA-I3, University of Liege and CHU of Liege, Liege, Belgium).

Abstract:

The immunopeptidome is the repertoire of MHC-I-associated peptides (MAPs) that define the immune self for CD8+ T cells. Recently, advances in proteogenomics have revolutionized immunopeptidomic studies by demonstrating that non-coding genomic regions, such as endogenous retroelements (EREs), generate significant numbers of MAPs and by using transcriptome-informed mass spectrometry to identify tumor-associated MAPs. Therefore, annotating MAPs biological properties as well as quantifying their RNA expression in malignant and benign cells has become a frequent challenge for immunologists. Here, we introduce BamQuery, a computational tool that counts RNA-seq reads from all genomic regions capable of encoding given MAPs from a large number of user-provided aligned RNA-seq reads, as well as annotates their genomic-based features. By evaluating the expression of previously reported MAPs, BamQuery showed that most canonical-reported MAPs derive from a single (coding) genomic region whereas most ERE-reported MAPs can be generated by numerous (average=861) distinct transcripts. In addition, many ERE MAPs could derive from highly expressed regions other than those previously annotated, including canonical regions, resulting in an ambiguous classification of their biotype. Similarly, we show that published tumor-specific antigens presented as mutated neoantigens and proteasomal spliced-peptides, can in fact be generated by other non-mutated, non-coding, non-spliced, highly expressed transcripts in normal tissues. We also present BamQuery as a powerful tool to evaluate the sharing of tumor antigens' source RNAs among cancer patients. Considering these observations, we conclude that BamQuery is convenient for providing immunopeptidome quantification and biological annotation in any proteogenomic study and thus in any tumor antigens identification pipeline.

Genes

Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data

Sini Junttila (University of Turku), Johannes Smolander (University of Turku) and Laura Elo (University of Turku).

Abstract:

Single-cell RNA-sequencing (scRNA-seq) enables the quantification of gene expression from thousands of cells simultaneously and identification of transcriptomic changes between the cells. scRNA-seq datasets increasingly include multi-subject, multi-condition experiments to study cell-type-specific differential states (DS) between conditions. This can be performed by first annotating all the cells in the dataset and then by performing a DS analysis between the conditions within each cell type. Naïve single-cell DS analysis methods that treat cells statistically independent and do not model the subjects in any way are subject to false positives in the presence of variation between biological replicates, an issue known as the pseudo-replicate bias. While several methods have already been introduced to carry out the statistical testing in multi-subject scRNA-seq analysis, comparisons that include all available method types are currently lacking. We have performed a comprehensive comparison of 18 methods for the identification of DS changes between conditions from multi-subject scRNA-seq data. Our results suggest that pseudo-bulk methods performed generally best. Both pseudo-bulks and mixed models that model the subjects as a random effect were superior compared with naive single-cell methods. While the naive models achieved higher sensitivity than the pseudo-bulk methods and the mixed models, they exhibited a high number of false positives. In addition, accounting for subjects through latent variable modeling did not improve the performance of the naive methods.

Genes

Beyondcell 2.0: dissecting therapeutic heterogeneity in spatial single-cell RNA-seq data

María José Jiménez-Santos (Centro Nacional de Investigaciones Oncológicas (CNIO)), Coral Fustero-Torre (Centro Nacional de Investigaciones Oncológicas (CNIO)), Santiago García-Martín (Centro Nacional de Investigaciones Oncológicas (CNIO)), Carlos Carretero-Puche (Centro Nacional de Investigaciones Oncológicas (CNIO)), Luis García-Jimeno (Centro Nacional de Investigaciones Oncológicas (CNIO)), Tomás Di Domenico (Centro Nacional de Investigaciones Oncológicas (CNIO)), Gonzalo Gómez-López (Centro Nacional de Investigaciones Oncológicas (CNIO)) and Fátima Al-Shahrour (Centro Nacional de Investigaciones Oncológicas (CNIO)).

Abstract:

Intratumor heterogeneity (ITH) is characterized by the presence of different cell types within the same tumor, including different cancer subpopulations and the tumor microenvironment (TME). Single-cell RNA-seq (scRNA-seq) is one of the most popular strategies to study ITH at fine resolution because it makes possible to transcriptionally characterize each cell in the tumor. Moreover, spatial scRNA-seq is a cutting-edge technology that makes morphological and context data available, allowing to visualize the cells in the tissue and identify cell-to-cell interactions. Beyondcell [1] (https://github.com/cnio-bu/beyondcell) is an R package for the identification of different drug vulnerabilities in scRNA-seq which enables to study ITH by clustering cells in therapeutic clusters (TCs) according to their predicted response to a set of drug signatures of interest. Here we present Beyondcell 2.0, the latest version of this software that includes new functionalities like the ability to analyze spatial scRNA-seq data for dissecting the tumor and TME therapeutic heterogeneity and study their spatial distribution and interactions. As a demonstration, we characterized the tumor and TME subpopulations in different spatial datasets based on their gene expression and the pathologist annotations. Furthermore, we computed the TCs using Beyondcell and identified different TME subpopulations with distinct inferred drug vulnerabilities that could be explained by their location in the tissue and cell-to-cell interactions with cancer cells.

1. Fustero-Torre C, Jiménez-Santos MJ, García-Martín S, Carretero-Puche C, García-Jimeno L, Ivanchuk V, Di Domenico T, Gómez-López G, Al-Shahrour F. Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-seq data. Genome Med. 2021 Dec 16;13(1):187. doi: 10.1186/s13073-021-01001-x.

Genes

Cancer signatures for reproducible gene expression analysis data: the computational way to achieve precision medicine

Stefania Pirrotta (University of Padova), Laura Masatti (University of Padova), Fabiola Pedrini (University of Padova), Chiara Romualdi (University of Padova) and Enrica Calura (University of Padova).

Abstract:

Cancer is a complex disease, characterized by genomic aberrations with an impact on gene regulation and cell processes. Many studies proposed gene expression signatures as a valuable tool for understanding cancer mechanisms, defining subtypes and showing cancer activities while they are happening. However, a computational implementation for several signatures proposed in literature is missing. That would provide detailed signature definition and assure reproducibility, dissemination and usability. To achieve this, we developed signifinder, an R package that collects and implements a compendium of gene expression signatures in cancer, with the aim to ensure an easy and reproducible computation.

A list of signatures covering numerous cancer hallmarks was collected from the literature under the selection of stringent criteria. We collected and stored the essential information for the signature computation (the list of genes with eventually their corresponding coefficients or attributes) inside the signifinder package and implemented a dedicated function for each signature.

Signifinder implements more than 40 signatures from cancer literature. It attributes single-sample scores per signature covering many different tumor aspects, such as predicting the response to therapy or the survival association, as well as quantifying multiple microenvironmental conditions, such as hypoxia or the immune activity.

For the first time multiple cancer signatures can be easily and efficiently calculated in different cancer types. With the promise of tailored predictions for individual patients, signatures collected within signifinder can help in automatically investigating tumor samples in a more easy and reproducible way, leading us a step closer to precision medicine.

Genes

Cell type specificity of Hi-C interaction network topology and gene expression patterns

Juris Viksna (Institute of Mathematics and Computer Science, University of Latvia), Lelde Lace (Institute of Mathematics and Computer Science, University of Latvia), Gatis Melkus (Institute of Mathematics and Computer Science, University of Latvia), Peteris Rucevskis (Institute of Mathematics and Computer Science, University of Latvia), Edgars Celms (Institute of Mathematics and Computer Science, University of Latvia), Sandra Silina (Institute of Mathematics and Computer Science, University of Latvia) and Andrejs Sizovs (Institute of Mathematics and Computer Science, University of Latvia).

Abstract:

Chromatin conformation capture (Hi-C) is rapidly developing class of technologies that are capable to provide insights about chromatin spatial organisation. High quality datasets of Hi-C data are increasingly becoming available, including studies that cover chromatin interactions collected by uniform protocol across different cell types. Previously we have studied variation of topologies of chromatin interaction networks obtained from publicly available PCHi-C dataset covering 17 human primary haematopoietic cell types and proposed a set of 6 well-defined topological features (Base6 metrics) that provide good discriminatory power between cell types and could be assigned plausible biological explanations for their significance.

The discriminatory power of these metrics, however, was assessed against cell type distances that were related to PCHi-C technology itself. In this work we analyse the relations of topological features of chromatin interaction networks and distances between cell types defined by proximity measures based on gene expression data. The study confirms that network topological features well correlate with cell type distances based on gene expression, and, at the same time, there are notable variations depending on expression measurement units (FPKM or TPM) and the specific proximity measure (with recently proposed Transcriptomic Signature Distance overall providing the best results). The largest discriminatory power remains provided by the same Base6 metrics that we have proposed earlier. Not surprisingly stronger correlations can be obtained using distances based only on expression of genes that are directly included in PCHi-C data, nevertheless they remain significant also for subsets of genes that are completely absent in interaction networks.

Genes

Cell type-specific gene co-expression modules and expression signatures define tumor heterogeneity in melanoma patients

Lars Bosshard (ETH Zürich), Franziska Singer (ETH Zürich) and Michael Prummer (ETH Zürich).

Abstract:

Gene co-expression networks are governing cellular processes in health and disease. However, the presence or absence of correlated gene pairs is difficult to interpret in bulk samples. The co-occurrence of two cell types can lead to an apparent co-expression of two genes even when they are completely independent within each individual cell. In bulk, the observation of positively or negatively correlated expression of a pair of genes requires low within sample variability. In contrast, in single cell experiments, the observation of positively or negatively correlated expression of a pair of genes requires low within cell variability, and high between cell variability. In addition, in single cell experiments, an observed correlation between a pair of genes is truly present within one cell.

We used single cell transcriptomics to discover disease-specific co-expression networks in different cell types from tissue biopsies of melanoma patients. We analyze each sample independently to arrive at patient-specific networks and subsequently compare them across the cohort. Co-expression sub-networks are identified in each patient using community detection principles. Recurring as well as unique co-expression modules are compared to gene ontology terms to assign a biologically meaningful label. Any meaningful difference of the disease and cell type-specific module composition from common gene sets may give new insight into disease causing mechanisms or novel treatment options. Many of the curated gene sets used for enrichment analysis were derived from bulk samples of healthy individuals. Moreover, patient-specific gene expression programs in various cell types may give rise to personalized treatment recommendations.

Genes

Comparative study of dynamic changes in gene expression profiles induced by PPARa ligands

Yayoi Natsume-Kitatani (National Institutes of Biomedical Innovation, Health and Nutrition), Ken-Ichi Aisaki (National Institute of Health Sciences), Satoshi Kitajima (National Institute of Health Sciences) and Jun Kanno (National Institute of Health Sciences).

Abstract:

The Percellome database [1], which collects gene expression profiles (microarray data) induced by toxic chemicals, is a useful resource for elucidating the mechanisms of chemical-induced toxicity from the gene expression cascade. The database contains estimated mRNA copy number per cell for each exposure dose and exposure time in mice, allowing us to extract what kind of molecular network changes are associated with chemical exposure from the dynamic changes in gene expression levels. We compared the dynamic variation in gene expression profiles of the known or inferred PPAR ligands (clofibrate, valproic acid, and estragole) and detected patterns common to and unique to these three chemicals. The results showed that PPARa target genes were commonly up-regulated at 2, 4, and 8 hours, that there was no significant overlap in the list of genes up-regulated by the three chemicals, and that among the up-regulated genes, those with common functions detected by enrichment analysis were the most commonly up-regulated among the three chemicals. The results also suggest that classification based on patterns of gene expression variation and in-depth domain knowledge are required to detect secondarily elicited biological responses, while induction of gene expression via nuclear receptors, such as PPARa activation, is relatively easy to detect. The results suggest the need for an ontology for toxicology, different from pathways and existing gene ontology, in inferring the mechanism of toxicity from gene expression profiles.

Genes

Comparison of Deconvolution Techniques for Spatially Resolved Transcriptomics Data

Carolin Walter (Westfälische Wilhelms-Universität Münster), Sarah Sandmann (Westfälische Wilhelms-Universität Münster), Ada Anike Pohlmann (Westfälische Wilhelms-Universität Münster), Alina Sophie Disch (Westfälische Wilhelms-Universität Münster), Kornelius Kerl (Westfälische Wilhelms-Universität Münster) and Julian Varghese (Institute of Medical Informatics).

Abstract:

Spatially Resolved Transcriptomics (ST) is a recent molecular profiling technique that allows new insights into the spatial composition of transcriptome data, and thus offers new opportunities for the study of tumor heterogeneity and the composition of tumor microenvironment data (TME) [1]. In contrast to regular single-cell RNA-seq (scRNA) data, current ST techniques typically provide a multi-cell resolution, where the input of different cells is mixed for small spatial areas, or “spots”. Consequently, computational methods are needed to approximate the expression and type of the involved cells per ST spot. While algorithms like STdeconvolve and SpatialGE offer functions for spot-wise ST deconvolution, the Seurat pipeline uses an integration approach to predict the cell composition of ST spot data.

We present a comparison of computational deconvolution and integration algorithms for spatial transcriptomics data on a set of eight original human 10X Visium retinoblastoma samples, and published murine and human 10X Visium spatial transcriptomics data. For each original sample, a matching single-cell RNA-seq (scRNA) retinoblastoma sample with detailed cluster annotations was available for comparison purposes and as computational reference dataset, if applicable. We evaluated the effects of different algorithm parameters on the deconvolution results, assessed the proportions of retinoblastoma and TME cell types in the ST and scRNA data, and tested the chosen algorithms' run time and general usability.

[1] Ahmed R et al. Single-Cell RNA Sequencing with Spatial Transcriptomics of Cancer Tissues. Int J Mol Sci. 2022 Mar 11;23(6):3042.

Genes

Computational analysis of differential splicing and transcript alternations in severe COVID-19 infection

Sunanda Biswas Mukherjee (Faculty of Medicine, Bar-Ilan University) and Milana Frenkel-Morgenstern (Faculty of Medicine, Bar-Ilan University).

Abstract:

Viral infections could modulate the widespread alternations of cellular splicing, which favor the viruses replicating within the host cells by overcoming host immune responses. How the SARS-CoV-2 induces the host cell differential splicing, and the landscape of transcript alternation in severe COVID-19 infection remains elusive. Understanding the differential splicing and transcript alternations in severe COVID-19 infection could improve the molecular insights into the SARS-CoV-2 pathogenesis. In this study, we analyzed the publicly available blood transcriptome data of severe COVID-19 patients, recovered COVID-19 patients at 12-, 16-, and 24-weeks post-infection, and healthy controls. We identified 1385 genes that undergo significant transcript isoform switching events. Among these, 414 genes were found to be differentially expressed in gene expression analysis, while 971 genes are not undergoing any changes in expression levels, but they altered at the transcripts level. Altered transcripts show the significant loss of the open reading frame (ORF), functional domains, and changed the coding to the non-coding transcript, impacting normal cellular functions. We identified the expression of several novel recurrent chimeric transcripts in the samples from severe COVID-19 patients. Further, analysis of the isoform switching in recovered COVID-19 patients highlights that there is no significant isoform switching in 16 -, and 24-weeks post-infection, and the expressed chimeric transcripts are also less. This finding emphasizes that SARS-CoV-2 severe infection could induce widespread splicing in the host cells, which could help the viruses alter the host immune responses that might facilitate the viruses to replicate within the host and translate viral proteins efficiently.

Genes

Computational Mapping of the Human-SARS-CoV-2 Protein-RNA Interactome

Marc Horlacher (Helmholtz Center Munich), Svitlana Oleshko (Helmholtz Center Munich), Yue Hu (Helmholtz Center Munich), Giulia Cantini (Helmholtz Center Munich), Patrick Schinke (Helmholtz Center Munich), Mahsa Ghanbari (Max Delbruck Center for Molecular Medicine), Ernesto Elorduy Vergara (Helmholtz Center Munich), Florian Bittner (Knowing01 GmbH), Nikola Mueller (Knowing01 GmbH), Uwe Ohler (Max Delbruck Center for Molecular Medicine), Lambert Moyon (Helmholtz Center Munich) and Annalisa Marsico (Helmholtz Center Munich).

Abstract:

RNA-binding proteins (RBPs) are critical host factors for viral infection. Here, we investigated the role of human RBPs in the context of SARS-CoV-2 infection by constructing an in silico binding map of human RBPs to the SARS-CoV-2 RNA at nucleotide-resolution using two deep learning methods, Pysster and DeepRiPe, for more than 100 human RBPs.

Model interrogation via integrated gradients revealed that high-scoring positions coincide with known binding motifs of RBPs, suggesting that our predictions represent bona fide binding sites. Using this binding map, we investigated shared patterns of binding and highlighted relationships between functionally relevant RBPs and local genomic features.

Furthermore, the deep learning models enabled the fast interrogation of mutational impact on RBP binding, which is of high relevance for monitoring the emergence of specific viral lineages during the pandemic. We scored the impact of variants from 11 viral strains on protein-RNA interaction, thus identifying a large set of gain-and loss of binding events that could contribute to changes in fitness of these strains. We expanded this analysis by predicting the impact of hypothetical variants via systematic in silico mutagenesis of the SARS-CoV-2 reference genome.

Additionally, by extending our analysis to other human coronaviruses, we identified conserved and differential binding between SARS-CoV-1, SARS-CoV-2 and MERS. Lastly, we linked RBPs to OMICs and patient data from other studies, and identified MBNL1 and FTO as potential biomarkers. Our results contribute towards a deeper understanding of how viruses hijack host cellular pathways and open new avenues for therapeutic intervention.

Genes

Computational method prioritising gene-specific cis-regulatory elements reveals new aspects of transcriptional regulation in B cells

Amber Emmett (University of Leeds), Matthew Care (University of Leeds), Amel Saadi (University of Leeds), Gina Doody (University of Leeds), Reuben Tooze (University of Leeds) and David Westhead (University of Leeds).

Abstract:

Cellular differentiation is guided by epigenetic, and transcriptional reprogramming, orchestrated through interactions between transcription factors and cis-regulatory elements. Here we present an integrative 'omics approach designed to identify and prioritise gene-specific cis-Regulatory Elements Across Differentiation (cisREAD.) We show how our two-pronged approach of community detection and LASSO regression predicts co-acting candidate regulatory elements, which regulate transcription of key differentiation genes, through the activity of lineage-specifying transcription factors. We demonstrate the application of our method to time-course ATAC-seq and RNA-seq datasets to identify thousands of gene-specific regulatory elements driving transcription along the B cell lineage. Leveraging predicted regulatory interactions we link transcription factor occupancy to cell-specific regulatory clusters and gene co-expression modules, revealing novel aspects of transcriptional regulation in B cell activation and plasma cell differentiation.

Genes

consICA: a package for reference-free deconvolution reveals oncogenic processes in "omics" data and highlights the benefits of multimodal data analysis for patient stratification

Maryna Chepeleva (Luxembourg Institute of Health & Belarusian State University), Tony Kaoma (Luxembourg Institute of Health), Aliaksandra Kakoichankava (University of Luxembourg & Luxembourg Institute of Health), Yue Zhang (Luxembourg Institute of Health), Reka Toth (Luxembourg Institute of Health) and Petr Nazarov (Luxembourg Institute of Health).

Abstract:

In cancer research, the analysis of patient-derived molecular profiles, such as gene expression or DNA methylation, is not straightforward because of the tumor heterogeneity that masks important molecular signatures. This can be caused by natural variability in cell type proportions or by clonal variation between species of tumor cells. Additionally, technical biases between experimental platforms may limit the direct comparison of patient data to large public datasets.

We previously presented a reference-free deconvolution algorithm based on consensus independent component analysis (consICA). It allows reproducible mapping of the small clinical data into a space of biologically relevant components defined by a large reference dataset, simultaneously correcting for technical biases. The method was tested on cell line data and showed its potential in extracting realistic biological signals and predicting tumor phenotypes. Importantly, the resulting weights of independent signals (loadings) may have diagnostic or prognostic power.

We applied consICA to 33 tumor cohorts presented in the TCGA dataset and established recommendations on the use of specific data modalities for better prediction of patient survival: mRNA, microRNA and DNA methylation data were considered. More specifically, mRNA should be used for SKCM, KIRC, PAAD cohorts, while DNA methylation for LGG/GBM, KIRP, SARC patients. BRCA and UCEC cohorts strongly benefit from the combination of transcriptomic and methylation data. We also built a connectivity network uniting clusters of signals specific to cancer hallmarks (proliferation, invasion, angiogenesis, inflammation, metabolism deregulation).

The package is available on GitHub (biomod-lih/consICA) and it is undergoing Bioconductor review.

Genes

Context-Aware Transcript Discovery and Quantification from Long Read RNA-Seq data with Bambu

Ying Chen (Genome Institute of Singapore), Andre Sim (Genome Institute of Singapore), Yuk Kei Wan (Genome Institute of Singapore), Keith Yeo (Genome Institute of Singapore), Michael Love (Department of Genetics, University of North Carolina-Chapel Hill) and Jonathan Goeke (Genome Institute of Singapore).

Abstract:

More and more RNA seq samples and conditions are being generated now with the long read RNA sequencing technology. However, being able to accurately discover novel transcripts from these RNA-Seq data while minimizing the impact of non-expressed isoforms on transcript abundance quantification has remained a challenge. To address this problem, we developed Bambu, an R package that performs context specific transcript discovery and quantification, allowing multi-sample processing with one command. Bambu has the following two modules: 1) bambu employs a machine learning model using an eXtreme gradient boosting framework to identify the most confident novel transcript models passing a novel discovery rate (NDR) threshold; 2) bambu estimates the abundance of transcript models with an expectation-maximization algorithm while retaining the full-length and unique estimates to provide inference for active transcripts. With these two features together, Bambu not only demonstrated improved performance in transcript discovery but also more accurate quantification estimates with context specific transcripts over other contemporary methods. Bambu is implemented in R, enabling simple, fast, and accurate analysis of long read transcriptome profiling data.

Availability:

http://bioconductor.org/packages/bambu/

https://github.com/GoekeLab/bambu

Genes

Curare and GenExVis: A toolkit for analyzing and visualizing RNA-Seq data

Patrick Blumenkamp (Justus Liebig University Giessen), Raphael Müller (Justus Liebig University Giessen) and Alexander Goesmann (Justus Liebig University Giessen).

Abstract:

The yearly increasing citations of DESeq2, edgeR, and limma show that differential gene expression (DGE) analyses are still on an emerging path. The vast amount of data generated by current sequencing platforms underpins the need for automated and reproducible analysis pipelines.

Thus, we introduce a simple-to-use toolkit for analyzing and visualizing RNA-Seq data with a focus on DGE analyses. For the processing of high-throughput transcriptomics data, we developed the customizable and reproducible analysis pipeline for RNA-Seq Experiments (Curare). Curare divides a typical analysis into preprocessing, quality control, mapping, and analysis and offers multiple options for each of these steps. Due to the usage of Snakemake internally, it is built for high-throughput analyses and is fully reproducible and parallelizable. For a fast and straightforward exploration and visualization of DGE results, we developed the gene expression visualizer GenExVis. It can display various charts and tables from simple count tables and DESeq2 result files without the necessity of uploading any data or installing any packages.

Both tools combined create an environment that supports the entire process of data analysis, from the initial handling of RNA-seq raw data to the final DGE analyses and result visualization.

Genes

De novo gene evolution in bacteria: a case study of taxonomically restricted genes in Bacillus

Wojciech Karlowski (Adam Mickiewicz University), Deepti Varshney (Adam Mickiewicz University) and Andrzej Zielezinski (Adam Mickiewicz University).

Abstract:

Taxonomically restricted genes (TRGs) are unique for a defined group of organisms and may act as potential genetic determinants of their unique lineage-specific, biological properties. One of the most puzzling questions associated with TRGs concerns their origin. We explore the TRGs of highly diverse and economically important Bacillus bacteria by examining commonly used TRG identification methods, parameters and data sources. We show the significant effects of sequence similarity thresholds, composition, and the size of the reference database in the identification process. We applied stringent TRG search parameters and expanded the identification procedure by incorporating an analysis of noncoding and nonsyntenic regions of non-Bacillus (outgroup) genomes. A multiplex annotation pipeline minimized the number of false positive TRG predictions and showed nearly one-third of the alleged TRGs could be mapped to genes missed by genome annotations pipelines. We traced the putative origin of TRGs by identifying homologous, noncoding genomic regions in non-Bacillus species and detected sequence changes that could transform these regions into protein-coding genes. The results indicate that TRGs represent an intermediate state between genes that are conserved across multiple taxa and nonannotated peptides encoded by open-reading frames.

Genes

decoupleR: ensemble of computational methods to infer biological activities from omics data

Pau Badia i Mompel (Heidelberg University, Faculty of Medicine and Heidelberg University Hospital, Institute of Computational Biomedicine;) and Julio Saez-Rodriguez (Heidelberg University, Faculty of Medicine and Heidelberg University Hospital, Institute of Computational Biomedicine;).

Abstract:

Many methods allow us to extract biological activities from omics data using information from prior knowledge resources, reducing the dimensionality for increased statistical power and better interpretability. Here, we present decoupleR, a Bioconductor and Python package containing computational methods to extract these activities within a unified framework. decoupleR allows us to flexibly run any method with a given resource, including methods that leverage mode of regulation and weights of interactions, which are not present in other frameworks. Moreover, it leverages OmniPath, a meta-resource comprising over 100 databases of prior knowledge. Using decoupleR, we evaluated the performance of methods on transcriptomic and phospho-proteomic perturbation experiments. Our findings suggest that simple linear models and the consensus score across top methods perform better than other methods at predicting perturbed regulators.

decoupleR's open-source code is available in Bioconductor (https://www.bioconductor.org/packages/release/bioc/html/decoupleR.html) for R and in GitHub (https://github.com/saezlab/decoupler-py) for Python. The code to reproduce the results is in GitHub (https://github.com/saezlab/decoupleR_manuscript) and the data in Zenodo (https://zenodo.org/record/5645208).

Genes

Diana - ROOT: A categorization tool related to the Region Of Origin biotype.

Nikos Perdikopanis (PASTEUR INSTITUTE /UNIVERSITY OF THESALY), Georgios Georgakilas (University of Thessaly) and Artemis Hatzigeorgiou (University of Thessaly).

Abstract:

DIANA - ROOT (Region Of Origin bioType) is a genomic intervals' categorization tool relating to the biotype of the genomic neighborhood. It can be applied for microRNAs and arbitrary sets of intervals provided by the user.

Genomic position plays a critical role for microRNA transcription regulation and function. Various studies have shown a strong regulatory relationship between microRNAs (or small RNAs in general) and host or nearby genes. Intergenic microRNAs commonly are produced by independent transcriptional units, while intragenic may be transcribed along with host gene.

DIANA-ROOT, automates the process of cross-referencing genomic intervals with annotation information. It provides graphical and analytical information. A hierarchical tree structure, that is based on the biotype of overlapping genes and the exonic/intronic/antisense/intergenic localization, is constructed. For each defined category, it creates files with details for member-intervals, ready for further processing. A great number of species (Mus musculus, Drosophila melanogaster, Bos taurus, Caenorhabditis elegans, Danio rerio, gallus gallus, Monodelphis domestica, Macaca mulatta, Rattus norvegicus, Tetraodon nigroviridis and Xenopus tropicalis.) are currently supported. New species can be rapidly enabled through a fully automated backend mechanism.

We believe that DIANA-ROOT will be a valuable resource that can significantly accelerate the day-to-day research tasks in laboratories that produce voluminous Next Generation Sequencing data.

Genes

Differential Abundance analysis of microbiome data: which tool and how to choose it?

Marco Cappellato (Department of Information Engineering, University of Padova, Padova, Italy), Giacomo Baruzzo (Department of Information Engineering, University of Padova, Padova, Italy) and Barbara Di Camillo (Department of Information Engineering, University of Padova, Padova, Italy).

Abstract:

The microbiome comprises all of the genetic material within a microbiota (microorganisms that inhabit an ecologiacal niche). Efficient and cost-effective high throughput DNA sequencing techniques has enhanced the study of these complex microbial systems, leading to important conclusions in different fields. Differential abundance (DA) analysis finds a microbial signature looking at differences in taxa abundances between classes of subjects or samples. Several bioinformatics methods have been specifically developed for microbiome data. However, there's no consensus about the best approach.

In this work we develop a novel simulation framework exploiting metaSPARSim, a microbial sequencing count data simulator. We simulate DA features between experimental groups with great effort to resemble real data characteristic, e.g. compositionality, sparsity and taxa intensity-variability relationship, simulating DA taxa in term of differences in absolute and relative abundance. We perform an extensive comparison of 12 recently developed and established DA methods on a common benchmark. The performance overview includes scenarios and covariates not yet investigated by previous studies such as the combined effect of sample size, percentage of DA taxa, sequencing depth, fold change, variability of taxa, low abundance DA taxa, normalisation and different ecological niches.

Our work aims to provide researchers recommendations on how to properly conduct DA analysis in their own datasets. Mainly, we find that methods show a good control of the type I error and, generally, also of the false discovery rate at high sample size, while recall seem to depend on the dataset and sample size.

Genes

Disentangling Cellular Heterogeneity with Multimodal single-cell Integration

Jules Samaran (Computational Systems Biology Team, IBENS, CNRS, INSERM, ENS, Université PSL, Sorbonne Université, Collège Doctoral), Gabriel Peyré (Département de Mathématiques et Applications de l'ENS, CNRS, ENS, Université PSL) and Laura Cantini (Computational Systems Biology Team, IBENS, CNRS, INSERM, ENS, Université PSL).

Abstract:

Single-cell transcriptomics has revolutionized biology and medicine by unraveling the diversity of the cells constituting human tissues. The single-cell technological development has now shifted to the measurement of other modalities (e.g. chromatin accessibility, proteomics). Integrating these complementary sources of information is expected to provide a more comprehensive view of the cell's regulatory states and explain how these regulatory states contribute to different cellular phenotypes. We here propose a method designed to map groups of cells profiled using different single-cell sequencing technologies to a shared low dimensional latent space. By using one Variational Autoencoder (VAE) on each modality, we compute low-dimensional and biologically meaningful embeddings. In order to obtain a space where the proximity between groups of cells is only related to their biological similarity, we enforce a constraint derived from optimal transport (OT) which forces samples from different modalities to mix in the latent space. The main advantages of our approach are (i) its modularity: we propose a framework in which any VAE-based model can be plugged for each modality; (ii) its robustness: using the unbalanced relaxation of the original OT problem our method allows cell subpopulations present only in one of the modalities to remain separated from the rest of the cells; (iii) its scalability: we leverage the power of GPUs to speed up computations and avoid memory limitations by processing the data sets per batches.

Genes

Dissecting the impact of genetic variants on transcriptional and post-transcriptional gene regulation

Anneke Brümmer (University of Lausanne) and Sven Bergmann (University of Lausanne).

Abstract:

Gene expression is regulated at the transcriptional and post-transcriptional level. Typically, expression quantitative trait loci (eQTL) studies only consider exon expression levels, which are often dominated by transcriptional effects, and ignore intron expression levels, despite their information on RNA metabolism.

To better understand the genetic influences on different gene regulatory processes, we analysed QTLs of exon (exQTL) and intron (inQTL) expression levels and their ratio (ex-inQTL) in lymphoblastoid cell lines (n=901). Overall, we detected 10721 genes with QTLs, of which 57% showed shared effects between QTL types. While exQTLs localized more often upstream of genes than other cis-QTLs, ex-inQTLs were mostly within gene bodies. exQTLs, especially when shared with inQTLs, had the largest overlap with sites regulating transcription. In contrast, exQTLs shared with ex-inQTLs overlapped 3'UTRs and QTLs for alternative splicing and polyadenylation. Analyzing the trans-effects of cis-QTLs for transcription factors (TFs) or RNA-binding proteins (RBPs) we found that RBPs had fewer trans-exQTL and more trans-ex-inQTL associations than TFs, and RBPs' trans-effects were lower on exon than on intron levels and their ratio. On average, adding inQTLs and ex-inQTLs almost doubled the number of GWAS trait variants colocalizing with cis-QTLs, suggesting similar functional relevance as for exQTLs. Furthermore, cis-QTLs of all types with strongest effects tended to be more prevalent among common disease cases, indicating potential disease relevance.

Altogether, including inQTLs and ex-inQTLs expands, and allows disentangling, the effect of genetic variants on gene regulatory processes, and it may help elucidating the modulation of human traits through gene regulation.

Genes

Diving into tumor heterogeneity

Marina Mendiburu-Eliçabe (IBSAL), Adrián Blanco-Gomez (Cancer Research UK Manchester Institute), Diego Alonso-Lopez (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL), Roberto Corchado-Cobos (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL), Natalia García-Sancha (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL), Sonia Castillo-Lluva (Departamento de Bioquímica y Biología Molecular; Facultad de Ciencias Químicas, Universidad Complutense), Andrés Castellanos-Martín (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL), María del Mar Sáez-Freire (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL), Julie Milena Galvis-Jiménez (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL), Alejandro Jiménez-Navas (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL), Manuel Jesús Pérez-Baena (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL), Asunción García-Sánchez (Servicio de Bioquímica Clínica, Hospital Universitario de Salamanca and Instituto de Biomedicina de Salamanca (IBSAL)), María Isidoro Martín (Servicio de Bioquímica Clínica, Hospital Universitario de Salamanca and Instituto de Biomedicina de Salamanca (IBSAL)), Martin Pérez-Andrés (Departamento de Medicina. Universidad de Salamanca. and Instituto de Biomedicina de Salamanca (IBSAL)), Alberto Orfao (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL), Jian-Hua Mao (Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States), Javier De Las Rivas (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL) and Jesús Pérez Losada (Instituto de Biología Molecular y Celular del Cáncer (IBMCC) de Salamanca. CSIC. IBSAL).

Abstract:

Background: Why do patients have different clinical pathophenotypes of the disease? Why do patients with the same histopathological type of cancer have different evolution? Why do patients respond differently to treatment?

Description: In order to answer these questions, learn more about the disease, achieve a more personalized medicine, it is essential to carry out a detailed study of the heterogeneity of tumors, identifying the factors responsible. In the present work we studied the susceptibility and evolution of breast cancer in a murine population with controlled genetic heterogeneity, generated by a backcross strategy. The complex phenotypes, at the systemic, tissue, cellular, transcriptomic and genomic levels, are responsible for the clinical behavior of breast cancer. These phenotypes are in turn modified by multiple intermediate phenotypes that are determined at the genetic level.

Conclusions: In our research, we propose a strategy of transcriptomic analysis of breast tumors originating from transgenic mice. We look for gene signatures associated with the clinical pathophenotypes analyzed during breast cancer evolution and the intermediate phenotypes that participate in its pathogenesis. In this way, we out biological networks with genetic, molecular and cellular determinants that could define the different evolution and response to breast cancer treatment and explain the disease's complexity, evolution of complex-trait disease, response to treatment and identifying a part of the missing heritability.

Genes

EmbedPVP: Prioritizing causative variants by integrating functional embedding and biological annotations for genes

Azza Althagafi (KAUST) and Robert Hoehndorf (KAUST).

Abstract:

Whole-exome and genome sequencing has become a common tool in diagnosing patients with rare diseases. Despite its success, this approach leaves many patients undiagnosed. It is argued that more disease variants still await discovery or the novelty of disease phenotypes results from a combination of variants of multiple known disease-related genes. Interpreting the phenotypic consequences of genomic variants relies on information about gene functions, expression, and other genomic features. Phenotype-based methods to identify variants involved in genetic diseases combine molecular features with prior knowledge about the phenotypic consequences of altering gene functions. While phenotype-based methods have been successfully applied to prioritizing single nucleotide variants, such methods are based on known gene-diseases association as training data. In addition, the difference in phenotypes that come from clinicians with the phenotypes in the public databases makes it more challenging to predict the causing variants. Furthermore, single nucleotide variants, as well as short insertions and deletions, can affect a large number of coding regions, and phenotype information may not be available for all of them. We developed Embedding PhenomeNET Variant Predictor (EmbedPVP), a computational method to prioritize variants involved in genetic diseases by combining genomic information and clinical phenotypes and leveraging a large amount of background knowledge from human and invertebrate models. We incorporate phenotypes linked to genes, functions of gene products, and gene expression, and systematically relate them to their phenotypic effects through deep and knowledge-based learning models. We demonstrate EmbedPVP's efficacy on a large set of synthetic genomes and genomes matched with clinical information.

Genes

EmpiReS: Differential Analysis of Gene Expression and Alternative Splicing

Gergely Csaba (LMU Munich), Evi Berchtold (LMU Munich), Armin Hadziahmetovic (LMU Munich), Markus Gruber (LMU Munich), Constantin Ammar (LMU Munich) and Ralf Zimmer (LMU Munich).

Abstract:

While absolute quantification is challenging in high-throughput measurements, changes of features between conditions can often be determined with high precision. Therefore, analysis of fold changes is the standard method sufficient for differential expression, but often, the analysis of "changes of changes" is required. Differential alternative splicing is an application of such a doubly differential analysis. EmpiReS is a quantitative approach for various kinds of omics data based on fold changes for appropriate features of biological objects. Empirical error distributions for these fold changes are estimated from Replicate measurements and used to quantify feature fold changes and their directions. We assess the performance of EmpiReS to detect differentially expressed genes applied to RNA-Seq using simulated data. It achieved higher precision than established tools at nearly the same recall level. Furthermore, we assess the detection of alternatively Spliced genes via changes of isoform fold changes on distributionfree simulations and on experimentally validated splicing events. EmpiReS achieves the best precision-recall values for simulations based on different biological datasets. We propose EmpiReS as a general, quantitative and fast approach with high reliability and an excellent trade-off between sensitivity and precision for both differential expression and differential alternative splicing.

Genes

EvolclustDB: exploring eukaryotic gene clusters with evolutionary conserved synteny.

Uciel Chorostecki (Barcelona Supercomputing Centre (BSC) and Institute for Research in Biomedicine), Marina Marcet-Houben (Barcelona Supercomputing Centre (BSC) and Institute for Research in Biomedicine), Ismael Collado-Cala (Barcelona Supercomputing Centre (BSC) and Institute for Research in Biomedicine), Andrés Garisoain-Zafra (Barcelona Supercomputing Centre (BSC) and Institute for Research in Biomedicine), Diego Fuentes Palacios (Barcelona Supercomputing Centre (BSC) and Institute for Research in Biomedicine) and Toni Gabaldón (Barcelona Supercomputing Centre (BSC) and Institute for Research in Biomedicine).

Abstract:

Gene order in eukaryotic genomes is often poorly conserved through evolution. Despite this, multiple examples of metabolic gene clusters in eukaryotes are known, particularly among fungi and plants. Furthermore, certain groups of genes remain close in the genome over long evolutionary distances, which suggests that selection acts to maintain their genomic co-localization.

We developed EvolClustDB, a browsable database of evolutionarily conserved gene clusters pre-computed with the EvolClust algorithm (Marcet-Houben et al., 2020, Bioinformatics). This algorithm identifies groups of neighboring genes whose proximity is significantly conserved across evolution compared to the genome average. This prediction is performed pairwise in and all against all comparisons of genomes of interest, and then clusters are grouped into multi-species families. We inferred ~40,000 cluster families in 838 different species from Fungi, Plants, Metazoans, Insects and Protists. EvolClustDB allows browsing through the clusters, searching by protein, species, cluster or sequence. Visualization allows inspecting gene order per species in a phylogenetic context, enabling inference of potential evolutionary events.

All in all, EvolClustDB constitutes a resource of broad interest and applicability. EvolClustDB server is freely available, without registration, at http://evolclustdb.org/.

Genes

Exploring mRNA isoform diversity in mouse

Agata Muszyńska (Institute of Automatic Control, Electonics and Computer Science, Silesian University of Technology, Gliwice, Poland), Ryszard Przewłocki (Department of Molecular Neuropharmacology, Institute of Pharmacology Polish Academy of Sciences, Kraków, Poland) and Paweł P. Łabaj (Małopolska Centre of Biotechnology UJ, Kraków, Poland).

Abstract:

Almost all multiexonic genes in mammals undergo alternative splicing, a process in which exons are joined in different ways, resulting in new mRNAs or ncRNAs. This process is critical for cell development and differentiation, and its dysfunction is associated with numerous diseases. It is also the mechanism that increases the complexity of the transcriptome. We present the results of studying splicing events in data from an experiment focused on neuropathic pain in mice. One of the tools that provides the opportunity to study this is Spladder. It builds an augmented splicing graph based on the current annotation and then expands it with new events. In our study, we focused on finding commonalities in a collection of 88 fairly diverse samples to expand the currently known mouse transcriptomic landscape. We hypothesize that the inclusion of multiple pathological factors should allow the detection of novel alternative splicing events (nASEs) specific for spinal cord irrespective of the stress conditions. In our study, we were able to find nASEs common to all samples. In addition, the results of the functional analysis, both the GO terms and further changes at the protein level, showed a clear link to the nervous system. We were also able to track changes in functional annotation, depending on the type of events considered. This result may suggest that the mouse reference model lacks information for brain tissue but also reflects the expected neuroplasticity.

Genes

Exploring the functional landscape of soft tissue sarcoma

Miriam Payá-Milans (FISEVI-HUVR), María Peña-Chilet (FPS-HUVR), Carlos Loucera (FPS-HUVR), Marina Esteban-Medina (FPS-HUVR) and Joaquín Dopazo (FPS-HUVR).

Abstract:

Soft tissue sarcomas (STSs) are a group of rare cancers that are difficult to treat. The most frequent procedure is surgery, which in most cases does not prevent a relapse. An alternative would be Neoadjuvant Chemotherapy (NCT), which has the potential to downstage current disease and target micrometastases, thus reducing the risk of relapse. Current status of NCT use reflects the lack of consensus for treatment of STS, finding a huge variability in the treatments used. In our group, we are interested in finding new targets that could serve as biomarkers to improve the cure rates.

We used RNA-Seq data from The Cancer Genome Atlas (TCGA) Program and The Genotype-Tissue Expression (GTEx) project through the resource Recount3, where data was preprocessed with the same pipeline. Analyses we performed include: differential subpathway activation between STS and sarcomogenic normal tissues after conversion of gene expression into activity of KEGG signaling pathways with the method HiPathia; activation of transcription factors (TFs) through the enrichment of their targets ranked after differential expression analysis; survival analysis of relevant circuits and genes.

Data preanalysis shows good separation of tissues, and activity of TCGA normal tissues close to GTEx ones. Analysis of commonly deregulated subpathways in STS highlights the activation of the immune response, also linked to higher patient survival, and suppression of metabolic processes. These processes are further confirmed by deregulated genes and TFs. Extraction of shared mechanisms on STS facilitates the establishment of a common background in order to help with treatment.

Genes

GAZE: A single-cell gene regulatory inference framework from transcriptomics and epigenomics data

Shamim Ashrafiyan (Goethe University Frankfurt), Fatemeh Behjati Ardakani (Goethe University Frankfurt), Dennis Hecker (Goethe University Frankfurt) and Marcel Schulz (Goethe University Frankfurt).

Abstract:

Single-cell sequencing has become a prevalent approach to interrogate cell-type specific signatures and cellular heterogeneity, which assists researchers to unravel the underlying complexities of diseases. This, however, creates a need for integrating single-cell omics data through building specialized machine learning approaches that are capable of inferring key regulatory players at single-cell granularity. Although there have been numerous methods proposed for discovering transcriptional regulation on the basis of scRNA-seq data, they lack delivering a comprehensive view of the whole regulatory landscape.

Here, we address these limitations by incorporating diverse single-cell modalities. We have established a versatile statistical framework, called GAZE, that guarantees a comprehensive analysis of single-cell data in an integrative fashion.

This allows us to broaden the current understanding of transcriptional regulatory mechanisms through identifying the key players involved in differential regulation of various cell types.

Interrogating these models (regression coefficients or SHAP values) enables us to reveal interesting and novel regulatory aspects. Additionally, we designed adept tests for investigating the inferred regulatory activities to identify key genes or TFs driving cell regulation.

Finally, we have implemented an R shiny application to easily visualize and retrieve important regulators at single-cell or meta-cell level.

Genes

Genetic determinants of blood transcript splicing and impact on molecular phenotypes in 4732 healthy individuals

Alex Tokolyi (Wellcome Sanger Institute), Elodie Persyn (University of Cambridge), Katie Burnham (Wellcome Sanger Institute), Artika Nath (University of Cambridge), Jonathan Marten (University of Cambridge), David Roberts (University of Oxford), Emanuele Di Angelantonio (University of Cambridge), John Danesh (University of Cambridge), Adam Butterworth (University of Cambridge), Mike Inouye (University of Cambridge), Dirk Paul (University of Cambridge) and Emma Davenport (Wellcome Sanger Institute).

Abstract:

Untangling the pathways by which genetic variants modulate molecular phenotypes and disease risk requires comprehensive integrated analyses in large, deeply phenotyped cohorts of individuals. Splicing quantitative trait loci (sQTLs) are major contributors to complex traits, with a similar contribution as those affecting gene expression levels. However, the mechanisms that link these variants to downstream molecular and disease phenotypes through transcript splicing remain to be explored. INTERVAL is a deeply phenotyped cohort of healthy blood donors across England, including 4,732 individuals with matched genotypes and whole-blood RNA-seq, as well as proteomic, lipidomic, and metabolic measurements.

Here we utilize the split reads present in RNA-seq to describe the genetic architecture of splicing in whole blood, yielding 31,319 sQTLs (at FDR<0.05) across 8,799 genes. Comparing and colocalizing these sQTLs to expression QTLs (eQTLs) mapped in the same individuals reveals 47.9% of tested genes possess a colocalizing sQTL. Splicing also appears to share genetic regulation with plasma protein QTLs (pQTLs), assayed by SOMAscan, with 50.3% of 346 proteins having genetic signals colocalizing with sQTLs after conditional association analysis. Subsequent colocalization of sQTLs and pQTLs with COVID-19 HGI summary statistics recapitulates established associations of splicing and protein levels with disease risk in OAS1, as well as providing novel splicing associations in risk genes such as IFNAR2 and OAS3.

This largest to-date splicing QTL catalog and associated preliminary analyses will be a useful resource for future study of the genetic architecture of transcript splicing and the subsequent impact on molecular phenotypes and disease risk.

Genes

Genome-wide high-resolution identification of regulatory small RNAs in bacteria based on hybrid transcriptome sequencing data

Muhammad Elhossary (ZB MED - Information Centre for Life Sciences), Konrad Förstner (ZB MED - Information Centre for Life Sciences), Lauren Walling (Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)), Kai Papenfort (Friedrich Schiller University of Jena) and Gisela Storz (Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)).

Abstract:

In bacteria, small RNAs (sRNA) are important post-transcriptional regulators, and a single small RNA can influence the activity of numerous mRNAs by a variety of mechanisms. Bacteria typically express hundreds of these regulatory small RNAs. While extensively studied in a few bacterial species, their abundance, evolution, and biological functions remain largely unknown in most bacteria despite their vital role. Here we describe the application of high-throughput sequencing approaches together with computational data integration for the genome-wide annotation of sRNAs in 20 bacterial species of the phylum Gammaproteobacteria. For this purpose, we have developed a generic workflow in which sequencing data of differential RNA-Seq (dRNA-Seq) and Term-Seq are jointly analyzed to precisely call the 5’ ends and 3’ ends of the small RNAs, and confidently produce high resolution annotations. The approach includes a scoring and ranking method to quantify the confidence in predicted sRNAs. The method was successfully benchmarked against manually-curated sRNAs and is capable of detecting sRNAs encoded in various locations including those overlapping with open reading frames.

Genes

High-dimensional Differential Explorer

Felix Offensperger (LMU Munich), Markus Joppich (LMU Munich) and Ralf Zimmer (LMU Munich).

Abstract:

High-dimensional sequencing experiments measure many samples from combinations of several conditions, several knockdowns, various perturbations, many time points - all combinations with several replicates. In order to understand the differences in complex regulations (sometimes across different species) many sample combination measurements and heterogeneous techniques are required to resolve even simple mechanisms and possible causal interactions. Thus, an overview of a large number of measurements and the interpretation of differential and multi-differential analyses and the derivation of possible causal relationships is difficult but crucial.

To address these issues, High-dimensional Differential Explorer (HiDifE) introduces a number of extensions of state-of-the-art visualisations to enhance the interpretation of differential analyses. As differential analyses are difficult to interpret due to either too few or too many (inconsistent) results, the exploitation of double or even high-dimensional differential analysis is a challenge. The HiDifE categorizes the observed measurements per dimension into interpretable fuzzy values via fuzzification and derives error measures and significant differences for the individual categories and category combinations.

HiDifE provides several visualisations: (1) PCA/UMAP/TSNE maps, where the conditions along a dimension are connected by arrows indicating the trend of respective samples dimensions along certain other dimensions. This yields an overview of global changes in the data irrespective of individual genes. (2) Extended scatter plots for double differential analysis via comparison of two differential analyses. (3) Sankey plots to visualize individual gene changes with respect to the defined categories along certain dimensions. Every plot can spotlight gene sets or pathways for detailed analysis.

Genes

Identification and characterization of aneuploid cells applied to 5q deleted Myelodysplastic Syndromes

Guillermo Serrano (Center for Applied Medical Research), Aintzane Díaz (Center for Applied Medical Research), Nerea Berastegui (Center for Applied Medical Research), Marina Ainciburu (Center for Applied Medical Research), Sofia Huerga (Clinica Universidad de Navarra), José María Lamo de Espinosa (Clínica Universidad de Navarra), Mikel San Julián (Clínica Universidad de Navarra), Pamela Acha (Universitat Autonoma de Barcelona), Tamara Jimenez (Hospital Universitario de Salamanca-IBSAL), An

Abstract:

Deletion 5q (5q-) is a rare form of Myelodysplastic Syndromes (MDS) where a fragment of the long arm of chromosome 5 is lost from one allele.

Since differentially expressed genes are the downstream results of all the genetic machinery, the deletion of just one allele makes uncovering and studying these abnormal cells from single-cell RNA sequencing (scRNA-Seq) challenging due to genetic compensation by the other allele, and the shallowness scRNA-Seq.

In this work, we started by showing that single-cell signature-transfer methods coupled with clustering methods were not capable of identifying 5q-deleted cells from scRNA-Seq data.

To address this issue, we first uncovered (5q-)-characterizing genes from FACS-isolated CD34+ cells from isolated (5q-)-patients, and low-risk MDS cases with anemia and normal karyotype. Based on such genes, we developed a scoring function to quantify the (5q-)-ness of single cells that when applied to two independent bulk MDS datasets showed that (5q-)-MDS samples scored significantly higher (p-adj < 1e-5) than normal MDS and healthy cases. We showed that highly (5q-)-scored cells were enriched in cells showing CNVs on chromosome 5q as inferred by CASPER and CopyKat. These CNV inference methods are known to produce slightly noisy results precluding the fine identification of 5q-deleted cells. Thus, when combined with the developed score we were able to yield a robust set of (5q-)-cells.

Therefore, the proposed signature method seamlessly complemented previously-proposed CVN inference methods and allowed, for the first time, to clearly identify (5q-)-cells from scRNA-Seq, enabling future analyses to characterize this malignancy both transcriptionally and functionally.

Genes

Identification of biomarkers associated with cisplatin resistance in testicular germ cell tumours

Dominik Hadzega (nstitute of Molecular Biology, Slovak Academy of Sciences, Bratislava / Medirex Group Academy n. p. o., Nitra), Xavier Tominisec Vandoren (Department of Sciences and Technology, Haute École en Hainaut, Mons), Katarina Kalavska (2nd Department of Oncology, Faculty of Medicine, Comenius University and National Cancer Institute, Bratislava), Silvia Schmidtova (2nd Department of Oncology, Faculty of Medicine, Comenius University and National Cancer Institute, Bratislava), Luc

Abstract:

Germ cell tumours (GCTs) are the most common tumours in young males between 20 to 40 years. Although, it is usually curable disease, some patients develop resistance for standard therapy procedures. Our study aimed to describe changes participating in cisplatin resistance in GCTs, in so called cisplatin resistant testicular GCTs (TGCTs) and identifying of associated biomarkers. Although several processes were linked to cisplatin resistance, exact mechanism remains to be clarified. Parental sensitive and corresponding cisplatin resistant TGCTs cell lines previously derived by our research team (6 parental and 26 cisplatin resistant variants) were sequenced using both RNA-seq and whole exome sequencing and compared to search for responsible genes and pathways. For the purpose of identification of potentially deregulated genes, multiple computational tools were used in differentially expressed genes analysis. Subsequently, results were further analysed for Gene Ontology and pathway analysis. For the purpose of identification of sequence variants, variant calling was done using multiple pipelines and variants unique for cisplatin resistant cell lines were identified. This work was supported by grant APVV-20-0158.

Genes

Identification of novel RNA Switches using inverse RNA folding

Sumit Mukherjee (Ben-Gurion University) and Danny Barash (Ben-Gurion University).

Abstract:

RNA switches or riboswitches are conserved structural RNA sensors which are mainly found to regulate a large number of genes/operons in bacteria. More than 40 classes of riboswitches have been discovered in bacteria. Only the TPP riboswitch class is also detected in eukaryotes like fungi, plants, and algae. One of the most important challenges in riboswitch research is to discover additional riboswitch classes in eukaryotes. Traditional search methods for riboswitch detection, such as covariance models (CM) and profile hidden markov models (pHMM) failed to detect additional riboswitches in eukaryotes. We developed a novel method based on inverse RNA folding that attempts to find sequences that match the shape of the target structure with minimal sequence conservation based on key nucleic acids such as those that interact directly with the ligand. Then, to support our matched candidates, we expanded single sequence results into a covariance model representing similar sequences preserving the structure. Each matching sequence is aligned to the consensus secondary structure in the model and verifies that the secondary structure shape has not been lost and that the sequence constraints are still present within their structural context. Our method transforms a structure-based search into a sequence-based search and has been validated by detecting known bacterial riboswitches available in Rfam. Using this method, we identified several potential eukaryotic riboswitch candidates and bacterial riboswitches which are not available in Rfam. Furthermore, their folding prediction was checked to ensure the ligand could potentially bind to them before experimental verification.

Genes

Improving genome annotations with RNA-seq data: a scalable and reproducible workflow for Transcripts And Genes Assembly, Deconvolution, Analysis (TAGADA).

Cyril Kurylo (INRAE), Cervin Guyomar (INRAE), Sarah Djebali (INSERM) and Sylvain Foissac (INRAE).

Abstract:

Genome annotation aims to provide a comprehensive, accurate and nonredundant catalog of all reported genes and transcripts for a given species. Research projects worldwide routinely generate new transcriptome data to be integrated into existing annotations in a consistent manner. While many bioinformatics pipelines can build disparate de novo transcriptomes or quantify gene expression levels from a provided annotation, they usually do not properly update a reference annotation using new RNA-seq data, in particular for large numbers of samples.

Here we present TAGADA, an RNA-seq pipeline for Transcripts And Genes Assembly, Deconvolution, Analysis. Given a genomic sequence, a reference annotation and RNA-seq short reads, TAGADA extends the set of provided gene and transcript models with new ones, generating an improved annotation. In addition, it computes expression values for both the reference and the novel annotation, detects long non-coding transcripts (lncRNAs), and generates a comprehensive report with multiple quality controls. Built with Nextflow DSL2, TAGADA is easy to use and provides a containerized environment to ensure reproducibility across computing platforms. TAGADA has been used in the context of the FAANG action (Functional Annotation of ANimal Genomes) to detect thousands of new coding and non-coding genes in several livestock species. The code is available at https://github.com/FAANG/analysis-TAGADA.

“This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement No. 817998”.

Genes

Inferring aberrant expression dynamics across early myeloid differentiation to discover potential therapeutic targets in myelodysplastic syndromes

Aintzane Díaz-Mazkiaran (Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain), Jesús De la Fuente (TECNUN, Universidad de Navarra, Spain), Guillermo Serrano (Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain), Paula Garcia-Olloqui (Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain), Nerea Berastegui (Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain), Marina Ainciburu (Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain), Ana Alfonso (Clínica Universidad de Navarra, Spain), Amaia Vilas-Zornoza (Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain), Patxi San Martin (Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain), Jose Lamo de Espinosa (Clínica Universidad de Navarra, Spain), Mikel San-Julian (Clínica Universidad de Navarra, Spain), Pamela Acha (Myelodysplastic Syndromes, Josep Carreras Leukaemia Research Institute, Universitat Autònoma de Barcelona, Spain), Francesc Solé (Myelodysplastic Syndromes, Josep Carreras Leukaemia Research Institute, Universitat Autònoma de Barcelona, Spain), Tamara Jimenez (Hematology, Hospital Universitario de Salamanca-IBSAL, Salamanca, Spain), Félix López (Hematology, Hospital Universitario de Salamanca-IBSAL, Salamanca, Spain), María Díez-Campelo (Hematology, Hospital Universitario de Salamanca-IBSAL, Salamanca, Spain), Antonieta Molero (Department of Hematology, Vall d'Hebron Institut Oncologic (VHIO), University Hospital Vall d'Hebron, Barcelona, Spain), María Julia Montoro (Department of Hematology, Vall d'Hebron Institut Oncologic (VHIO), University Hospital Vall d'Hebron, Barcelona, Spain), David Valcarcel (Department of Hematology, Vall d'Hebron Institut Oncologic (VHIO), University Hospital Vall d'Hebron, Barcelona, Spain), Teresa Ezponda (Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain), Mikel Hernaez (Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain) and Felipe Prósper (Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain).

Abstract:

Introduction: Myelodysplastic syndromes (MDS) represent hematological malignancies characterized by defective differentiation of hematopoietic stem and progenitor cells (HSPCs). Molecular studies on MDS have focused chiefly on genomic alterations, but such lesions fail to fully explain disease development. Hence, a detailed transcriptional characterization of HSPCs could help discover novel mechanisms underlying early myeloid differentiation abrogation in MDS.

Results: We developed a computational model to infer trajectory disruptions across myeloid differentiation, enabling the identification of key genes whose expression dynamics are altered in MDS. To unravel transcriptional alterations across early hematopoiesis, we performed MARS-sequencing on healthy individuals' and MDS patients' HSPCs (HSCs, CMPs, GMPs, MEPs). The developed model identified 579 and 711 disrupted genes along the monocytic/granulocytic (HSC-CMP-GMP) and the megakaryocytic/erythrocytic (HSC-CMP-MEP) lineages, respectively.

The vast majority of these genes were either positively or negatively disrupted across both differentiation pathways. Ontology studies on negatively disrupted genes suggested their participation in myeloid cells' differentiation and functionality, including neutrophil activation/degranulation and gas transport. Further, the application of TraRe, a Gene-Regulatory-Network inference method, uncovered transcription factors that could be globally driving transcriptomic dysregulations in MDS. Among them, ZNF350 and ZMAT2 happened to be associated with key disrupted genes, and genome-wide CRISPR-Cas9 screenings showed that their inhibition promoted differentiation of MDS cell lines, thus representing potential therapeutic targets for reverting differentiation blockage in MDS.

Conclusions: Collectively, these findings offer a new approach in the study of MDS pathogenesis and shed light into novel drivers, contributors or by-passers of aberrant hematopoiesis in the disease.

Genes

Influence of Topologically Associating Domains on gene expression variation

Patrycja Rosa (University of Warsaw) and Aleksander Jankowski (University of Warsaw).

Abstract:

Gene expression profiles differ between tissues in an organism, and also within heterogeneous populations of cells forming a single tissue. The variation in gene expression may be caused by different mechanisms, and can also be decisive for the fate of cells. We aimed to test whether the variation in gene expression could possibly be regulated by the spatial chromatin structure, especially by Topologically Associating Domains.

For our study, we used published scRNA-seq data from Drosophila melanogaster ventral nerve cord. We grouped 24,199 cells into 142 cell clusters, corresponding to specific subpopulations. To obtain the information about gene location in TAD structure, and to calculate the distance from TAD boundaries, we integrated them with previously published Hi-C data from Kc167 cells. We calculated the average gene expression level in each cell cluster, and further used the mean and standard deviation of these averages to quantify gene expression variability.

We found that mean and standard deviation of expression are mostly correlated with each other, which justified further use of their ratio (coefficient of variation) as one of the features. Overall, genes located close to the TAD boundary are less variable in their expression, and those transcribed in the direction away from the TAD boundary have higher variability. We further considered pairs of genes at different distance ranges, calculated their spatial autocorrelation, and compared them to a model with permuted gene expression profiles. We observed an enrichment indicative of a relationship between gene expression variation and the gene location in the chromatin structure.

Genes

Integrating multi-omics and biophysical data to explore cellular responses against phenolic compounds in yeast

Natalia Coutuouné (University of Campinas | UNICAMP), Rafael Boni (LNBR - CNPEM | UNICAMP), Cleyton Biffe (LNNano -CNPEM | UNICAMP), Ingrid Barcelos (Sirius - CNPEM), Silvana Rocco (Sirius - CNPEM), George Jackson de Moraes Rocha (LNBR - CNPEM) and Leandro Vieira dos Santos (SENAI Biotecnologia).

Abstract:

The environmental situation of the planet requires that we urgently rethink our way of production. A promising approach is using agroindustrial-waste biomass to develop an integrated biorefinery for producing biofuels and high-value biochemicals from renewable sources. In Brazil, the sugarcane biorefinery is based on the yeast's fermentation of sugars released from polysaccharides in the plant biomass. Sugarcane depolymerization also releases toxic compounds: the inhibitors, such as phenolics, a heterogeneous group derived from the lignin decomposition and one of the most abundant and toxic components. We developed a method to obtain an enriched fraction of phenolic (EFP) compounds from sugarcane hydrolysate. The main component of the EFP was identified using NMR, GC/MS, and -LC/MS techniques. Also, we supplemented the EFP obtained in a synthetic medium. We performed an RNA-seq and proteomic analysis to study the cellular responses in the presence of phenolics on an engineered high-performing industrial yeast strain. Additionally, the preliminary results suggest that the most pronounced effects are related to the structural components of yeast. Furthermore, we detected responses caused by oxidative stress and alterations in the expression profile of genes related to ergosterol homeostasis and autophagy. Finally, we integrate molecular analysis with biophysical approaches (Bio-AFM, FTIR) to investigate structural and nanomechanical alterations in the yeast cell. These findings will be critical to the rational design of more tolerant strains, which have a key role in converting agroindustrial-waste residues into a sustainable solution.

Genes

Investigating the Transcriptome of Adenomatous Colorectal Polyps in the Context of Metachronous Recurrence

Simon Fisher (Canon Medical Research Europe), Russell Hung (Canon Medical Research Europe), Ditte Andersen (BioClavis), Gerard Lynch (University of Glasgow), Noori Maka (NHS Greater Glasgow and Clyde), Jennifer Hay (University of Glasgow), Jakub Jawny (University of Glasgow), William Sloan (NHS Greater Glasgow and Clyde), Stephen McSorley (NHS Greater Glasgow and Clyde), Joanne Edwards (University of Glasgow) and Ian Poole (Canon Medical Research).

Abstract:

Adenomatous polyps are aberrant gland-like growths of the colon or rectum. These growths can potentially progress to cancer, thus understanding the prognostic risk markers of index polyps and how they may correlate with downstream recurrence is likely to have significant patient benefits.

In this study, we obtained TempOSeq transcriptomic count data of index polyps, alongside clinical information including metachronous polyp incidence, as part of the Integrated Technologies for Improved Polyp Surveillance (INCISE) project (n=1826, total). We performed differential expression analysis with DESeq2 for participants stratified by metachronous occurrence. Furthermore, we performed unsupervised weighted-gene-co-expression-analysis (WGCNA) and correlated gene modules with clinical traits, including metachronous occurrence and time-to-event. Finally, we analysed enriched hub genes for their ability to predict disease free survival (DFS) by Kaplan Meier analysis.

A total of 241 differentially expressed genes (DEGs) were detected, 100 downregulated and 141 upregulated, in individuals with metachronous polyps versus healthy controls ($PAdj < 0.05$). A total of 11 modules were detected by WGCNA, one of which correlated with metachronous polyps. Gene set enrichment analysis for both DEGs and this module identified enrichment toward extracellular matrix invasion and immune regulation. Six overlapping hub genes between both methods, HMOX, COL1A1, CD163, CTSL, FN1 and LOX, were found. By median split, none of these genes had significant prognostic potential in predicting metachronous events ($p > 0.05$).

Index polyps of individuals who develop metachronous events may have greater infiltrative and immunoregulatory capacity, but the prognostic signal of identified drivers is poor.

Genes

IsoAnnot: a pipeline for functional annotation of isoforms

Alessandra Martínez Martín (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia, Spain), Francisco Jose Pardo-Palacios (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia, Spain.), Pedro Salguero (Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Valencia, Spain), Ángeles Arzalluz-Luque (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia, Spain.), Lorena de la Fuente Lorente (PerkinElmer Informatics, R&D, Tres Cantos, Spain.) and Ana Conesa (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia, Spain.).

Abstract:

Post-transcriptional regulation (PTR) mechanisms are essential for creating transcriptome and proteome complexity in eukaryotic cells. Previous studies suggest that alternative exons modulate properties such as subcellular localization or mRNA stability. However, how these mechanisms imprint distinct functional characteristics on the resulting set of isoforms to define the observed phenotype remains poorly understood.

Functional profiling is the most broadly adopted genome-wide approach for characterization of the functional relevance of gene expression regulation. However, the gene-centric nature of functional annotation resources such as Gene Ontology prevents the study of functional consequences of differential splicing in specific contexts. The emergence of third-generation sequencing technologies has allowed the characterization of full-length isoform sequences and motivates the development of methods for isoform-centered functional analysis.

Here we document the development of IsoAnnot, a new pipeline for the functional characterization of isoforms. This tool considers an extensive variety of functional properties, both at RNA and protein level. Importantly, most of the functional labels are defined by protein/RNA coordinates which enables the direct mapping of splicing events to functional elements. The pipeline uses transcript sequences to construct an isoform-resolved database of functional annotations by integrating information stored in public databases such as Uniprot and PhosphositePlus and tools based on sequence-prediction such as UTRscan and RepeatMasker. The main advantage of IsoAnnot is its ability to annotate known and novel isoforms obtained from long-read sequencing. IsoAnnot have been developed using Nextflow and Conda environments to ensure scalability and reproducibility.

Genes

K-nearest neighbour correction for confounding effects in gene expression measurements

Franco Simonetti (Fundacion Instituto Leloir), Macarena Alonso (Fundacion Instituto Leloir), Cristina Marino Buslje (Fundación Instituto Leloir), Saikat Banerjee (Department of Statistics, University of Chicago) and Johannes Soeding (Max Planck Institute for Multidisciplinary Sciences).

Abstract:

Gene expression measurements can be dominated by strong confounding effects such as technical details of RNA recovery, sample conservation, sequencing, environmental and biological factors. Typical tools for removing confounding effects from expression data use linear regression models to regress out known confounders. Statistical tools such as PCA or PEER analysis can infer new covariates that can be later regressed out in the same way as with known confounders.

We developed a method based on an unsupervised K-nearest neighbour (KNN) approach where we assume that confounding effects dominate the gene expression. If the samples are close to one another in the expression space, we expect them to be close to one another in the confounder space and hence, to be able to correct them.

We applied gene expression corrections using current standard methods and our KNN approach. We trained PrediXcan models for each gene using RNA-seq data and matched genotypes from GEUVADIS dataset, and tested the models accuracy by predicting gene expression levels on GTEx matching tissue (LCL).

We compared the accuracy in gene expression prediction (Pearson rho) after training with different confounder correction settings. Our initial results indicate that KNN correction provides comparable results to PEER (< 0.01 avg R2 difference, KS-test p-value=0.99), although some genes show better prediction R2 with one method or the other. This could be due to over- or under-correction in either method applied. The approach is completely unsupervised, it only depends on the parameter K and can run >100x faster than PEER.

Genes

Lung tissue multi-layer network in Chronic Obstructive Pulmonary Disease

Núria Olvera (Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS); Barcelona Supercomputing Center (BSC)), Jon Sánchez (Barcelona Supercomputing Center (BSC)), Iker Núñez (Barcelona Supercomputing Center (BSC)), Guillaume Noell (Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS)), Sandra Casas (Centro de Investigación Biomédica en Red de Enfermedades Respiratorias M.P. (CIBERES)), Alejandra Lopez (Respiratory Institute, Hospital Clínic, University of Barcelona), Angela Guirao (Respiratory Institute, Hospital Clínic, University of Barcelona), Rosalba Lepore (Barcelona Supercomputing Center (BSC)), Davide Cirillo (Barcelona Supercomputing Center (BSC)), Àlvar Agustí (Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS); CIBERES; Universitat de Barcelona (UB)), Alfonso Valencia (Barcelona Supercomputing Center (BSC)) and Rosa Faner (Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS); CIBERES; Universitat de Barcelona (UB)).

Abstract:

Background: Chronic Obstructive Pulmonary Disease (COPD) is a highly heterogeneous condition. However, the biological mechanisms (endotypes) underlying this heterogeneity are still largely undefined.

Methods: In this study for the first time we integrate three omics levels (mRNA, miRNA and DNA methylation) determined in 135 lung tissue samples of ex-smokers with COPD. A patient network, identifying the molecular similarities between individuals, was built for each omic level resulting into a multi-layer network. Finally, communities of patients were detected within the multi-layer framework.

Results: Four stable multi-omics communities were identified with significant differences in relevant clinical features, including lung function parameters (FEV1 % ref.), Body Mass Index and blood eosinophils. Specifically, in one unique subtype enriched with severe patients we saw differences at all the layers. It had an hypomethylation of DNA regions that overlapped with genes that participate in the immune response and the same inflammatory pathways were upregulated at the mRNA level. These individuals also had a strong cilia dysfunction as observed in transcriptomics data, which was mediated by the downregulation of miR-34/449 family (required in ciliogenesis) at miRNA level. We could finally validate the mRNA and miRNA features of this group in the patients of an independent cohort from the Lung Genomics Research Consortium.

Conclusion: We show for the first time that the use of a multi-layer network based on the lung tissue patient similarities, uncovered communities that differ in clinical COPD characteristics and shed light on the biological mechanisms underlying the heterogeneity of the disease.

Genes

Machine leaning on whole blood RNAseq identifies interferon-regulated genes as key drivers in thrombotic primary Antiphospholipid syndrome

Kleio-Maria Verrou (National and Kapodistrian University of Athens), Petros Sfikakis (National and Kapodistrian University of Athens) and Maria Tektonidou (National and Kapodistrian University of Athens).

Abstract:

Objectives: Antiphospholipid syndrome (APS) is a rare autoimmune disease with significant morbidity and mortality, characterized by a wide range of thrombotic and pregnancy manifestations. Its pathogenesis is not fully elucidated. Here, we aimed to identify gene signatures characterizing thrombotic primary APS (thrPAPS).

Methods: We performed whole blood next-generation RNA-sequencing in 62 human patients with thrPAPS and 29 age-/sex-matched healthy controls (HCs), followed by deconvolution analysis for 22 immune cell subpopulations and differential gene expression analysis (DGEA). We identified the interferon regulate genes (IRGs) utilizing the Interferome database. Furthermore, we applied a stratified nested (n=3) cross-validation (k=10) classification methodology in Python using the deregulated genes. We tuned more than 1000 models for 3 classifiers, namely Support Vector Machine, Random Forest and k-Nearest Neighbours. Model selection was based on prediction accuracy. The methodology was also applied using only the differentially expressed IRGs.

Results: The deconvolution analysis returned no statistically significant differences in the cell subpopulations' abundances between thrPAPS and HCs. DGEA revealed 34 deregulated (Fold Change >|2|, meta p-value <0.05) genes in thrPAPS versus HCs; 33 were upregulated, and 14 out of those 33 were type I and II IRGs. Machine learning training applied with all deregulated genes returned 79% accuracy to discriminate thrPAPS from HCs, which increased to 86% when only IRGs were used.

Conclusion: Deregulated IRGs may better discriminate thrPAPS from HCs than all deregulated genes in peripheral blood. Taken together with DGEA data, IRGs may play a key role in thrPAPS regulation, with potential therapeutic implications.

Genes

Machine learning for improved immune cell type classification as a basis to better understand effects of cell type specific interferon stimulation

Bogac Aybey (University of Heidelberg - Merck Healthcare KGaA), Benedikt Brors (DKFZ), Sheng Zhao (Merck Healthcare KGaA) and Eike Staub (Merck Healthcare KGaA).

Abstract:

Robust immune cell gene expression signatures are central to the analysis of immuno-oncology single cell studies. Nearly all known sets of immune cell signatures have made use of single gene expression (scRNA-seq) datasets. Utilizing the power of multiple integrated datasets could lead to high quality immune cell signatures which could be used as inputs to machine learning (ML)-based cell type classification approaches.

We established a novel workflow for discovery of immune cell type signatures that leverages multiple datasets, here four scRNA-seq datasets from three different cancer types. We used these signatures in different own and published approaches, from naïve decision trees to Random Forest classification, to classify cells into immune cells. We included small sets of genes coming from our immune cell signatures in our random forest model. Our prediction model showed comparable results with commonly used methods in benchmarking dataset. Finally, we applied our cell type classification models to public and internal scRNA-seq studies to identify immune cell-specific interferon signaling. In both datasets, our random forest approach based on our immune cell signatures helped us to eliminate biases in the analysis and discover genes stimulated by different interferons in specific cell types.

We demonstrated the quality of our immune cell signatures and their utility for a ML-based cell typing approach. We argue that classifying cells based on our comparably slim sets of genes accompanied by a ML-based approach not only matches or outperforms widely used public approaches, but also allows unbiased downstream statistical analyses of genes between cell types.

Genes

Machine learning methods applied to genotyping data capture interactions between single nucleotide variants in late onset Alzheimer's disease

Magdalena Arnal Segura (Sapienza University of Rome) and Gian Gaetano Tartaglia (Sapienza University of Rome).

Abstract:

Genome-wide association studies (GWAS) in LOAD provide lists of individual genetic determinants associated with disease. However, GWAS are not designed to capture the synergistic effects among multiple genetic variants and lack specificity due to the multiple testing problem and linkage disequilibrium. These limitations hinder the discovery of multiple causal SNVs that most contribute to the disease phenotype, especially in polygenic risk diseases like LOAD. In this regard, machine learning methods (ML) can add an important layer of information to the disease-related variants obtained with population genomic approaches such as GWAS. We propose the use of tree-based ML in combination to SNVs reported in GWAS to classify individuals with LOAD and controls and discover interactions among SNVs. Rather than being a black box, tree-based ML retrieve the most important predictors used to reach an efficient discrimination of the classes and are particularly appropriate to classify based on categorical predictors such as SNVs. In our work we discovered a set of interactions linked with LOAD in UK Biobank with ML and validated them in an external dataset (ADNI) using generalized linear models. We show ML are relatively robust to linkage disequilibrium and perform a prioritization of variants not only based on the individual enrichment of each SNV in the different classes, but also considering interactions between groups of SNVs. We expect to continue this work and apply ML to discover other synergistic effects across neurodegenerative and cardiovascular diseases.

Genes

Macrophage subpopulations identification using single-cell RNA sequencing data

Mercè Alemany-Chavarria (Vall d'Hebron Institute of Oncology), Marta Lalinde (Vall d'Hebron Institute of Oncology), Joaquín Arribas (Vall d'Hebron Institute of Oncology) and Lara Nonell (Vall d'Hebron Institute of Oncology).

Abstract:

Single-cell RNA sequencing (scRNAseq) is an excellent technique to study expression profiles of individual cells and assess similarities and differences between cell populations. In oncology, scRNAseq analysis can help to characterize the tumor microenvironment, elucidate tumor progression or resistance to therapy (Nieto et al., 2021).

The purpose of our analysis was to identify the populations in a mice breast cancer immunological environment and how the epithelial and macrophages populations evolved throughout the experiment timeline. Of special interest was the identification and progression of M1-like and M2-like macrophages subpopulations.

The sequencing of the cells was performed by 10XGenomics and the alignment and quantification with Cell Ranger. From the count matrices, a quality control was applied to the samples, they were normalized, scaled and integrated before clustering all cells (Seurat, Hao et al. 2021). An automated cluster annotation was performed (SingleR, Aran et al. 2019) using as reference the ImmGenData expression dataset from (celldex, Aran et al. 2019). Once the macrophages were identified, a subclustering step was performed and the reference was refined using an accurate marker selection-based method. Clusters were relabeled to include the subpopulations of interest, according to the expression levels of specific markers.

Populations of M1-like and M2-like were identified successfully and their progression across conditions assessed.

Genes

Molecular and functional atlas of sex-differences in multiple sclerosis subtypes analising single cell and single nucleus transcriptomic data

Irene Soler-Sáez (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF), 46012, Valencia, Spain), Zoraida Andreu (Foundation Valencian Institute of Oncology (FIVO), 46009, Valencia, Spain), José Francisco Català-Senent (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF), 46012, Valencia, Spain), Rubén Grillo-Risco (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF), 46012, Valencia, Spain), Adolfo López-Cerdán (Bioinformatics and Biostatistics Unit CIPF and Biomedical Imaging Unit FISABIO-CIPF 46012, Valencia, Spain), Almudena Neva-Alejo (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF), 46012, Valencia, Spain), Borja Gómez (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF), 46012, Valencia, Spain), Héctor Carceller (Biomedical Imaging Unit FISABIO-CIPF; Institute Biotechnology and Biomedicine, Universitat de València-Burjassot-Spain), María de la Iglesia-Vayá (Biomedical Imaging Unit FISABIO-CIPF Valenciana, 46012, Valencia, Spain), Marta R. Hidalgo (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF), 46012, Valencia, Spain) and Francisco García-García (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF), 46012, Valencia, Spain).

Abstract:

Multiple sclerosis (MS) is the commonest cause of non-traumatic disability among young adults. MS hallmark underlies myelin damage induced by defective autoimmune responses, leading to the neurodegeneration of the central nervous system. Sex differences in MS have been reported at several epidemiological and clinical levels as prevalence, progression and response to treatment. However, the molecular mechanisms underneath those differences remain poorly understood.

To exhaustively characterise sex bias in MS by cell type, we performed an in silico analysis of scRNA-seq and snRNA-seq data using R programming language. Firstly, we performed a systematic review implementing PRISMA guidelines [PMID:33780438] in public repositories. Then, we processed each selected dataset through quality control filtering, normalisation, high variable genes selection, dimensionality reduction, clustering and cell type annotation. Finally, we characterise each cell type by differential gene expression and functional profiling analyses, evaluating for the latter biological functions from the Gene Ontology [PMID:10802651] and pathways from the KEGG [PMID:10592173] databases.

Three datasets, each representing a different subtype of MS, were spotted. Nervous tissue dataset (n=1) stored astrocytes, microglia, neurons, oligodendrocytes, and oligodendrocyte precursor cells, whilst blood datasets (n=2) included diverse types of lymphocytes, dendritic cells and monocytes. We found significant sex-differential and sex-specific gene expression patterns, biological functions, and pathways for almost all cell types in each dataset. Some significant features were shared among cell types with similar or opposite patterns, whilst others were cell type exclusive. Therefore, this atlas enhances personalised medicine by unveiling molecular and functional sex-dependent prospective biomarkers.

Genes

Molecular mechanisms governing CAR-T cell response in MM patients at single cell level

Lorea Jordana (CIMA Universidad de Navarra), Guillermo Serrano (CIMA Universidad de Navarra), María Erendira Calleja (CIMA Universidad de Navarra), Patxi San Martín-Úriz (CIMA Universidad de Navarra), Amaia Vilas-Zornoza (CIMA Universidad de Navarra), Asier Ullate-Agote (CIMA Universidad de Navarra), Aintzane Zabaleta (CIMA Universidad de Navarra), Diego Alignani (CIMA Universidad de Navarra), Aina Oliver-Caldes (Hospital Clinic de Barcelona), Marta Español-Rego (Hospital Clinic Barcelona), Mariona Pascal (Hospital Clinic de Barcelona), Teresa Lozano (CIMA Universidad de Navarra), Bruno Paiva (CIMA Universidad de Navarra), Susana Inoges (Clinica Universidad de Navarra), Ascension Lopez-Diaz de Cerio (Clínica Universidad de Navarra), Juan Jose Lasarte (CIMA Universidad de Navarra), Carlos Fernadez de Larrea (Hospital Clinic de Barcelona), Mikel Hernaez (CIMA Universidad de Navarra), Juan Roberto Rodriguez-Madoz (CIMA Universidad de Navarra) and Felipe Prosper (CIMA Universidad de Navarra, Clínica Universidad de Navarra).

Abstract:

As immunotherapy and biological therapies gain increasing repercussion in the field of oncology, single cell technologies are emerging as powerful tools to understand the molecular mechanisms by which they act. The application of single cell RNA sequencing (scRNA-seq) to study immunotherapy against hematological malignancies has shed light into important aspects of the in vivo evolution of CAR-T cells; however, a deep analysis of the gene regulatory networks (GRN) governing the expansion of CAR-T cells is missing.

In the present work we have used SimiC, a machine learn algorithm that infers GRNs from transcriptomic data, coupled with scRNA-seq and single cell TCR sequencing (scTCR-seq) to interrogate CAR-T cells isolated from bone marrow (BM) and peripheral blood (PB) of patients with Multiple Myeloma at different times after treatment. We have found that CAR-T cells suffer a dramatic shift in their transcriptomic profile after infusion into patients, and we have identified key GRNs as potentially responsible for these phenotypic changes. Moreover, our analysis showed that CAR-T cells in BM present a more cytotoxic phenotype with increased exhausted features than their PB counterparts. Besides, the combination of single cell TCR sequencing (scTCR-seq) with scRNA-seq and SimiC has allowed the identification and characterization of a hyperexpanded clone with immunosuppressor features in the BM of a relapsed patient.

Thus, our results show that single cell technologies coupled with machine learning algorithms have the potential to decipher important events in the evolution of CAR-T cells after infusion into patients that could have clinical relevance.

Genes

NetRank Recovers Known Cancer Hallmark Genes as Universal Biomarker Signature for Cancer Outcome Prediction

Ali Al-Fatlawi (TU Dresden) and Michael Schroeder (BIOTEC, TU Dresden).

Abstract:

Gene expression can serve as a powerful predictor for disease progression and other phenotypes. Consequently, microarrays, which capture gene expression genome-wide, have been used widely over the past two decades to derive biomarker signatures for tasks such as cancer grading, prognosticating the formation of metastases, survival, and others. Each of these signatures was selected and optimized for a very specific phenotype, tissue type, and experimental set-up. While all of these differences may naturally contribute to very heterogeneous and different biomarker signatures, all cancers share characteristics regardless of particular cell types or tissue as summarized in the hallmarks of cancer. These commonalities could give rise to biomarker signatures, which perform well across different phenotypes, cell and tissue types. Here, we explore this possibility by employing a network-based approach for pan-cancer biomarker discovery. We implement a random surfer model, which integrates interaction, expression, and phenotypic information to rank genes by their suitability for outcome prediction. To evaluate our approach, we assembled 105 high-quality microarray datasets sampled from around 13,000 patients and covering 13 cancer types. We applied our approach (NetRank) to each dataset and aggregated individual signatures into one compact signature of 50 genes. This signature stands out for two reasons. First, in contrast to other signatures of the 105 datasets, it is performant across nearly all cancer types and phenotypes. Second, It is interpretable, as the majority of genes are linked to the hallmarks of cancer in general and proliferation specifically. Many of the identified genes are cancer drivers with a known mutation burden linked to cancer. Overall, our work demonstrates the power of network-based approaches to compose robust, compact, and universal biomarker signatures for cancer outcome prediction.

Genes

New genes and splicing isoforms revealed with native RNA in fission yeast

José Carlos Montañés (Evolutionary Genomics Group, Hospital del Mar Medical Research Institute (IMIM) Universitat Pompeu Fabra (UPF)), M. Mar Albà (Evolutionary Genomics Group, Hospital del Mar Medical Research Institute (IMIM) Universitat Pompeu Fabra (UPF)), Marta Huertas (Evolutionary Genomics Group, Hospital del Mar Medical Research Institute (IMIM) Universitat Pompeu Fabra (UPF)), Simone G. Moro (Evolutionary Genomics Group, Hospital del Mar Medical Research Institute (IMIM) Universitat Pompeu Fabra (UPF)), Williarm R. Blevins (CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona, Spain), Mercè Carmona (Oxidative Stress and Cell Cycle Group, Universitat Pompeu Fabra (UPF), Barcelona, Spain.), José Ayte (Oxidative Stress and Cell Cycle Group, Universitat Pompeu Fabra (UPF), Barcelona, Spain.) and Elena Hidalgo (Oxidative Stress and Cell Cycle Group, Universitat Pompeu Fabra (UPF), Barcelona, Spain.).

Abstract:

Schizosaccharomyces pombe is a unicellular model organism used in splicing studies due to half of its genes have introns. However, little is known about the impact of alternative splicing on gene regulation and proteome diversification. Here we leverage Oxford Nanopore Technologies native RNA sequencing (dRNA) plus ribosome profiling data to investigate the full range of polyadenylated transcripts and translated open reading frames. We find that lowly expressed transcripts, including mRNAs encoding meiosis-related proteins but also intron retention isoforms, tend to have longer poly(A) tails than transcripts that are expressed at very high levels. We identify 332 alternative isoforms in 262 different genes 97 of which with higher frequencies than the 20% compared to the reference isoform suggesting their functionality to the organism. Intron retention events make about 80% of the cases; these events may be involved in the regulation of gene expression and, in some cases, generate novel protein isoforms, as supported by ribosome profiling data in 18 of the intron retention isoforms. One example is the rpl22 gene which is related to protein synthesis and whose intron retention new isoform is predicted to encode for a peptide of only 13 amino acids. In addition, we identify 214 completely new transcripts, including 158 antisense sequences of previously annotated genes, some of which with ribosome profiling signals indicating their translation. In summary, this study described new methodologies to study the regulation of gene splicing in a simple eukaryotic organism.

Genes

Optimal Transport improves cell-cell similarity inference in single-cell omics data

Geert-Jan Huizing (ENS PSL), Gabriel Peyre (CNRS and Ecole Normale Supérieure) and Laura Cantini (Institut de Biologie de l'Ecole Normale Superiore).

Abstract:

High-throughput single-cell molecular profiling is revolutionizing biology and medicine by unveiling the diversity of cell types and states contributing to development and disease. The identification and characterization of cellular heterogeneity are typically achieved through unsupervised clustering, which crucially relies on a similarity metric. We propose the use of Optimal Transport (OT) as a cell–cell similarity metric for single-cell omics data. OT defines distances to compare high-dimensional data represented as probability distributions. To speed up computations and cope with the high dimensionality of single-cell data, we consider the entropic regularization of the classical OT distance. We then extensively benchmark OT against state-of-the-art metrics over 13 independent datasets, including simulated, scRNA-seq, scATAC-seq and single-cell DNA methylation data. First, we test the ability of the metrics to detect the similarity between cells belonging to the same groups (e.g. cell types, cell lines of origin). Then, we apply unsupervised clustering and test the quality of the resulting clusters. OT is found to improve cell–cell similarity inference and cell clustering in all simulated and real scRNA-seq data, as well as in scATAC-seq and single-cell DNA methylation data. This work was recently published in Bioinformatics. A Python package and Jupyter notebooks ensure reproducibility of our analyses.

Genes

Power analysis of cell-type deconvolution across human tissues

Anna Vathrakokoili Pournara (EMBL-EBI), Zhichao Miao (EMBL-EBI) and Irene Papatheodorou (EMBL-EBI).

Abstract:

Cell-type deconvolution methods aim to infer cell-type heterogeneity and the cell abundances from bulk RNA-sequencing data. Since a plethora of methods have been developed, there is an urgent need for guidance on method selection. Although previously suggested benchmarks have paved the way to better understand the performance of deconvolution methods, the proposed tests remain theoretical and have only been applied to a handful of datasets. At the same time there is pressing interest to achieve decomposition of database-level transcriptomic data of different tissues, conditions and species, scenarios in which deconvolution methods have not been tested thoroughly. Here, we propose a large-scale, multi-level assessment of 28 available deconvolution methods, leveraging 44 single-cell RNA-sequencing (scRNA-seq), from 8 organs and tissues. We extend previous benchmarks, suggesting a comprehensive simulation framework to evaluate deconvolution across a wide range of scenarios and we provide guidelines on the normalisation and transformation strategies. We also show that regression-based deconvolution methods such as MuSiC, FARDEEP, DWLS and nnls are performing well but their performance is highly dependent on the reference selection and the tissue type. Lastly, we provide a modularised benchmarking pipeline that will speed up the evaluation of newly published methods and we showcase its applicability and significance in the large-scale decomposition of available tissue data.

Genes

Predicting off-target effects of antisense oligomers targeting bacterial mRNAs with the MASON webserver

Jakob Jung (Institute for Molecular Infection Biology (IMIB), University of Würzburg, Würzburg, Germany), Linda Popella (Institute for Molecular Infection Biology (IMIB), University of Würzburg, Würzburg, Germany), Phuong Thao Do (Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research, Würzburg, Germany), Patrick Pfau (Faculty of Medicine, University of Würzburg, Würzburg, Germany), Jörg Vogel (Institute for Molecular Infection Biology (IMIB), University of Würzburg, Würzburg, Germany) and Lars Barquist (Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research, Würzburg, Germany).

Abstract:

Antisense oligomers (ASOs) such as peptide nucleic acids (PNAs), designed to inhibit the translation of essential bacterial genes, have emerged as attractive sequence- and species-specific programmable RNA antibiotics. Yet, potential drawbacks include unwanted side effects caused by their binding to transcripts other than the intended target.

To facilitate the design of PNAs with minimal off-target effects, we developed MASON (Make AntiSense Oligomers Now), an open-source webserver for the design of PNAs that target bacterial mRNAs. MASON generates PNA sequences complementary to the translational start site of a bacterial gene of interest and monitors critical sequence attributes, such as melting temperature while also screening for potential off-target sites. We tailored MASON's off-target predictions based on experiments in which we treated Salmonella enterica serovar Typhimurium with a series of 10mer PNAs that are based on a widely-used PNA targeting the essential gene acpP but carry two serial mismatches from the N- to the C-terminus.

Growth inhibition and RNA-sequencing (RNA-seq) data revealed that PNAs with terminal mismatches are still able to target acpP, suggesting wider off-target effects than anticipated. Comparison of these results to a published RNA-Seq dataset from uropathogenic Escherichia coli (UPEC) treated with eleven different PNAs confirmed that these off-target effects are not unique to Salmonella. We believe that this off-target assessment, which is incorporated into MASON, will improve the design of specific PNAs and other ASOs and hope that the webserver will become an evolving resource for the bacterial ASO community. The MASON website can be freely accessed at https://www.helmholtz-hiri.de/en/datasets/mason.

Genes

Predicting Transcription Factors Biological Roles in P. vulgaris

Liudmyla Kondratova (University of Florida), Eduardo Vallejos (University of Florida) and Ana Conesa (Spanish National Research Council).

Abstract:

Phaseolus vulgaris or common bean is a model organism for studying the evolution of crops due to the existence of two gene pools developing independently for more than 165,000 years. However, little is known about the regulation of genes involved in evolutionary and agriculturally important traits, which limits our understanding of regulatory processes leading to domestication. Studying the differences in regulation of traits such as starch accumulation or oligosaccharides biosynthesis between wild and domesticated accessions is important to building effective breeding programs.

Here, we combined comparative genomics, affinity motif scanning, and functional profiling to infer a functionally informed Transcription Factor regulatory network in common bean. We screened homologous promoters of orthologous genes using an adaptation of a conservation test which is anticipated to select evolutionary conserved transcription factor binding sites. Thus, we were able to identify evolutionary conserved motifs between P. vulgaris and phylogenetically close species such as G. max, V. angularis, and V. radiata. Additionally, we used an affinity-based motif scanning approach to identify transcription factor binding sites that could have been missed by the conservation test. Biological roles of transcription factor families were assigned based on enriched functional annotations within genes with conserved binding motifs to a specific TF family and a regulatory network was constructed based on shared functional roles. We validated our predictions using available knowledge about the starch biosynthesis pathway in plants.

Genes

Rational design of a novel respiratory syncytial virus reporter of early events of its viral cycle

Marcio Andrés De Ávila (Fundación Universidad del Norte), Jose Luis Villarreal Camacho (Universidad Libre), Christian Cadena (Fundación Universidad del Norte), Leidy Hurtado Gómez (Fundación Universidad del Norte), Laura Piñeres Santos (Universidad del Atlántico), Amner Muñoz Acevedo (Fundación Universidad del Norte) and Homero San Juan Vergara (Fundación Universidad del Norte).

Abstract:

The Respiratory Syncitial Virus (RSV) is the most common cause of lower respiratory tract infection in children that generate significant morbidity and mortality, nevertheless, there is not an effective antiviral treatment or vaccine developed yet, the viral process as fusion, transcription and replication are important targets against RSV, the above highlight the growing needs for new tools to reveal molecular mechanism in the antiviral activity, we have developed a simple and fast assay for the detection of the successfully fusion of RSV in the NHBE cells. We generated a three-dimensional model for the protein P and BlaM (Uniport code P03422 and C5I4X2, respectively) using a Colab server with a slightly simplified version of AlphaFold v2.0 which omit existing structural templates. The reliability of the predictions was assessed trough the score of local differences distance test (LDDT) inform for each structure, additionally, we compare the in silico protein P-BlaM against the experimental crystallographic structure of P-protein (PDB: 6PZK). This approach provide a platform for recovering a recombinant RSV, that contains the P-BlaM

Gene next to the RSV-P gen and expresses the P-BlaM reporter protein that cleave the Beta-lactam ring, as a reference assay we use CCF2-AM, which consist of 2 florescent molecules linked by a Beta-lactam ring that is cleave once the reporter protein enter to the cell due to the virus fusion, generating a fluorescente detection by flow cytometry, florescence microscopy or UV photometry.

Genes

RBPNet: Predicting Protein-RNA Interaction via CLIP-Seq Sequence-to-Signal Learning

Marc Horlacher (Helmholtz Center Munich), Nils Wagner (Technical University Munich), Marco Salvatore (University of Copenhagen), Julien Gagneur (Technical University Munich), Lambert Moyon (Helmholtz Center Munich), Ole Winther (University of Copenhagen) and Annalisa Marsico (Helmholtz Center Munich).

Abstract:

RNA-binding proteins (RBPs) are crucial actors of post-transcriptional regulation and have been associated with an abundance of human diseases, in particular nervous system diseases and psychiatric disorders. Uncovering binding preferences and RNA targets of RBPs is crucial for understanding the role of RBPs in regulatory pathways and for quantifying the impact of their dis-regulation in the context of human disease. Experimental protocols such as Cross-Linking Immunoprecipitation followed by Sequencing (CLIP-Seq) and its derivatives allow for accurate transcriptome-wide profiling of protein-RNA interaction at near single-nucleotide resolution. However, only a fraction of transcripts may be expressed in the experimental cell type at a given time, creating the need for imputation via computational methods.

We present RBPNet, a deep learning method which learns a direct mapping of RNA sequence to CLIP-seq signal and predicts the distribution of protein-RNA crosslinking events at single-nucleotide resolution. Using a large cohort of eCLIP experiments, we demonstrate that RBPNet reaches replicate-level performance on the majority of datasets. By modelling the control signal as an auxiliary task, RBPNet accounts for potential experimental CLIP-seq biases and approximates the unobserved protein-specific signal. Through model interrogation, RBPNet identifies highly predictive sub-sequences corresponding to known and novel RBP binding motifs. Finally, RBPNet enables impact-scoring of sequence variants on protein-RNA interaction, thus prioritising SNPs that may disrupt post-transcriptional processes.

We believe that by capturing the full bandwidth of experimental variance, RBPNet represents a significant advancement over previous, classification-based approaches.

Genes

ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments

Fayrouz Hammal (TAGC INSERM U1090 AMU), Benoit Ballester (TAGC INSERM U1090 AMU), Pierre De Langen (TAGC INSERM U1090 AMU), Fabrice Lopez (Aix Marseille Univ, INSERM, UMR U1090, TAGC, Marseille, France) and Aurélie Bergon (TAGC INSERM U1090 AMU).

Abstract:

ReMap is a resource of transcriptional regulators binding regions available in four different species: Arabidopsis thaliana, Mus musculus, Drosophila melanogaster, Homo sapiens. Our work aims to combine all publicly available ChIP-seq of transcriptional regulators (TRs) and create a catalog of manually curated and uniformly processed datasets. By its size and complexity, the ReMap catalogs allow a better understanding of the regulatory landscape. Currently we have annotated, curated, and uniformly processed a total of 8,103 ChIP-seq datasets for the human genome covering a total of 1,210 transcriptional regulators across 182 million peaks. The mouse catalog also contains a large amount of experiments with 5,503 datasets of 648 transcriptional regulators across 123 million of peaks.

For the four species available in our ReMap catalog the same steps were used. The ChIP-seq datasets were retrieved from GEO and ENCODE. To create this regulatory atlas, we have manually curated and uniformly processed the ChIP-Seq datasets. Due to the heterogeneity of datasets, the pipeline assesses the quality of the datas and filters them accordingly . The pipeline used for the processing, filtering and control quality is available on github (https://github.com/remap-cisreg) as a snakemake pipeline.

The four regulatory catalogs are available at https://remap.univ-amu.fr in bed format. The ChIP-seq peaks of the datasets are also available for each of the biotypes and transcription factor. The datasets are also browsable as native tracks in UCSC genome browser. Complex filtering features on targets or biotypes can be applied to improve visualization of ReMap peaks.

Genes

rfPhen2Gen: A machine learning based association study of brain imaging phenotypes to genotypes

Muhammad Ammar Malik (University of Bergen), Alexander Selvikvåg Lundervold (Western Norway University of Applied Sciences) and Tom Michoel (University of Bergen).

Abstract:

Imaging genetic studies aim to find associations between genetic variants and imaging quantitative traits. Traditional genome-wide association studies (GWAS) are based on univariate statistical tests, but when multiple traits are analyzed together they suffer from a multiple-testing problem and from not taking into account correlations among the traits. An alternative approach to multi-trait GWAS is to reverse the functional relation between genotypes and traits, by fitting a multivariate regression model to predict genotypes from multiple traits simultaneously. However, current reverse genotype prediction approaches are mostly based on linear models. Here, we evaluated random forest regression (RFR) as a method to predict SNPs from imaging QTs and identify biologically relevant associations. We learned machine learning models to predict 518,484 SNPs using 56 brain imaging QTs. We observed that genotype regression error is a better indicator of permutation p-value significance than genotype classification accuracy. SNPs within the known Alzheimer disease (AD) risk gene APOE had lowest RMSE for lasso and random forest, but not ridge regression. Moreover, random forests identified additional SNPs that were not prioritized by the linear models but are known to be associated with brain-related disorders. Feature selection identified well-known brain regions associated with AD, like the hippocampus and amygdala, as important predictors of the most significant SNPs. In summary, our results indicate that non-linear methods like random forests may offer additional insights into phenotype-genotype associations compared to traditional linear multi-variate GWAS methods.

Genes

RNA-seq analysis: primary breast tumour and circulating tumour cells

Dominik Hadzega (Medirex Group Academy n.p.o., Nitra / Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava), Gabriel Minarik (Medirex Group Academy n.p.o., Nitra), Andrea Soltysova (Biomedical Research Center SAS, Bratislava / Department of Molecular Biology, FNS, Comenius University, Bratislava), Petra Nemcova (Medirex Group Academy n.p.o., Nitra), Katarina Kalavska (Translational research unit, Faculty of Medicine, Comenius University and National Cancer Institute, Bratislava), Marian Karaba (Department of Oncosurgery, National Cancer Institute, Bratislava), Juraj Benca (Department of Oncosurgery, National Cancer Institute / Department of Medicine, St. Elizabeth University, Bratislava), Tatiana Sedlackova (Institute of Molecular Biomedicine, Faculty of Medicine, Comenius University, Bratislava), Daniel Pindak (Department of Oncosurgery, National Cancer Institute, Bratislava), Lubos Klucar (Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava) and Michal Mego (2nd Department of Oncology, Faculty of Medicine, Comenius University and National Cancer Institute, Bratislava).

Abstract:

Circulating tumour cells (CTC) are cells with origin in tumour tissue, which are able to leave their original tissue and travel to distant organs, where they can establish new metastasis. Epithelial to mesenchymal transition (EMT) is believed to influence cells' ability to become CTC. In our research we studied gene expressions in breast tumour and normal breast tissue of breast cancer patients. We investigated which genes expressions are altered for tissue of patients with CTC EMT (positive on mesenchymal markers) in their blood. Gene expressions were obtained by whole transcriptome RNA-sequencing of the fresh frozen primary tumour samples. The study included 18 patients with primary breast cancer and 5 donors of normal breast tissue. We used standard procedure for differentially expressed genes analysis (using tools in Galaxy and R environments) and then used set of differentially expressed genes for gene enrichment analysis in purpose of relevant pathways identification. From RNA-seq analysis, we found 70 genes to be up-regulated and 17 to be down-regulated, under the conditions of adjusted p-value<0.1 and log2FC>1. Downregulated genes were showed to be related to immunology, while up-regulated genes to various signalling pathways and cell-cell interactions. This research was funded by APVV-16-0010 project and the OPII programme as the project PROMEDICOV-19, code ITMS: 313011ATA2, co-financed by the ERDF.

Genes

Sex differences in the molecular basis of multiple sclerosis: meta-analysis of transcriptomic data

Jose Francisco Català-Senent (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF)), Zoraida Andreu (Foundation Valencian Institute of Oncology (FIVO)), Marta R. Hidalgo (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF)), Roig Francisco José (Faculty of Health Sciences. San Jorge University), Natalia Yanguas-Casás (Instituto de Investigación Sanitaria Puerta de Hierro-Segovia de Arana (IDIPHISA), Grupo de Investigación en Linfomas), Almudena Neva-Alejo (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF)), Adolfo López-Cerdán (Biomedical Imaging Unit FISABIO-CIPF), Irene Soler-Sáez (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF)), María de la Iglesia-Vayá (Biomedical Imaging Unit FISABIO-CIPF) and Francisco García-García (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF)).

Abstract:

Multiple sclerosis (MS), an auto-immune, inflammatory, and degenerative disorder of the central nervous system, affects both males and females; however, females are at an increased risk of developing MS (2-3:1 ratio compared to males). The factors behind these sex differences are not still clear. Therefore, the aim of this work has been to explore the role of sex in MS to identify potential molecular mechanisms underlying sex-based differences.

To this end, we performed a systematic review in public databases of transcriptomic studies, in nervous tissue and blood. Next, we performed 3 meta-analyses that allowed us to detect MS alterations in females, in males and between both sexes.

As a result of our work, we selected 9 studies (4 of nerve tissue and 5 of blood) containing 474 individuals. Our meta-analyses identified some genes and functions altered on a sex-specific or between-sex basis. Among them, highlight 15 genes that showed a significantly different expression pattern between sexes in some of the tissues analyzed: (KIR2DL3 in blood; ARL17B, CECR7, CEP78, IFFO2, LOC401127, NUDT18, RNF10, SLC17A5, STEMP1, TRAF3IP2-AS1, UBXN2B, ZNF117, ZNF488 in nervous tissue; LOC 102723701 in both tissues).

This work evidences the existence of sex differences in MS at the transcriptomic level and, moreover, open the door to future applications leading to more sex-specific treatments.

Genes

Shigella-type IpaH ubiquitin ligases in non-human-host enteroinvasive Escherichia

Olga Bochkareva (IST Austria), Natalia Dranenko (IITP RAS) and Maria Tutukina (IITP RAS).

Abstract:

Until recently, Shigella and enteroinvasive Escherichia coli were thought to be primate-restricted pathogens. The base of their pathogenicity is the type 3 secretion system (T3SS) encoded by the pINV virulence plasmid, which facilitates host cell invasion and subsequent proliferation. A large family of T3SS effectors, E3 ubiquitin-ligases encoded by the ipaH genes, have a key role in the Shigella pathogenicity through the modulation of cellular ubiquitination that degrades host proteins. However, recent genomic studies identified ipaH genes in the genomes of Escherichia marmotae, a potential marmot pathogen, and an E. coli extracted from fecal samples of bovine calves, suggesting that non-human hosts may also be infected by these strains, potentially pathogenic to humans. We performed a comparative genomic study of the functional repertoires in the ipaH gene family in Shigella and enteroinvasive Escherichia from human and predicted non-human hosts. Non-human host IpaH proteins had a diverse set of substrate-binding domains and, in contrast to the Shigella proteins, two variants of the NEL C-terminal domain. Inconsistencies between strains phylogeny and composition of effectors indicate horizontal gene transfer between E. coli adapted to different hosts. These results provide a framework for understanding of ipaH-mediated host-pathogens interactions and suggest a need for a genomic study of fecal samples from diseased animals.

Genes

SNEEP: SNP exploration and functional analysis using epigenomics data

Nina Baumgarten (Institute for Cardiovascular Regeneration), Chaonan Zhu (Institute for Cardiovascular Regeneration), Ting Yuan (Institute for Cardiovascular Regeneration), Meiqian Wu (Institute for Cardiovascular Regeneration), Minh Duc Pham (Institute for Cardiovascular Regeneration), Despina Stefanoska (Institute for Cardiovascular Regeneration), Thorsten Kessler (German Heart Centre Munich), Stefanie Dimmeler (Institute for Cardiovascular Regeneration), Jaya Krishnan (Institute for Cardiovascular Regeneration) and Marcel H. Schulz (Institute for Cardiovascular Regeneration).

Abstract:

Genome-wide association studies (GWAS) indicate that most single nucleotide polymorphisms (SNPs) appear in non-coding genomic regions. These SNPs may alter gene expression by interrupting Transcription Factor Binding Sites (TFBS) and lead to functional consequences like various traits or diseases. To understand the underlying molecular mechanisms, it is crucial to identify which variations are involved and how they affect TF binding.

Our SNEEP workflow prioritizes SNPs as targets of Transcription Factors (TFs) and infer whether a gene's expression is influenced by the change in the TF binding behavior. The impact of a SNP to a potential TFBS is evaluated by calculating a probabilistic differential binding score for the difference in TF binding in wild type versus a mutated sequence. To associate SNPs to potential target genes, SNEEP can use gene-regulatory elements based on Hi-C data or catalogued elements. SNEEP easily handles large collections of SNPs and allows to incorporate customized epigenetic data. A novel background sampling approach is applied to adjust for variable allele frequencies of SNPs. SNEEP provides a comprehensive report supporting TF- and gene-centric analyses.

We illustrate the application of SNEEP on cardiovascular GWAS from the EBI GWAS catalogue, identifying a number of long-noncoding RNAs (lincRNAs) associated with SNPs in gene-regulatory elements. Using a heart organoid model, we find that lincRNA IGBP1P1 drives cardiac hypertrophy and contractile dysfunction.

Summarizing, SNEEP prioritize GWAS SNPs to study the impact of genetically induced transcriptional mis-regulation in human diseases and other phenotypes and shows great promise to identify lincRNAs involved in cardiovascular diseases.

Genes

Statistical test to detect equivalence between feature lists based on the Sorensen-Dice index and GO term enrichment

Pablo Flores (Escuela Superior Politecnica de Chimborazo (ESPOCH); Research Group in Data Science CIDED), Jordi Ocaña Rebull (Departament de Genetica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona), Miquel Salicrú Pages (Departament de Genetica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona) and Alex Sanchez (Departament de Genetica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona).

Abstract:

Modern omics technologies yield large amounts of biological data related with the biological problem being studied that can be described as “feature lists”, where this term is used to name, genes, proteins or other molecular characteristics. A common goal of many bioinformatic analysis is establishing biological similarity between these lists, and methods such as goProfiles allow to do the comparison based on the projection of the terms in an ontology such as the GO. In this work we present a recently published test of equivalence, that extends the previous method so that it can be used on the results of enrichment analysis instead of using all the annotations. The test is posed as an equivalence test, where “equivalence” is understood as “equality” except for irrelevant deviations.

The test starts from a contingency table summarizing the enriched and non-enriched GO terms of the feature lists to be compared. A Sorensen dissimilarity, dS, is computed to measure the degree of mutual enrichment. The fundamental idea is that two feature lists will be called “similar” if they share a sufficiently high proportion of common enriched GO terms. We have derived the sampling distribution of dS, and used it to build an equivalence test with null hypothesis “dS>=d0” vs alternative “dS<d0”. That is, if dS is not big enough ("enough": >= d0) we can conclude that the dissimilarity is equivalent to zero, which means biological similarity between compared lists.

The test is implemented in the goSorensen R-package (submitted to Bioconductor).

Genes

Studying the dynamics of relative RNA localization - From nucleus to the cytoplasm

Vasilis F. Ntasis (Centre for Genomic Regulation (CRG)) and Roderic Guigó Serra (Centre for Genomic Regulation (CRG), Universitat Pompeu Fabra (UPF)).

Abstract:

The precise coordination of important biological processes, such as differentiation and development, is highly dependent on the regulation of expression of the genetic information. The flow of the genetic information is tightly regulated on multiple levels. Among them, RNA export to cytoplasm is an essential step for the production of proteins in eukaryotic cells. Hence, estimating transcript relative localization, that is the proportion of total RNA molecules in eukaryotic cells present in the cytosol or in the nucleus, is of major significance. However, most studies with a focus on transcriptome analysis ignore subcellular RNA localization. Those with an effort to take that into account, utilize RNA sequencing (RNA-seq) in combination with cellular fractionation. Nevertheless, transcript quantification estimates obtained independently from nuclear and cytosolic RNA cannot be compared (as the total amount of RNA in each of these cellular compartments is usually unknown). Here we show that if, in addition to nuclear and cytosolic RNA-seq, whole cell RNA-seq is also performed, then accurate estimations of the relative localization of transcripts can be obtained. We first establish the theoretical basis that supports this by formalizing mathematically the relationship between the different RNA abundances. Based on this, we designed a method that estimates for every transcript its relative localization. Then, we evaluate our methodology on simulated and real bulk RNA-seq data from the ENCODE project, and from a time course differentiation process.

Genes

T-cell receptors as biomarkers for early diagnosis of colorectal cancer in Lynch Syndrome

Maria S. Benitez-Cantos (Centro Pfizer-Universidad de Granada-Junta de Andalucía de Genómica e Investigación Oncológica (GENYO)), Sonia Garcia-Rodriguez (Centro Pfizer-Universidad de Granada-Junta de Andalucía de Genómica e Investigación Oncológica (GENYO)), Marta Cuadros (Centro Pfizer-Universidad de Granada-Junta de Andalucía de Genómica e Investigación Oncológica (GENYO)), Carlos Cano (Universidad de Granada), Antonio Poyatos-Andújar (Hospital Universitario San Cecilio), Carmen Sánchez-Toro (Hospital Virgen de Las Nieves), Pilar Carrasco-Salas (Hospital Universitario Juan Ramón Jiménez), Ana M. Serrano-Mira (Hospital Universitario Juan Ramón Jiménez), Carmelo Diéguez-Castillo (Hospital Universitario Torrecárdenas), Ana Delgado-Maroto (Hospital Universitario Torrecárdenas) and Paul Lizardi (Centro Pfizer-Universidad de Granada-Junta de Andalucía de Genómica e Investigación Oncológica (GENYO)).

Abstract:

Lynch syndrome (LS) confers an 80% lifetime risk of developing colorectal cancer (CRC). It results from germline mutations in the DNA mismatch repair genes, which cause an accumulation of frameshift mutations in microsatellites, even before the onset of cancer. These neoantigens can be recognized by T-cells in premalignant lesions through their antigen-specific T-cell receptors (TCRs), making LS an ideal model to study the feasibility of early cancer diagnosis by detecting neoantigen-specific TCRs in peripheral blood. In this study, we performed a TCR-Seq of circulating CD4+ T-cells from 20 healthy LS mutation carriers and 19 non-carriers to identify neoantigen-specific TCRs that might be biomarkers for early CRC diagnosis. We generated a dataset of ~3.9M TCRs, that we complemented with ~2.9K TCRs extracted from public RNA-Seq data of 125 LS-CRC tumor biopsies (caTCRs) and ~10M TCRs from peripheral blood of an existing cohort of 666 healthy donors. To identify putative neoantigen-specific TCRs, we run a large-scale sequence clustering with all the datasets and extracted 89 clusters composed exclusively of caTCRs and carrier TCRs. Of those, we selected 7 clusters as candidate LS-neoantigen biomarkers based on the number of carriers contributing to them and the higher relative frequency of those TCRs in their repertoires, which may indicate T-cell activation and antigen recognition. We also predicted which LS-neoantigens are more likely to be presented by MHC class II alleles and we plan to test them against our 7 TCR candidates in a multiplexed T-cell receptor antigen specificity assay for an experimental validation.

Genes

The adapted Activity-By-Contact model for enhancer-gene assignment and its application to single-cell data

Dennis Hecker (Goethe University Frankfurt), Fatemeh Behjati Ardakani (Goethe University Frankfurt) and Marcel Schulz (Goethe University Frankfurt).

Abstract:

Identifying regulatory regions in the genome is of great interest for understanding the epigenomic landscape in cells. One fundamental challenge in this context is to find the target genes whose expression the regulatory regions affect. A recent successful method is the Activity-By-Contact (ABC) model (Fulco et al., Nature Genetics 51 (2019)) which scores enhancer-gene interactions based on enhancer activity and the contact frequency of an enhancer to its target gene. It requires two types of assays to measure enhancer activity, which limits the applicability. The ABC score describes regulatory interactions entirely from a gene's perspective, and does not account for all the candidate target genes of an enhancer. Moreover, there is no implementation available which would allow for an integration with transcription factor (TF) binding information or an efficient analysis of single-cell data.

We show that the ABC-model is comparably accurate with only one assay to measure enhancer activity. Further, we demonstrate that it can yield a higher accuracy by adapting the enhancer activity according to the number of contacts the enhancer has to its candidate target genes. We combined our adapted ABC-model with TF binding information and illustrate an analysis of a single-cell ATAC-seq data set of the human heart (Hocker et al., Science Advances 7 (2021)). We were able to characterise cell type-specific regulatory interactions as well as to prioritise candidate TFs that drive cell type-specific expression. All executed processing steps are incorporated into our new computational pipeline STARE (https://github.com/schulzlab/STARE).

Genes

The Enrichment of Genes Associated with T-cell Proliferation and Memory Formation of B-cell Co-stimulatory Domain Potentiate CD19CAR-T cell Functions

Socheatraksmey Ung (Prince of Songkla University), Jakrawadee Julamanee (Prince of Songkla University), Wannakorn Khopanlert (Prince of Songkla University), Kajornkiat Maneechai (Prince of Songkla University), Surasak Sangkhathat (Prince of Songkla University) and Pongsakorn Choochuen (Prince of Songkla University).

Abstract:

CD28 or 4-1BB co-stimulated CD19 chimeric antigen receptor (CAR) T-cell has demonstrated remarkable outcomes in B-cell malignancies. Recently, the novel CD19CAR incorporated B-cell co-stimulatory molecules, CD79A/CD40, has demonstrated the superior anti-tumor activity in B-cell lymphoma model compared to CD28 or 4-1BB. Here, we assessed mRNA expression using RNA-sequencing (RNA-seq) to interrogate the intrinsic transcriptional gene underlying CD19.79a.40z CAR-T cell response and identify the differentially expressed genes (DEGs) compared to CD19.28z or CD19.BBz CAR-T cell following CD19-specific antigen exposure. A total of 374 DEGs (232 up-regulated and 142 down-regulated DEGs) and 1 up-regulated DEG were identified in CD19.79a.40z vs CD19.28z CAR-T cell groups and CD19.79a.40z vs CD19.BBz CAR-T cell groups respectively. Functional enrichment analysis illustrated that CD19.79a.40z CAR-T cells consisted of down-regulated genes mediating apoptosis and up-regulated genes correlated with T-cell activation, proliferation, positive regulation of interferon production, and NF-κB pathway. Comparing to either CD19.28z or CD19.BBz CAR-T cells, CD19.79a.40z CAR-T cells were strongly enriched in genes associated with T-cell proliferation at the transcription level supported by the expression of IL2, ZP3, TFRC, IRF1 and CD70. Regarding metabolic pathway, GSEA revealed that CD19.79a.40z CAR-T cells were enriched in both glycolysis and fatty acid metabolism comparing to CD19.BBz CAR-T cells. In terms of T-cell differentiation, CD19.79a.40z CAR-T cells were particularly enriched in naïve and memory-related genes compared to CD19.BBz CAR-T cell. In conclusion, this study provides a comprehensive insight into the understanding of gene expression and its related pathways that potentiates the superior anti-tumor functions by CD19CAR-T cell incorporated B-cell co-stimulatory domain.

Genes

The least diverged orthologue conjecture

Alex Warwick Vesztrocy (University of Lausanne).

Abstract:

The orthologue conjecture - that orthologous genes are functionally more similar than paralogous genes - has been the subject of much debate. However, annotation bias leads to issues when studying the orthologue conjecture in previous studies using gene ontology annotations.

In this study the gene families from the PANTHER database [1] are combined with expression data from the Bgee database [2]. Using expertly curated ancestral anatomical entity similarity annotations [3], this enables the reconstruction of ancestral gene expression profiles. Then, with the application of a simple evolutionary model to allow for different rates of gene family evolution, it is possible to compare sequence evolution with changes in gene expression profile.

This enables the investigation of a specific case of the orthologue conjecture: that after gene duplication the least evolutionary diverged copy maintains the ancestral function, whilst the other copy is no longer under selective pressure to maintain function and is free to diverge. We call this the least diverged orthologue conjecture.

[1] Mi, Huaiyu, et al. "PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API." Nucleic acids research 49.D1 (2021): D394-D403.

[2] Bastian, Frederic B., et al. "The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals." Nucleic Acids Research 49.D1 (2021): D831-D847.

[3] “Anatomical Similarity Annotations.” BgeeDB, github.com/BgeeDB/anatomical-similarity-annotations. Accessed 30 June 2022.

Genes

The Multilayer Community Structure of Medulloblastoma

Iker Núñez Carpintero (Barcelona Supercomputing Center), Marianyela Petrizzelli (Institut Curie), Andrei Zinovyev (Institut Curie), Davide Cirillo (Barcelona Supercomputing Center) and Alfonso Valencia (Barcelona Supercomputing Centre BSC).

Abstract:

Biomedical multilayer networks offer a wide range of possibilities for the interpretation of the molecular basis of diseases; a particularly challenging task in the case of rare diseases, where the number of cases is small in comparison with the size of the associated multi-omics datasets. In this work, we present a dimensionality reduction methodology to identify the minimal set of genes that characterize the clinical subgroups of medulloblastoma, a rare childhood brain tumor.

For that purpose, we constructed a multilayer gene graph that integrates general biomedical knowledge from 5 databases (Reactome, Recon3D Virtual Metabolic Human, BioGRID, KEGG BRITE "Target-based Classification of Compounds" and Monarch Disease Ontology) and performed a multilayer community trajectory analysis using the R package CmmD, that we implemented.

By applying CmmD, it is possible to consider co-existent community structures from different modularity resolution limits, as well as tracking different events throughout the associated process of network community decomposition. Such events can already be used as features for gene clustering or other machine learning tasks, such classification and prediction.

Our approach recapitulates known medulloblastoma subtypes (accuracy > 94%), offering a clear characterization of the functional landscape affected in each clinical subtype, with the downstream implications for diagnosis and therapeutic interventions.

We verified the general applicability of our method on an independent dataset, achieving very high performances (accuracy > 98%). Overall, our approach demonstrates the potential of multilayer-based methods to overcome the specific dimensionality limitations of rare disease datasets.

Genes

topRegNet: an analysis workflow for detecting enriched regulatory elements in RNA-seq datasets

Anshupa Sahu (University Hospital Bonn), Sugirthan Sivalingam (Core Unit for Bioinformatics Data Analysis, University Hospital Bonn), Farhad Shakeri (Core Unit for Bioinformatics Data Analysis, University Hospital Bonn), Svetozar Nesic (Core Unit for Bioinformatics Data Analysis, University Hospital Bonn) and Andreas Buness (Core Unit for Bioinformatics Data Analysis, University Hospital Bonn).

Abstract:

Understanding gene regulation is essential to better elucidate fundamental molecular processes underlying cell differentiation, disease progression and pathogenesis, etc. Enhancers and transcription factors are the two main regulators of gene expression. Enhancers are 50-1500 bp long cis-regulatory elements that regulate the expression of genes by recruiting transcription factors in a cell-specific manner. However, an analysis workflow for identifying tissue-specific regulatory elements in RNA-seq datasets is lacking.

We present topRegNet, an analysis workflow for tissue-specific identification and analysis of regulatory elements. Using existing public databases, topRegNet extracts enhancer and TF information for the tissue and genes of interest. It then scores the enhancers and transcription factors based on their frequency across databases and the quality of experimental evidence. Currently, topRegNet identifies enhancers and transcription factors, however, in the future we also plan to include other regulatory elements such as microRNAs.

For a given list of differentially expressed genes, topRegNet reports both high and low confidence enhancers and TFs that have been reported to regulate the genes for the given tissue. If the enhancer sequence is still unknown for a gene, topRegNet predicts the enhancer sequence based on the frequency of TF binding sites and reports it along with transcription factors. topRegNet can be included as a part of bulk as well as scRNA-seq downstream analysis. Hence, providing valuable insights regarding the transcriptome regulation landscape in disease and development.

Genes

Towards sex-specific polygenic risk scores

María Morales Martínez (Barcelona Supercomputing Center), Hira Shahid (Barcelona Supercomputing Center), Kathleen Imbach (Josep Carreras Leukaemia Research Institute and associated to Barcelona Supercomputing Center) and Eduard Porta Pardo (Josep Carreras Leukaemia Research Institute and associated to Barcelona Supercomputing Center).

Abstract:

A large number of germline variants associated with complex diseases have been identified by genome-wide association studies (GWAS). The results of such GWAS have oftentimes been used to create polygenic risk scores (PRS), which quantify the genetic predisposition of each individual towards the trait studied in the GWAS. However, the sex disparities in genetic predisposition in general and in cancer in particular, are still understudied. Here we first show how PRS obtained from GWAS with both males and females for glioma, kidney, thyroid, and colorectal cancer, show widely divergent incidences in males and females that have the same genetic risk. Next, we present GUIDANCE 2.0, an automated GWAS pipeline developed at BSC which, unlike existing approaches, is able to perform haplotype phasing, genotype imputation using multiple reference panels and association testing with different models of inheritance. Finally, we show how by using GUIDANCE 2.0 to re-analyze the genotypes from publicly available GWAS we can identify new sex-specific cancer risk variants that can then be used to create sex-specific PRS that have higher predictive power.

Genes

Transcription start site signal profiling improves transposable element expression analysis at loci level

Natalia Savytska (German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany), Peter Heutink (German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany) and Vikas Bansal (German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany).

Abstract:

In recent years, the transcriptional activity of Transposable Elements (TEs) became suspect in contributing to multiple degenerative pathologies such as amyotrophic lateral sclerosis and fronto-temporal lobar degeneration. However, measuring TEs activity remains a challenging task for short-read sequencing technologies due to erroneous mapping to multitude of highly similar genomic loci derived from singular TEs subfamilies. A number of specialised tools were developed to quantify the expression of TEs that either rely on probabilistic re-distribution of multimapper count fractions, or allow for discarding multimappers altogether. Nevertheless, no consensus in the field was reached as for the best performing strategy and benchmarking across those tools was largely limited to aggregated expression estimates over whole TEs subfamilies. In this study, we compared performance of recently published tools (SQuIRE, TElocal, SalmonTE) with simplistic quantification strategies (featureCounts in unique, fraction and random modes) at the individual loci level. Using simulated datasets, we examined the false discovery rate and its main driver in the optimal quantification strategy. Our findings suggest a high false discovery number, that exceeds the total number of correctly recovered active loci, for all of the quantification strategies, including the best performing tool TElocal. As a remedy, filtering on the minimum number of read counts improved the F1 score and significantly decreased the number of false positives. Furthermore, we demonstrated that additional profiling of Transcription Start Site (TSS) mapping statistics could greatly improve the performance of TElocal and allow for reliable TEs quantification even for lowly expressed elements.

Genes

Transcriptomic effects of cigarette smoking across human tissues

José Miguel Ramírez Cardeñosa (Barcelona Supercomputing Center), Marta Mele Messeguer (Barcelona Supercomputing Center), Rogério Ribeiro (University of Porto) and Pedro Ferreira (University of Porto).

Abstract:

Tobacco smoke is the leading modifiable cause of death worldwide as it increases the predisposition to develop many chronic diseases and types of cancers. Despite the relevance of this clinical trait, most transcriptomic studies have only focused on particular tissues like blood, airway epithelium or adipose tissue, and have ignored its combined effect with other demographic traits like ancestry or age. Here, we used the Genotype-Tissue Expression dataset to systematically analyze the effect of smoking on gene expression and alternative splicing across 46 human tissues and 838 individuals. Our analysis reveals a major dysregulation of the lung, thyroid, and esophagus, and a common transcriptomic signature across 26 tissues. We observed a big similarity between the effects of smoking and aging across tissues, suggesting that smoking contributes to tissue ageing. Furthermore, by using ex-smoker samples and machine learning, we identified most of the smoking effects as reversible or partially reversible upon smoking cessation. Overall, our multi-tissue approach characterises tobacco smoking effects in new organs and suggests a common signature across tissues that can be reversed upon smoking cessation.

Genes

TranSNPs: A class of functional SNPs affecting mRNA translation potential revealed by fraction-based allelic imbalance

Samuel Valentini (University of Trento, CIBIO), Caterina Marchioretti (University of Padova, DBS), Alessandra Bisio (University of Trento, CIBIO), Annalisa Rossi (University of Trento, CIBIO), Sara Zaccara (Weill Medical College, Cornell University), Meriem Hadjer Hamadou (University of Trento, CIBIO), Elisa Pettinà (University of Trento, CIBIO), Giacomo Fantoni (University of Trento, CIBIO), Alberto Inga (University of Trento, CIBIO) and Alessandro Romanel (University of Trento, CIBIO).

Abstract:

Single Nucleotide Polymorphisms (SNPs) are the largest class of human genetic variations and are believed to be one of the most significant contributors to inter-individual phenotype variations.

SNPs can modify allelic gene expression by altering epigenetic or transcriptional regulatory elements or impacting post-transcriptional, translational, or post-translational processes.

We developed an approach to identify SNPs that can mark allele-specific mRNA translation potential and could represent sources of inter-individual variation in disease risk.

Using MCF7 cells and mock, doxorubicin, and Nutlin treatments, we performed polysomal profiling followed by RNA-sequencing of both total and polysome-associated mRNA fractions. Taking advantage of SNPs that are heterozygous in the MCF7 genome, we designed and implemented a computational approach to identify SNP alleles that show a significant change in the allelic balance between total and polysomal mRNA fractions.

We identified 147 SNPs imbalanced by our criteria, 39 of which located in UTRs that were considered further. Allele-specific differences at the translation level were confirmed in transfected MCF7 cells by reporter constructs containing either SNP allele within cloned UTR regions of the CDKN1A, ATF6, and BRIP3BP genes. Exploiting SNP linkage disequilibrium data and clinical data of a large cohort of Breast Cancer patients from TCGA we identified 33 UTR SNPs that clustered patients for distinct prognosis features, a subset of which was also predicted to alter binding sites of RNA binding proteins.

Our approach produced a catalog of tranSNPs, a new class of functional SNPs associated with allele-specific translation and potentially endowed with prognostic value for disease risk.

Genes

Using multi-omics to identify predictive markers of aggressive neuroendocrine cancer

Dimitria Brempou (King's College London), Louise Izatt (Guy's and St Thomas' NHS Foundation Trust), Cynthia Andoniadou (King's College London) and Rebecca Oakey (King's College London).

Abstract:

Pheochromocytomas (PCC) and Paragangliomas (PGL), collectively referred to as PPGLs, are rare neuroendocrine tumours associated with genetic pathogenic variants. More than 40% of PPGLs are inherited, but only some variant carriers go on to develop the disease with a subset presenting an aggressive phenotype with devastating consequences. So far, no biomarkers of aggressive phenotype have been identified.

To address the lack of predictive tools for disease progression, I study the transcriptome and DNA methylation of PPGLs caused by variants in the genes SDHA, SDHB, SDHC, SDHD and SDHAF2, collectively known as SDHx. SDHB mutations are associated with higher rates of aggressive disease compared to other known mutations and are, therefore, the main focus of my research. Preliminary results indicate that there is an epigenetic signature of aggressive PPGL. Biomarkers of aggressive phenotype will enable personalised monitoring and treatment for patients and variant carriers.

The aim of this interdisciplinary project is to deploy advanced computational methods to decode the molecular mechanism of aggressive disease progression. Previous research on PPGL contains statistical analysis of the transcriptome and DNA methylation of PPGL tumours independently. However, DNA methylation is an epigenetic mechanism regulating gene expression and, therefore, the two are interconnected. To understand the mechanism of aggressive disease, we will integrate EPIC arrays and RNA sequencing data and study the system alterations in aggressive PPGLs through networks. Moreover, we will apply advanced machine and transfer learning techniques to reveal the complex signature of aggressive disease. This approach will unlock new insights into PPGL phenotype.

Genes

Variability of nonsense-mediated mRNA decay (NMD) pathway efficiency in human cancers

Guillermo Palou Márquez (Institute for Research in Biomedicine (IRB Barcelona)) and Fran Supek (Institute for Research in Biomedicine (IRB Barcelona)).

Abstract:

The nonsense-mediated mRNA decay (NMD) pathway is a critical mRNA surveillance mechanism responsible for the degradation of transcripts containing premature termination codons (PTCs), reducing the production of potentially harmful truncated proteins. Interestingly, the NMD efficiency (NMDeff) can vary between PTCs and transcripts due to distinct NMD-eliciting features. This includes the position of the PTC along the gene, or other PTC-unrelated features that apply to endogenous natural NMD target transcripts such as having upstream open reading frames (uORFs) or splice sites in 3'UTR. There is some evidence anticipating that NMDeff can also vary across human individuals, reported in small-scale studies focussing on several cell lines, tissues or genetic diseases. Here, we present a systematic quantification of NMDeff variability across 33 tumor types and ~10.000 individuals, and validating the results in 54 normal tissues. We use matched whole-exome sequencing (WES) and RNA-seq data from the TCGA and GTEx databases to estimate a NMDeff value per individual using three independent statistical methods: i) PTCs-containing transcript levels; ii) endogenous natural NMD target transcript levels and iii) Allele-Specific Expression of PTCs. We show how NMDeff significantly varies across tumors and tissues. For instance, the microsatellite-instable (MSI) colorectal and uterus tumors have high NMDeff, suggesting that NMD may be needed to protect the cell from the high burden of frameshifted toxic peptides. In conclusion, we have implemented three statistical methods to quantify the NMDeff variability across individuals and tissues and detect some of its genetic underpinnings.

Genes

WASP - A versatile, web-accessible single cell RNA-seq processing platform

Andreas Hoek (Justus Liebig University Giessen), Katharina Maibach (Justus Liebig University Giessen), Ebru Özmen (Justus Liebig University Giessen), Torsten Hain (Justus Liebig University Giessen), Susanne Herold (Justus Liebig University Giessen) and Alexander Goesmann (Justus Liebig University Giessen).

Abstract:

Single cell RNA sequencing (scRNA-seq) enables exploration of cellular transcriptomes in an unprecedented resolution. Compared to traditional bulk RNA sequencing, analysis of gene expression profiles at the single cell level allows e.g., the identification of previously undiscovered rare cell populations with corresponding marker genes, detection of heterogeneity between cells of the same type or to follow transcriptional programs of cells during differentiation. Current single cell protocols combined with high-throughput sequencing techniques yield enormous amounts of data of up to hundreds of thousands of cells per experiment that need to be analyzed. Hence, there is need for bioinformatic analysis solutions adapted to the specific challenges deriving from scRNA-seq data.

To meet these challenges, we have implemented WASP - a versatile application designed for the management, analysis and interpretation of scRNA-seq high-throughput data. The software was published in 2021 and addresses all aspects from initial quality control, demultiplexing and reference alignment to downstream statistical evaluation. It can be used with any gene expression matrix or also with raw data derived from various single cell protocols such as 10x, ddSeq or Dolomite Nadia. Furthermore, it provides an automated workflow suitable for both non-bioinformaticians as well as experts. The software is available via a web-based interface and supports deployment to local as well as cloud-based compute infrastructures and is also freely available as web service. In the past, WASP has already successfully been applied to various data sets derived from different organisms and tissue types.

Genes

κ-velo improves single-cell RNA-velocity estimation

Valérie Marot-Lassauzaie (Berlin Institute for Medical Systems Biology, Max Delbrück Center in the Helmholtz Association, Berlin, Germany), Brigitte Bouman (Berlin Institute for Medical Systems Biology, Max Delbrück Center in the Helmholtz Association, Berlin, Germany), Fearghal Donaghy (Berlin Institute for Medical Systems Biology, Max Delbrück Center in the Helmholtz Association, Berlin, Germany) and Laleh Haghverdi (Berlin Institute for Medical Systems Biology, Max Delbrück Center in t

Abstract:

Single-cell transcriptomics has been used to study dynamical processes such as cell differentiation. RNA velocity (La Manno et. al. 2020) was a breakthrough towards obtaining a more complete description of the dynamics of such processes. Here, simultaneous measurement of new unspliced and old spliced mRNA adds a temporal dimension to the data. The change in mRNA abundance, called RNA velocity, is used to infer the progression of cells through the dynamical process. However, reliable velocity analysis is still impeded by multiple computational issues. State-of-the-art methods for velocity inference (Bergen et. al. 2020) have issues in velocity inference as well as visualisation. Moreover, there are inconsistencies in current processing pipelines and the single-cell specific (stochastic) part of the dynamic is lost through multiple layers of data smoothing.

We introduce a new method for RNA velocity analysis that addresses some of the issues in velocity estimation. We also propose that visualisation of the velocities based on the Nystroem projection method represents the single-cell stochasticity better than current practices. Finally, we adjust the processing pipeline for consistency with downstream velocity estimation. We validate our model on simulation and on real data, and compare it to current state-of-the-art.

Genomes

3D modeling of genome reorganization upon senescence

Vasiliki Varamogianni Mamatsi (Universitätsmedizin Göttingen) and Argyris Papantonis (Universitätsmedizin Göttingen).

Abstract:

Senescence is a key biological process contributing towards aging. Recent works support that chromatin architecture is perturbed when cells enter senescence and this event is accompanied by alteration of the trascriptional profile, depletion of HMGB proteins and formation of CTCF phase separated compartments. Chromosome Conformation Capture techniques have certainly contributed to the understanding of the dynamic nature of chromatin folding in mammalian nuclei, but the commonly used interpretation approaches, lack of information concerning actual radial positioning alterations of the genetic loci. Such changes refer to those between chromosomes or TADs. Here, we have created in silico simulations of the 3D genome rearrangement upon senescence, introducing TADs as main building domains of the polyemers. Those polymers are distributed in the nuclei in a 3D scale. Therefore we are able track different chromatin components before and upon senescence establishment. We demonstrate in proliferating cells that genomic regions that present high content of the High mobility group box 1 (HMGB1), cluster together and the chromatin that is drifted alongside contains Senescence Associated Secretory Phenotype (or SASP) genes and proinflammatory genes. Upon senescence this clustering is maintained, even after the depletion of HMGB1. We present on the 3D scale additionally, the localization of CCCTC-binding factor (CTCF) in proliferating and senescent nuclei as well as of the HMGB2. Our analysis provide insights into the insulating role of genome architecture proteins and trascriptional clustering.

Genomes

A computational model of double strand breaks and repair explaining cut-and-paste structural variants

Bingxin Lu (University College London) and Chris Barnes (University College London).

Abstract:

Somatic structural variants (SVs) are large genomic rearrangements involving at least 50 nucleotides and play an important role in cancer evolution. Many studies have analysed and catalogued the patterns of SVs by grouping them into different classes, such as duplication, deletion, inversion, and other complex variants. However, there are still no quantitative models of how these SVs are generated and how the different classes are related to each other.

To decipher the mechanisms of SVs, we develop a computational model to imitate the generation and evolution of cut-and-paste SVs resulting from DNA fragmentation and re-ligation after double strand break (DSB) formation. We simulate cell divisions with a stochastic birth-death branching process and introduce random DSBs which are repaired via non-homologous end joining, accounting for cell cycle stage. The simulations allow stepwise visualization of SV accumulation and quantification of SV patterns. We find that chromothripsis-like patterns are generated in one or two cell cycles when breaks collocate on chromosomes with all of them being repaired or a fraction remaining unrepaired. There are also extrachromosomal circular DNAs (ecDNA) and breakage-fusion-bridges (BFBs) generated at the same time, suggesting the relationship of ecDNA and BFBs with chromothripsis. We are currently using this model to infer important parameters from real events with Bayesian inference and using the data to refine our model as well.

In summary, our approach contributes to the development of a unified framework to investigate the evolutionary processes leading to various patterns of SVs.

Genomes

A database of copy number variant frequencies in the Spanish population

Daniel López-López (Fundación progreso y salud), Gema Roldan (Fundacion Progreso y Salud), Jose Luis Fernandez-Rueda (Fundacion Progreso y Salud), Rosario Carmona (Fundación progreso y salud), Virginia Aquino (Fundación progreso y salud), Maria Peña-Chillet (Fundación progreso y salud), Ruben Garcia (Fundación progreso y salud), Rocio Nuñez (Centro nacional de investigaciones oncologicas), Anna Gonzalez (Centro nacional de investigaciones oncologicas), Guillermo Pita (Centro nacional de investigaciones oncologicas) and Joaquin Dopazo (Fundacion Progreso y Salud).

Abstract:

Copy number variations (CNVs) play a significant role in the development of cancer and other diseases. While significant progress has been made on detection of CNVs, they remain less well studied than single nucleotide variations, in part due to challenges in their reliable identification from short-read sequencing data. Consequently, current predictors tend to report a high number of false positives, which are greatly influenced by the prediction methodology. As a result, the allele frequency population, which has been demonstrated to be critical for the identification of clinical relevant genomic alterations, may be artificially altered.

Here we present the first Iberian CNVs reference database specifically developed to address allele frequency population particularities as well as CNV predictor-specific biases. The database is based on a cohort of 448 Iberian unrelated samples derived from the widely used CSVS database (http://csvs.clinbioinfosspa.es/), and contains CNVs predicted with a set of highly cited tools. Additionally, CNVs are annotated with a variety of clinical relevant databases.

This database was implemented in the freely available web tool http://csvs.clinbioinfosspa.es/spacnacs, allowing the scientific/medical community to explore the database in a genomic browser as well as filtering CNVs according to their allele frequency, variant type, sample gender and subgroup, sequencing technology, overlapped genomic region features/annotations and the pipeline/tool used. Additionally, a beacon web service has been implemented to facilitate the integration with other resources and tools.

Genomes

A phylum-wide genomic screen in cyanobacteria for RNA structure motifs adjacent to orthologues genes.

Adrian Geissler (University of Copenhagen), Elena Carrasquer-Álvarez (University of Copenhagen), Niels-Ulrik Frigaard (University of Copenhagen), Jan Gorodkin (University of Copenhagen) and Stefan E. Seemann (University of Copenhagen).

Abstract:

Cyanobacteria are multitalented photosynthetic microorganisms with promising applications in modern green biotech industries due to their capability of capturing CO2 and producing a wide range of products, such as biofuels. In order to leverage these capabilities, the study of cyanobacterial genomes and gene regulatory pathways is crucial. One mechanism of gene regulation involves cis-regulatory RNA structures that regulate the expression of cis-encoded genes dependent on various factors, such as temperature, metabolite concentrations, or protein binding. We analyzed ~200 complete genome assemblies of cyanobacterial organisms and estimate that they have a high potential for novel cis-regulatory RNA structure motifs (RSMs). Therefore, we identified genomic sequences adjacent to single copy orthologues genes universal in the cyanobacteria lineage (BUSCO) as candidate regions to harbor the RSMs. Next these genomic sequences (respectively for each gene) were screened for conserved RNA structure using CMfinder. Our preliminary analysis results in 4,700 putative RSMs, and we are in the process of filtering them based on structural alignment scores and phylogeny. One-third of these putative RSMs are further supported by R-scape, which finds their expected fraction of basepairs with significant covariation (compensatory basepair changes).

Genomes

A universal protein-coding gene finder

Ferriol Calvet (Centre for Genomic Regulation (CRG)), Jaume Reig (Centre for Genomic Regulation (CRG)), Emilio Righi (Centre for Genomic Regulation (CRG)), Francisco Camara (Centre for Genomic Regulation (CRG)) and Roderic Guigó (Centre for Genomic Regulation (CRG)).

Abstract:

The Earth Biogenome Project will sequence the genome of 1.8M eukaryotes. Gene identification is essential to uncover the biology encoded in genome sequences. To capture the transcriptional complexity of genomes, gene annotation methods incorporate deep RNA sequencing and other data into complex pipelines. These require substantial resources, available only at a few sites, and are only partially successful as illustrated by the fact that the human gene set has not yet been finalised. Moreover, functional inference on the biology of genomes can only be made from coding genes. Given the strong imprint that coding regions leave in genome sequences, ab initio methods can actually produce decent predictions of coding genes without the need for transcriptome data.

We have redesigned the program geneid to extremely efficiently predict the dominant isoform of protein-coding genes. This still captures a significant fraction of a genome's biology, as shown by the fact that when considering all splice isoforms there is only marginal gain in the functional assignment over considering just one. When using a non-redundant set of proteins and DIAMOND for protein-DNA comparisons, we show that geneid is comparable to top ab initio methods, but orders of magnitude faster in vertebrates' and arthropods' genomes annotated in Ensembl. The matrices describing the coding regions and the splice sites are estimated from the protein-DNA matches, making geneid a train-free program. The minimal computational need makes geneid carbon footprint aware and addresses the equity concerns in genome diversity projects, contributing to empowering local communities to perform genome analysis.

Genomes

Accurate de novo identification of biosynthetic gene clusters

Laura Carroll (European Molecular Biology Laboratory), Martin Larralde (European Molecular Biology Laboratory), Jonas Fleck (ETH Zürich), Ruby Ponnudurai (European Molecular Biology Laboratory), Alessio Milanese (ETH Zürich), Elisa Cappio (European Molecular Biology Laboratory) and Georg Zeller (European Molecular Biology Laboratory).

Abstract:

Biosynthetic gene clusters (BGCs) are enticing targets for (meta)genomic mining efforts, as they may encode novel, specialized metabolites with potential uses in medicine and biotechnology. We developed GECCO (GEne Cluster prediction with COnditional random fields; https://gecco.embl.de), a high-precision, scalable method for identifying novel BGCs in (meta)genomic data, which is both more accurate and over 4x faster than the state-of-the-art. We applied GECCO to >300,000 genomes and metagenomes derived from human gut-associated microbes, representing the most extensive characterization of the human gut microbiome biosynthetic potential to date. In the process, we identified 616,045 BGCs, which encompass previously unexplored regions of the human gut microbiome biosynthetic landscape. Using a high-throughput clustering method, we then identified 107,856 Gene Cluster families, half of which can be found in micro-organisms highly prevalent in the human gut. The method developed here represents a scalable, interpretable machine learning approach, which can identify BGCs de novo with high precision and provide unprecedented insight into microbial biosynthetic potential.

Genomes

Aggregated genomic data as cohort-specific allelic frequencies can boost variants and genes prioritization in non-solved cases of inherited retinal dystrophies

Ionut Florin Iancu (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Irene Perea-Romero (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Gonzalo Nuñez-Moreno (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Lorena de la Fuente (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Raquel Romero (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Almudena Ávila-Fernandez (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Maria Jose Trujillo-Tiebas (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Rosa Riveiro-Álvarez (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Berta Almoguera (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Inmaculada Martín-Mérida (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Marta Del Pozo-Valero (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Alejandra Damián-Verde (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Marta Cortón (Instituto de Investigación Sanitaria Fundación Jimenez Díaz), Carmen Ayuso (Instituto de Investigación Sanitaria Fundación Jimenez Díaz) and Pablo Minguez (Instituto de Investigación Sanitaria Fundación Jimenez Díaz).

Abstract:

The introduction of next-generation sequencing in the diagnosis of genetic diseases has increased the known repertoire of causal variants and genes involved, as well as the amount of genomic information produced, that is not always shared or reused. We built an allelic-frequency database for a heterogeneous cohort of genetic diseases to explore the aggregated genomic information and boost diagnosis in inherited retinal dystrophies (IRD). We retrospectively selected 5683 index-cases with clinical exome sequencing tests available, 1766 with IRD and the rest, with diverse genetic diseases. We calculated subcohort's IRD specific allele-frequencies and compare it with suitable pseudocontrols. Focusing on non-solved IRD cases, we prioritized variants with a significant increment of frequencies, among them 8 found in IRD non-solved cases that may contribute to explain the phenotype, and 10 out of 11 of uncertain significance that were reclassified as likely-pathogenic according to ACMG guides. Besides, we developed a method to highlight genes with more frequent pathogenic variants in non-solved IRD cases than in pseudocontrols weighted by the increment of benign variants in the same comparison. Thus, we identified 18 genes for further studies that provided new insights in five cases. Our resource can also help to calculate the carrier frequency of deleterious variants in IRD genes, being the most prevalent ABCA4 (~7%) and USH2A (~3%). A cohort-specific genomic database and phenotype-specific allele frequencies compared to controls can assist with variants and genes prioritization and operate as an engine that provides new hypothesis in specific non-solved cases, hence augmenting disease diagnosis rate.

Genomes

An evaluation of pipelines for DNA variant detection can guide a reanalysis protocol to increase the diagnostic ratio of genetic diseases

Raquel Romero (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Lorena de la Fuente (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Marta Del Pozo-Valero (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Rosa Riveiro-Álvarez (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), María José Trujillo-Tiebas (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Inmaculada Martín-Mérida (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Almudena Ávila-Fernández (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Ionut-Florin Iancu (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Irene Perea-Romero (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Gonzalo Núñez-Moreno (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Alejandra Damián (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Cristina Rodilla (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Berta Almoguera (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Marta Cortón (Instituto de Investigación Sanitaria Fundación Jiménez Díaz), Carmen Ayuso (Instituto de Investigación Sanitaria Fundación Jiménez Díaz) and Pablo Mínguez (Instituto de Investigación Sanitaria Fundación Jiménez Díaz).

Abstract:

Clinical exome (CE) sequencing has become a first-tier diagnostic test for hereditary diseases; however, its diagnostic rate is around 30-50%. In this study, we aimed to increase the diagnostic yield of CE using a custom reanalysis algorithm. Sequencing data were available for three cohorts using two commercial protocols applied as part of the diagnostic process. Using these cohorts, we compared the performance of general and clinically relevant variant calling and the efficacy of an in-house bioinformatic protocol (FJD-pipeline) in detecting causal variants as compared to commercial protocols. On the whole, the FJD-pipeline detected 99.74% of the causal variants identified by the commercial protocol in previously solved cases. In the unsolved cases, FJD-pipeline detects more INDELs and non-exonic variants, and is able to increase the diagnostic yield in 2.5% and 3.2% in the re-analysis of 78 cancer and 62 cardiovascular cases. These results were considered to design a reanalysis, filtering and prioritization algorithm that was tested by reassessing 68 inconclusive cases of monoallelic autosomal recessive retinal dystrophies increasing the diagnosis by 4.4%. In conclusion, a guided NGS reanalysis of unsolved cases increases the diagnostic yield in genetic disorders, making it a useful diagnostic tool in medical genetics.

Genomes

An Interactive Workflow for Exploratory Detection of Genomic Insertions on a Population Scale

Kedi Cao (Humboldt Universität zu Berlin), Nourhan Elfaramawy (Humboldt Universität zu Berlin), Matthias Weidlich (Humboldt Universität zu Berlin) and Birte Kehr (Leibniz-Institut für Immunotherapie, Regensburg).

Abstract:

Designing Data Analysis Workflows (DAWs) for scientific discovery is often an exploratory process.

The challenge of creating a DAW for population-scale data analysis is to ensure its adaptability for various data qualities and maintain its resource allocating ability. A good design of an automated DAW for genomic research offers intermediate output analysis and simple user interactions to increase the efficiency and accuracy of data analysis significantly. Our project is based on the bioinformatics tool PopIns and its updated version PopIns2. The tools are designed to discover and characterize non-reference sequence insertions in human genomes on a population scale. PopIns2, as the successor of PopIns, enables the application of PopIns to a larger sample size and higher precision. With PopIns in hands, we aim to (i) improve this already competitive tool by reducing the workflow execution cost, and (ii) increase its flexibility by developing an automated DAW that integrates all functions of PopIns along with intermediate data analysis and simple user interactions. Specifically, we present the PopinSnake workflow powered by Snakemake as the core workflow engine. Within this DAW, we further modularized the program design while implementing data visualization for intermediate outputs and simple user interactions. For example, we incorporated an optional sub-workflow that can be executed if intermediate results indicate a need for contamination removal from input samples. We anticipate such an exploratory workflow to simplify genomic data processing for researchers.

Genomes

An Orthogonal Approach for Genome Assembly Curation: Examples on Vertebrate Genome Assemblies

Fernando Cruz (Centro Nacional de Análisis Genómico (CNAG-CRG)), Lucía Álvarez-González (Genome Integrity and Instability Group (IBB), Universitat Autònoma de Barcelona (UAB)), Rut Lara-Hernández (Centro Nacional de Análisis Genómico (CNAG-CRG)), Jèssica Gómez-Garrido (Centro Nacional de Análisis Genómico (CNAG-CRG)), Margarida Barcelo-Serra (Institut Mediterrani d'Estudis Avançats, IMEDEA (UIB-CSIC)), Laura Baldo (Institute for Research on Biodiversity (IRBio). University of Barcelona (UB)), Aurora Ruiz-Herrera (Genome Integrity and Instability Group (IBB), Universitat Autònoma de Barcelona (UAB)), José Antonio Godoy (Estación Biológica de Doñana, CSIC) and Tyler Alioto (Centro Nacional de Análisis Genómico (CNAG-CRG)).

Abstract:

Current sequencing technologies, including long-read sequencing and Hi-C, allow us to achieve chromosome-level genome assemblies at a reasonable cost. However, these assemblies are not completely error-free as they can present mis-joins, artificial translocations and inversions. As de novo genome assembly is a sequential process, errors can be inherited from early stages up to HiC-scaffolding. Therefore, final assemblies always require detailed revision and curation to improve their final quality. Our team has gained experience with several genomes on how to resolve these issues, assign sex chromosomes, fix miss-assemblies, remove contaminants and rescue unlocalized contigs belonging to the “major chromosomes” assembled (i.e. super-scaffolds). This process involves the revision and editing of the Hi-C contact maps with PreTextView or Juicer, using additional sources of information (i.e. long-reads, density of telomeric repeats, synteny and conservation of chromatin conformation) to aid in decision-making. In general, syntenic breaks and rearrangements coincide with issues observed in the contact maps. However, in some cases, inspection of long-read alignments is required to confirm or reject breaks introduced during Hi-C-scaffolding. Here we showcase several examples of the curation of vertebrate genome assemblies – including a reptile (Podarcis lilfordi), a fish (Xyrichtys novacula) and four lynx species - using orthogonal approaches.

Genomes

Analysis of new immune systems using Deep Learning

Sven Hauns (Albert-Ludwigs-Universität Freiburg, Bioinformatics group), Omer Alkhnbashi (Albert-Ludwigs-Universität Freiburg, Bioinformatics group) and Rolf Backofen (Albert-Ludwigs-Universität Freiburg, Bioinformatics group).

Abstract:

Phages and plasmids are the most widespread biological entities, leading to frequent attacks on bacteria and archaea. As a result, bacteria have evolved numerous and diverse methods to defend themselves against such an onslaught. However, many of these immune systems have yet to be discovered. We have developed a deep learning tool that searches a genome and automatically identifies cassettes of defense genes. In addition, the model suggests potential new classes that may still be unknown. To identify potentially related systems and quantify the uncertainty of the prediction, we need to accurately determine the inaccuracy of the prediction by applying calibration methods.

The algorithm we developed uses five steps to achieve this goal: First, the genome is divided into genes, and then a deep-learning model classifies all candidates. The model output is calibrated in the third step to produce a likelihood-based classification metric. This classification metric is then used to reject data, accept the prediction or predict a new class. When multiple predictions of the same immune class occur nearby, a simple rule-based the mechanism enables the classification of immune cassettes in the final step. Using this method, we were able to identify immune system cassettes in both archaea and bacteria, of which about 5% were classified as potential new immune system types.

Genomes

Analysis of SNPs in E. Coli populations in murine gut metagenomes

Francisco Cerqueira (Instituto Gulbenkian de Ciência), João Costa (Instituto Gulbenkian de Ciência), Massimo Amicone (Instituto Gulbenkian de Ciência) and Isabel Gordo (Instituto Gulbenkian Ciência).

Abstract:

Analysis of occurring E. coli populations from metagenomes for SNPs detection was tested with inStrain and MIDAS tools. A bash pipeline was developed to process the metagenomic data from the raw reads, assembly, binning and the SNPs detection with the inStrain and MIDAS. Several scripts in R and python were incorporated into the pipeline to process several intermediate files, as well as the final output tables from inStrain, and MIDAS. Two fecal samples from one mice ob/ob with mutated ob gene, that encodes for Leptin, were collected at two distinct time-points (day 35 and 273), and sequenced. The metagenomic analysis revealed the estimated diversity coverage was above 90% in both samples. The co-assembly had a L50=445 and N50=28,026 with a cumulative length of 60,170,750 bp (contigs >= 1,000 bp). A total of 12 metagenome assembled genomes (MAGs) with completeness greater 50% and redundancy lower than 10%. Those MAGs along with two E. coli were used for inStrain and MIDAS. The results were compared with previous results obtained by E. coli populations WGS and then analyzed with breseq. From the 12 SNPs detected with breseq, 10 were detected both by inStrain and MIDAS at day 273.

Genomes

Assessment of supervised methods for lncRNA function prediction from expression data

Fatemeh Behjati (Uniklinikum and Goethe University Frankfurt), Hicham Saddiki (Uniklinikum and Goethe University Frankfurt), Matthias S. Leisegang (Goethe-Universität Frankfurt am Main), Theresa Gimbel (Uniklinikum and Goethe University Frankfurt), Frederike Boos (Uniklinikum and Goethe University Frankfurt), Stefanie Dimmeler (Uniklinikum and Goethe University Frankfurt), Ilka Wittig (Uniklinikum and Goethe University Frankfurt), Reinier Boon (Uniklinikum and Goethe University Frankfurt), Ralf P. Brandes (Uniklinikum and Goethe University Frankfurt) and Marcel H. Schulz (Uniklinikum and Goethe University Frankfurt).

Abstract:

Protein function prediction is essential for understanding disease mechanisms and can help discovering suitable drug targets. There have been many studies dedicated to predict the function of protein coding genes (PCGs), however there is an evident paucity of methods identifying the function of long non-coding RNA (lncRNA) genes; key molecules involved in many biological processes. Several studies addressed this problem by incorporating the genetic composition of lncRNA genes, but their strategies often lack the cell-type specificity and other dynamics involving regulation of lncRNA genes.

In this work, we explore the function prediction from a different angle yet inspired by existing approaches for PCGs. We tested different supervised classification methods for predicting the function of lncRNA genes from expression data measured in thousands of samples. This approach is based on the guilt-by-association approach in which genes with similar expression patterns will probably share similar functions, allowing to transfer annotations of PCGs to lncRNA genes with matching patterns. Using the GOslim annotations, we train multi-label classifiers that learn hundreds of molecular or biological functions from PCGs and transfer them when predicting for lncRNA genes.

We have experimentally validated our findings by cross comparing the predicted functions of some lncRNAs with a set of enriched functions obtained from protein pull-down experiments. Our results suggest a promising agreement between the predictions of our best classifier and experimentally determined functions. Our work paves the way for automated function prediction for lncRNA genes, but also suggests limitations for the prediction of some categories using expression data alone.

Genomes

Assessment of the structural differences of the JHEH from closely related insects as potential targets for safe pesticide design

Weronika Bagrowska (Tunneling Group, Biotechnology Centre, Silesian University of Technology), Tomasz Skalski (Tunneling Group, Biotechnology Centre, Silesian University of Technology), Maria Bzówka (Tunneling Group, Biotechnology Centre, Silesian University of Technology) and Artur Góra (Tunneling Group, Biotechnology Centre, Silesian University of Technology).

Abstract:

Commonly used pesticides have several limitations which are of great concern in the long perspective: they have a broad impact on a large variety of insects, are toxic for plants and animals and can accumulate in the environment. The most visible examples of the negative impacts are their harmful effect on pollinating insects and decline in biodiversity. One of the elegant ways, which was proposed to control the pest population, was based on pesticides that mimic insect hormones, however, the observed selectivity was far from expectations.

Some studies have reported that inhibiting the juvenile hormone epoxide hydrolase (JHEH), which plays an important role in metamorphosis, may be effective in selective insect control. Such selectivity is inextricably linked to differences in hormone binding sites. In order to answer the question of whether JHEH could indeed be a good molecular target for the design of selective inhibitors, we performed an analysis of the JHEH sequences available in the UniProt and NCBI databases for 184 insects. We selected 10 representative structures and compared their binding cavity surrounding. Our promising results indicate the key similarities and differences that can be used to design selective JHEH pest inhibitors. Observed differences suggest that the design of the selective inhibitors capable of distinguishing between closely related insects can be possible.

The work was supported by the National Science Centre, Poland (UMO-2020/39/B/ST4/03220) and in part by PL-Grid Infrastructure.

Genomes

Automating the genome assembly process with Snakemake

Jèssica Gómez-Garrido (Centro Nacional de Análisis Genómico (CNAG-CRG)), Fernando Cruz (Centro Nacional de Análisis Genómico (CNAG-CRG)), Marc Palmada-Flores (Institut de Biologia Evolutiva, Universitat Pompeu Fabra (CSIC-UPF)), Laura Baldo (Institute for Research on Biodiversity (IRBio), University of Barcelona (UB)) and Tyler Alioto (Centro Nacional de Análisis Genómico (CNAG-CRG)).

Abstract:

The availability of high-quality reference genomes for non-model species is essential for evolutionary studies, conservation genomics, the study of ecosystem interactions, or for molecular breeding. The adoption of long-read sequencing technologies by the genome assembly field has made sequencing and assembly easier and cheaper, and at the same time it has increased the quality of the resulting assemblies. However, obtaining de novo genome assemblies is still a complex process that requires the combination of multiple sequencing technologies and involves several steps, some of which are time and memory hungry. To simplify this process, we have developed a Snakemake pipeline that takes a combination of long Oxford Nanopore reads and short Illumina paired-end reads to assemble the genome of any eukaryotic species. Moreover, the pipeline can be configured to produce multiple assemblies using different assembly or polishing tools, and it automatically computes standard evaluation metrics to aid in the selection of the optimal assembly. The use of this pipeline for assembling many different organisms has resulted in high quality assemblies, which, with the further use of Hi-C data or genetic maps, we have been able to further scaffold to chromosome-level. Here we show how the pipeline integrates all of the required steps for generating a high-quality genome assembly, using as an example the genome of the Lilford's wall lizard (Podarcis lilfordi).

Genomes

Beacon Network goes v2

Dmitry Repchevsky (Barcelona Supercomputing Center), Sergi Aguiló-Castillo (Barcelona Supercomputing Center), Dominik Bruchner (Barcelona Supercomputing Center), Teemu Kataja (CSC - IT Center for Science Ltd.), Ville Muilu (CSC - IT Center for Science Ltd.), Juha Törmroos (CSC - IT Center for Science Ltd.), Babita Singh (Centre for Genomic Regulation), Jordi Rambla (Centre for Genomic Regulation), J. Dylan Spalding (CSC - IT Center for Science Ltd.), Michael Baudis (Swiss Institute of Bioinformatics), Salvador Capella-Gutierrez (Barcelona Supercomputing Center) and Josep Lluís Gelpí (Barcelona Supercomputing Center).

Abstract:

With its first concepts going back to 2014, in 2018 the Global Alliance for Genomics and Health (GA4GH) approved Beacon v1 protocol as its standard for genomics data discovery. Warmly welcomed by the genomics community, many institutions implemented the protocol to provide a standardized way to query over their genomic data. The main advantage of these efforts emerged through the federated data discovery enabled by networks of beacons.

Recently approved by the GA4GH, the completely redesigned Beacon v2 is a big step towards the adoption of Beacon API in clinical genomics and healthcare. The protocol changes require a corresponding redesign of the Beacon Network architecture and infrastructure in order to support new protocol features and datatypes.

The Barcelona Supercomputing Center (BSC) present together with IT Center for Science Ltd. (CSC), Centre for Genomic Regulation (CRG) and Swiss Institute of Bioinformatics (SIB) the prototype implementation of the ELIXIR Beacon Network v2 service that enables querying individual beacons in the network, the aggregation of Beacon responses, supports ELIXIR AAI with open, registered and controlled access tiers and the handover to ELIXIR Core data resources.

The service is jointly operated by BSC in Spain and CSC in Finland. We expect that other ELIXIR communities such as Federated Human Data, Rare Diseases, human Copy Number Variation, Proteomics, Plants, Cancer Data, and Health Data will benefit from this service either directly or through the experience from the ELIXIR Beacon Network v2 design and implementation.

Genomes

Benchmarking and improving imputation approaches for recurrent inversions in the human genome

Illya Yakymenko (Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Barcelona), Jon Lerga-Jaso (Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Barcelona) and Mario Cáceres (ICREA, Barcelona; Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Barcelona).

Abstract:

Inversions are a type of structural variant that has been usually involved in phenotypic differences among individuals. Due to certain features, such as the lack of loss or gain of DNA or the presence of large inverted repeats at their breakpoints, the characterization of inversions is quite difficult. It has been recently found that many human inversions are recurrent and they are not in linkage disequilibrium (LD) with other genomic variants. For this reason, the effect of recurrent inversions has been largely missed in current genome wide association studies (GWAS), and it is necessary to develop new methods to predict inversion genotypes accurately in the datasets of interest. Here, we report a benchmarking analysis of the genotype predictions among different imputation tools, comparing IMPUTE2 with other programs, such as IMPUTE5, BEAGLE or scoreInvHap. Imputation accuracy was calculated as the LD (r2) between observed and imputed genotypes. From our set of 130 experimentally genotyped inversions, we found out that 23 and 18 out of 55 recurrent inversions could be imputed with high accuracy (r2 > 0.8) in European and African populations, respectively, but results vary among programs. Nevertheless, this ratio increases to 26/55 for both populations when we filter out samples with a post-imputation genotype probability lower than 0.8. Finally, we are also testing a tool based on deep learning, which could avoid some of the HMM-based algorithm limitations, such as the effect of the variant position, linkage dependence or region complexity, to increase our catalogue of imputable inversions.

Genomes

Cell type specific drug response prediction from single cell data

Nikoletta Katsaouni (Institute of Cardiovascular Regeneration) and Marcel Schulz (Institute of Cardiovascular Regeneration).

Abstract:

Pharmacogenomics is an interdisciplinary field of pharmacology and genomics with the ultimate target to develop personalised clinical treatment for patients to assist precision medicine. The use of machine learning approaches, without any pre-assumptions on the drugs under investigation, has the potential to give rise to novel therapies for precision medicine. Previous studies tried to elucidate biological mechanisms of response to one or multiple medications with matrix factorization (MF) techniques, e.g. [1]. Here we are suggesting a new multimodal model for the prediction of drug response for cancer patients based on their genotype. The model assesses the effectiveness of already established, under evaluation and candidate monotherapies and drug combinations. scRNA-seq data from cancer cell lines and drug response data for different cell lines are integrated in the same latent space in order to prioritise drugs. We are using a probabilistic coupling for the miscellaneous data modalities. Hence, the proposed model allows the in vitro cell-type specific effect approximation for a huge variety of drugs, reducing in this way drastically the laboratory trials and facilitates novel drug discoveries.

Genomes

CESIM – A Simulator for Clonal Evolution

Sarah Sandmann (Institute of Medical Informatics), Silja Richter (Institute of Medical Informatics), Xiaoyi Jiang (Institute of Computer Science) and Julian Varghese (Institute of Medical Informatics).

Abstract:

Real data sets on clonal evolution, containing a variety of time points, different sources of data and evolution profiles, are sparse. Furthermore, ground truth is often unknown. Instead, the most likely clonal development is inferred from partly noisy, partly incomplete data. Therefore, realistic simulated data represent a key element in developing reliable algorithms for clonal evolution.

We present CESIM – a Clonal Evolution SIMulator. Our novel simulation tool is able to generate realistic variant calls for an underlying user-definable clonal evolution. All main types of clonal evolution are considered: linear, branched dependent and branched independent (of note, punctuated and neutral evolution can be simulated as specific forms of linear and branched dependent evolution).

CESIM reports variants together with information on their variant allele frequency, cancer cell fraction, associated clusters, as well as the clusters' parents. A user may select the number of clones and time points to be simulated. Furthermore, the number of variants, mean coverage of the sequencing experiment and purity of the tumour sample may be defined.

The output reported by CESIM can be taken as direct input for all common clonal evolution algorithms: 1) approaches performing clustering of variants, 2) reconstructing clonal evolution trees and 3) visualizing clonal evolution. Thereby, CESIM allows for systematic evaluation of available approaches as well as development of new optimized pipelines.

CESIM is written in R and provides an intuitive graphical user interface. The approach is publically available at https://imigitlab.uni-muenster.de/sandmans/cesim.

Genomes

Characterising human cell types using TF footprinting

Aybuge Altay (Max Planck Institute for Molecular Genetics), Yufei Zhang (Max Planck Institute for Molecular Genetics) and Martin Vingron (Max Planck Institute for Molecular Genetics).

Abstract:

Interactions between transcription factors (TFs) and their target DNA regulate the expression of downstream genes. Current high-throughput techniques such as RNA-seq and ATAC-seq let us peek into the mechanisms of these interactions. TF footprinting leverages the chromatin accessibility measurements and provides a new angle to understand these interactions. TF footprinting for ATAC-seq relies on the fact that one observes fewer transposase insertions where a TF protects the DNA. Consequently, read coverage is lower there and detection of these regions in turn indicates TF binding.

Here we report on our efforts to employ ATAC-seq based TF footprinting for the purpose of annotating clusters of single-cell ATAC-seq data with a putative cell type. We start with bulk ATAC-seq data and determine which TFs match the TF footprints. Using bulk ATAC-seq data of known cell-types we learn their active TFs and then transfer this information to sc clusters. We applied this approach for cell-type annotation in PBMC data and further extend it to human brain data, indicating the utility of TF footprinting in discriminating cell types.

Genomes

Characterization of the role of regulatory variants for trait pleiotropy

Aitor González (Aix-Marseille Université).

Abstract:

Genome-wide association studies (GWAS) have shown that pleiotropic genetic variants affecting multiple traits are relatively common in the genome. A large majority of these variants fall in non-coding regions and are likely gene regulatory variants.

In this poster, we have computed the colocalization of 400 GWAS studies and 125 eQTL studies to pinpoint 9099 regulatory variants associated with GWAS traits. Regulatory variants belonging to more than one GWAS trait category have been defined as pleiotropic. We find that pleiotropic variants are clustered around a few pleiotropic genomic regions. Pleiotropic variants are also associated with more gene expression variation in more tissues and bind more transcription factors.

In summary, these analyses clarify the role of regulatory variants for trait pleiotropy.

Genomes

CimpleG: Simple CpG methylation signatures

Tiago Maié (Institute for Computational Genomics, RWTH Aachen University Medical School), Marco Schmidt (Helmholtz-Institute for Biomedical Engineering, Stem Cell Biology and Cellular Engineering, RWTH Aachen University), Wolfgang Wagner (Helmholtz-Institute for Biomedical Engineering, Stem Cell Biology and Cellular Engineering, RWTH Aachen University) and Ivan G. Costa (RWTH Aachen University).

Abstract:

In a clinical setting, small molecular signatures capable of identifying disease or a certain phenotype are of great practical use for cost reduction. We are interested in the use of DNA methylation as a marker due to its high specificity for specific cell types. We, therefore, explore feature selection methods to obtain singular DNAm signatures (1 site per target class) to perform cell type classification and computational deconvolution.

In previous work, we have shown that these singular signatures can be used to estimate the composition of tissues and cellular mixes with an accuracy comparable to that of models using orders of magnitude more features (Schmidt, et al. 2020).

In this work, we define a feature selection heuristic, apply it within a cross-validation setup and devise a framework in the form of an R package (CimpleG), to systematically train, validate, test and select singular DNAm signatures from DNA methylation array data.

We benchmarked CimpleG on two different datasets that we have compiled from DNA methylation array data. One comprised of different somatic cells (576 samples) and the other with leukocyte subsets (365 samples). In total, we train for 16 different cell types and achieve great classification results (0.90~0.99 Area Under the Precision-Recall curve on the test data).

Finally, we use the CimpleG signatures on different cellular deconvolution problems to show the potential applications of this method.

References:

Schmidt, M. et al. "Deconvolution of cellular subsets in human tissue based on targeted DNA methylation analysis at individual CpG sites." BMC biology (2020).

Genomes

CloneTracer enables the identification of healthy and leukemic cells from single-cell transcriptomic data

Sergi Beneyto-Calabuig (Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology), Anne Kathrin Merbach (Department of Medicine, Hematology, Oncology and Rheumatology, University Hospital Heidelberg), Jonas Alexander Kniffka (Department of Medicine, Hematology, Oncology and Rheumatology, University Hospital Heidelberg), Chelsea Szu-Tu (Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology), Magdalena Antes (Department of Medicine, Hematology, Oncology and Rheumatology, University Hospital Heidelberg), Michael Scherer (Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology), Simon Raffel (Department of Medicine, Hematology, Oncology and Rheumatology, University Hospital Heidelberg), Carsten Müller-Tidow (Department of Medicine, Hematology, Oncology and Rheumatology, University Hospital Heidelberg) and Lars Velten (Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology).

Abstract:

Actute myeloid leukemia (AML) is a type of blood cancer characterized by the sequential accumulation of genomic alterations which gives rise to malignant populations referred to as clones. The identification and characterization of these clones is crucial for improving the understanding and treatment of AML. Single-cell DNA sequencing methods enable the assignment of individual cells to healthy and cancer clones, but are incapable of molecular profiling the clonal populations. Single-cell RNA sequencing (scRNAseq) on the other hand, allows the characterization of different cell populations based on their transcriptome, but does not inform on clonal identities. Here we developed CloneTracer, a generative model to identify healthy and malignant cells from scRNAseq. Our approach integrates information on nuclear single nucleotide variants (SNVs), mitochondrial SNVs (mtSNVs) and copy number variations (CNVs) obtained from the scRNAseq data as well as bulk information (e.g., exome sequencing). Accounting for the noise properties of scRNAseq measurements, CloneTracer infers the clonal hierarchy with the highest likelihood and probabilistically assigns single cells to clones. Our approach leverages stochastic variational inference and was written in pyro, a probabilistic programming language in python which uses pytorch as backend. Applied to a cohort of AML patients, CloneTracer unambiguously identified healthy and leukemic cells and, when present, different malignant subclones. Leveraging the clonal information, we performed intra-patient comparisons between clones and discovered known and novel expression markers enriched in healthy and malignant cells. Together, these results show the potential of CloneTracer to identify and characterize healthy and cancerous populations from scRNAseq data.

Genomes

Cluster-based conserved non-coding element (CNE) characterization

Silvia Fibi-Smetana (Graz University of Technology) and Leila Taher (Graz University of Technology).

Abstract:

The investigation of clusters of conserved non-coding elements (CNEs) is expected to provide insights into the mechanisms of gene regulation and our understanding of human disease.

With the aim of characterizing the evolutionary constrains acting on the distances that separate the CNEs found in clusters, we downloaded PhastCons elements based on 100-way alignments from the UCSC Genome Browser. These elements were then filtered and merged to obtain 1.2 million CNEs that were at least 24 bp long and did not overlap any protein-coding exons. Among these CNEs, 54% were intronic, 44% intergenic, and 2% overlapped with untranslated regions (UTR). Furthermore, we defined ~30,000 “clusters” containing CNEs with no conserved protein-coding elements in between them and bordered by the nearest conserved protein-coding elements upstream and downstream of each CNE in the cluster. On average, the clusters comprised 40 CNEs, and only ~8,000 CNEs were not part of any cluster, confirming previous reports indicating that CNEs are often found in clusters. Next, we used squared-change maximum parsimony to infer inter-CNE distances in the primate and mammalian ancestor. For comparison we also estimated the distance between the two protein-coding elements delimiting each cluster. Generally, we found that inter-CNE distances have contracted with respect to their loci. We hypothesize that epistatic interactions drive this pattern, and that clusters of CNEs have relevant functional or structural roles.

Genomes

CoLoRd: Compressing long reads

Marek Kokot (Silesian University of Technology), Adam Gudyś (Silesian University of Technology), Heng Li (Dana-Farber Cancer Institute; Harvard Medical School) and Sebastian Deorowicz (Silesian University of Technology).

Abstract:

The costs of maintaining exabytes of data produced by sequencing experiments every year has become a major issue in today's genomics. In spite of the increasing popularity of the third generation sequencing, the existing algorithms for compressing long reads exhibit minor advantage over general purpose gzip.

Here, we present CoLoRd, a compression algorithm for Oxford Nanopore and PacBio sequencing data. Its main contributions are (i) a novel method for compressing the DNA component of FASTQ files and (ii) a lossy processing of the quality stream. Since CoLoRd does not need a reference genome, it is suitable for different data types, including genomic, metagenomic, and transcriptomic. For the highest flexibility, the algorithm is also equipped with a reference-based mode.

CoLoRd revealed its potential on the latest ONT Bonito-base-called and PacBio HiFi data sets. By efficiently finding overlaps in the high fidelity reads, our algorithm reduced the size of the DNA stream by two orders of magnitude. This, accompanied by the quantization of quality levels, allowed squeezing FASTQ files to the 1/25 of their original size, which translated to 4-fold (ONT Bonito) and 10-fold (HiFi) advantage over lossless gzip compression. Importantly, the lossy compression of the quality stream did not affect the accuracy of downstream analyzes like variant calling or consensus generation.

We believe, that the presented research opens new opportunities in the field of long read sequencing, where maintaining gigantic data volumes has become one of the major contributors to the overall costs.

Genomes

Combining Epigenome and Genetic Mutation Data to Study Disease-Relevant Cell Types and Gene Sets

Ahlam Mallak (Goethe University Frankfurt), Dennis Hecker (Goethe University Frankfurt) and Marcel Schulz (Goethe University Frankfurt).

Abstract:

Genome-Wide Association Studies (GWAS) revealed the fact that most Single Nucleotide Polymorphisms (SNPs) appear in non-coding genomic regions. However, the relevant cell types and the mechanisms in which these mutations are linked to a certain phenotype remain complicated to unravel using GWAS alone.

Various studies have found that pre-defined functional categories such as cell-type specific enhancers are enriched in mutations related to disease etiology. However, untangling the intricacy behind the contribution of a multitude of cell types requires sophisticated methods that are able to address both overlap and correlation.

We have created a pipeline using Stratified LD-Score Regression (S-LDSC), an established method to partition genetic heritability, to compute the heritability enrichment of genomic functional categories or gene sets. This pipeline uses a large set of cardiovascular epigenome data to determine cell-type specific enhancer-gene interactions. We show how disease GWAS summary statistics, epigenome measurements and single cell expression data from patients can be combined to estimate cell-type specific contributions to different cardiovascular diseases.

In this work, we illustrate our approach with an in-depth analysis of GWAS summary statistics of various cardiovascular conditions along with 66 functional categories collected from 45 different cell-types. Our results reveal novel and already established links between cardiovascular cell-types and diseases such as coronary artery disease or ischemic stroke.

Genomes

Comparative Analysis of Long Versus Short-Read Sequencing Technologies in Terms of Structural Variants Inference for Trios

Sachin Gadakh (Center of New Technologies, University of Warsaw), Mateusz Chiliński (Warsaw University of Technology), Karolina Jodkowska (Center of New Technologies, University of Warsaw), Jan Gawor (DNA Sequencing and Oligonucleotide Synthesis Laboratory IBB PAS) and Dariusz Plewczynski (Center of New Technologies, University of Warsaw).

Abstract:

Structural Variants (SVs) are alterations in the human genome that may be linked to the development of human diseases. A wide range of technologies are currently available to detect and analyze SVs, but the restrictions of each of the methods are resulting in lower total accuracy. Therefore we aimed to develop a reliable computational pipeline to merge and compare the SVs from various tools to get accurate SVs for downstream analysis. In this study, we performed the whole genome sequencing (WGS) of 9 samples of TRIOS families from a 1000 genome project, using long-read sequencing technology such as Oxford Nanopore Technology (ONT). Further, We performed a detailed analysis of WGS data generated using ONT and Illumina, of these samples. The SVs were identified using 6 SV callers specific to long-read and 15 SV callers specific to short-read data. We created a machine learning tool for getting the consensus of SVs based on SVs obtained from given SV callers. Then, we used this tool on 9 samples to merge the results of these SV callers, trains them using neural networks, and benchmarked them on the gold set of Structural Variants to identify the reliable list of SVs for each member of the TRIOS families. We conclude that the consensus SVs detected by our algorithm over sequencing data from both sequencing technologies is not only highly accurate but also SVs from ONT are superior compared to Illumina in terms of distribution of the range of numbers and sizes.

Genomes

Comparison of Chromatin Contacts Detectability in Human ESC-H1 Mapped Using GAM or Hi-C

Teresa Szczepińska (Centre for Advanced Materials and Technologies, Warsaw University of Technology), Christoph Thieme (Berlin Institute for Medical Systems Biology, Max-Delbrück Centre for Molecular Medicine), Sachin Gadakh (Center of New Technologies, University of Warsaw), Alexander Kukalev (Berlin Institute for Medical Systems Biology, Max-Delbrück Centre for Molecular Medicine), Warren Winick-Ng (Berlin Institute for Medical Systems Biology, Max-Delbrück Centre for Molecular Medicine), Rieke Kempfer (Berlin Institute for Medical Systems Biology, Max-Delbrück Centre for Molecular Medicine), Thomas Sparks (Berlin Institute for Medical Systems Biology, Max-Delbrück Centre for Molecular Medicine), Miao Yu (Ludwig Institute for Cancer Research, La Jolla CA 92093, USA), Bing Ren (Ludwig Institute for Cancer Research, La Jolla CA 92093, USA), Dariusz Plewczynski (Faculty of Mathematics and Information Science, Warsaw University of Technology,) and Ana Pombo (Berlin Institute for Medical Systems Biology, Max-Delbrück Centre for Molecular Medicine).

Abstract:

Understanding the limitations of chromatin mapping techniques in detecting biological aspects of genome 3D structure is important for discovering its function, efficiently assessing long-range gene regulation, and making comparisons between cell types. Whereas Hi-C contact maps are based on the frequency of proximity ligation in millions of cells, Genome Architecture Mapping (GAM) extracts spatial information about 3D genome topology by sequencing the genomic content of hundreds to thousands ultra-thin, randomly oriented nuclear slices. We have performed GAM on H1 human embryonic stem cells and compared GAM data with Hi-C data from the same cell line available on 4D Nucleome repository. We have measured the detectability of each method along the linear genome at 50 kb genomic resolution. We devised methods to robustly remove regions with detectability outliers in either method. Moreover, we have characterized genomic windows according to the presence of functional features such as histone modifications, protein binding, gene density and expression (GRO-seq, RNA-seq), chromatin accessibility (ATAC-seq), enhancers and lamina association. Our analyses indicate that the two methods have good similarity at TAD level but show differences in the assigned compartments. We also observed different capabilities in detecting functionally active genomic regions. While more windows need to be cleared out from Hi-C than from GAM because of low detectability, we noticed that regions annotated with features are more often removed from one dataset exclusively.

Genomes

Comprehensive analysis of genomic patterns in different types of dengue virus

Myeongji Cho (Honam National Institute of Biological Resources), Xianglan Min (Seoul National University) and Hyeon S. Son (Seoul National University).

Abstract:

In this study, quantitative estimates of multiple indices of codon usage in 11 functional protein coding regions of dengue virus (DENV) types 1–4 were analyzed. Complete coding sequences of DENV1–4 all exhibited AT and AT3 bias, and showed A>T% and G>C%, overall and at the third codon position. Except for capsid and NS4a genes, AT3 values were generally higher than GC3 values. Uniquely, the capsid uniquely showed a strong GC3 bias and preference for G-end codons. The mean effective number of codon (ENC) values for each DENV gene suggested greater codon diversity in capsid and M genes of DENV-4. Apart from the M sequence of DENV-4, all Nc coordinates were plotted under the expected ENC curve, suggesting biased codon usage patterns. The Nc plot showed that the selection pressure on the M and capsid regions of DENV-2 was much greater than on the same genes for other DENV types. Meanwhile, the NS2a region of DENV-1 was subject to less selection pressure compared to other DENV types. In terms of codon usage similarity, DENV was more similar to Homo sapiens than to Aedes eagypti. The mean codon adaptation index (CAI) was highest (0.76) for capsid and lowest (0.71) for M. However, the CAI for the M region in DENV-2 was 0.76 (reference set: H. sapiens). Taken together with the Nc plot, the higher selection pressure for the M gene of DENV-2 could be attributed to the increased expression level of M protein.

Genomes

Containerised Pipelines for Sensitive Sequencing Data

Tina Visnovska (EpiGen, Medical Division, Akershus University Hospital, Lørenskog, Norway), Sadia Saeed (Department of Clinical Molecular Biology, EpiGen, Institute of Clinical Medicine, University of Oslo, Oslo, Norway), Lars la Cour Poulsen (EpiGen, Medical Division, Akershus University Hospital, Lørenskog, Norway), Torunn Rønningen (EpiGen, Medical Division, Akershus University Hospital, Lørenskog, Norway), Mai Britt Dahl (Department of Clinical Molecular Biology, EpiGen, Institute of Clinical Medicine, University of Oslo, Oslo, Norway), Junbai Wang (Department of Clinical Molecular Biology, EpiGen, Institute of Clinical Medicine, University of Oslo, Oslo, Norway), Tom Mala (Department of Endocrinology, Morbid Obesity and Preventive Medicine, Oslo University Hospital, Oslo, Norway), Jon A. Kristinsson (Department of Endocrinology, Morbid Obesity and Preventive Medicine, Oslo University Hospital, Oslo, Norway), Jens Kristoffer Hertel (Section for Morbid Obesity, Vestfold Hospital Trust, Tønsberg, Norway), Jøran Hjelmesæth (Section for Morbid Obesity, Vestfold Hospital Trust, Tønsberg, Norway), Matthias Blüher (Department of Medicine, University of Leipzig; HI-MAG, University of Leipzig & Leipzig University Hospital; Germany), Tone Gretland Valderhaug (Department of Endocrinology, Akershus University Hospital, Lørenskog, Norway) and Yvonne Böttcher (EpiGen, Medical Division, Ahus, Lørenskog; EpiGen, Institute of Clinical Medicine, UiO, Oslo; Norway).

Abstract:

When analysing human genomic and transcriptomic sequencing data in a hospital-based research environment, many restrictions regarding the process are in place to ensure that the patient's personal information is kept private. For our group this means that the analyses are executed on a high performance computing cluster dedicated to work with sensitive data. Developers have no access to internet from within and they have only limited possibilities to install standard bioinformatics tools. Taking also into account that similar datasets are generated in other research groups, we see added value in reusability of the pipelines when they can be easily deployed in different projects. These two factors led us to explore possibilities to develop and deploy containerised pipelines (with minimal requirements on installing additional software) for analysing next generation sequencing data.

Here we present a suite of publicly available pipelines to analyse RNAseq and ATACseq (Assay for Transposase-Accessible Chromatin using sequencing) data. We have used the code to analyse paired data from two different depots of human adipose tissue. The functionality is separated into three independent pipelines using singularity for containerisation and snakemake to orchestrate the pipelines' respective workflows. The first pipeline processes raw RNAseq reads to read counts for expressed genes and the second one processes raw ATACseq reads to read counts for accessible chromatin regions. The third pipeline takes the read counts for both feature types, performs differential analyses and combines them to identify co-occurrence of upregulated genes and open chromatin regions in the near proximity of the genes.

Genomes

Contaminant Detection using Differentiating k-mers

Alessia Petescia (Faculty of Mathematics, Physics and Informatics,Comenius University in Bratislava), Martina Nebohacova (Faculty of Natural Sciences, Comenius University in Bratislava), Jozef Nosek (Faculty of Natural Sciences, Comenius University in Bratislava), Brona Brejova (Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava) and Tomas Vinar (Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava).

Abstract:

When targeting a single genome for sequencing, there can still be potential contamination with non-target organisms. Their proper identification and removal is fundamental in order to avoid biases in further analysis. Typically, tools employed in metagenomics can be also used for contamination detection. This includes k-mer based methods (such as Kraken II) which offer a favourable trade-off between speed and sensitivity.

In our work, we examined differentiating k-mers in wild grapevine samples which were sequenced using the NovaSeq6000 platform. A k-mer is considered differentiating if either its frequency significantly differs between the samples or if it exclusively occurs in one of them. By systematically cataloguing differentiating k-mers in our samples, we discovered that many of them correspond to contaminants, such as grape leaves associated microbes. We propose that filtering sequencing reads spanning differentiating k-mers presents several advantages in detecting contaminants. A significantly decreased amount of reads in the analysis enables use of slower but more sensitive tools, such as BLAST, for precise identification of read origin. Moreover, it makes the sequence assembly a feasible option, potentially leading to even more precise classification. Last but not least, by leveraging differentiating k-mers between related samples, it is possible to perform the task of removing contaminants even without a reliable reference.

Genomes

Copy number-aware methylation deconvolution analysis of cancers from whole genome bisulfite sequencing data

Nana Mensah (The Francis Crick Institute), Elizabeth Larose Cadieux (The Francis Crick Institute), Tom Lesluyes (The Francis Crick Institute), Jonas Demeulemeester (The Francis Crick Institute) and Peter Van Loo (The Francis Crick Institute).

Abstract:

Aberrant DNA methylation is common across cancers and understanding the dynamics of these events is a grand challenge in cancer epigenetics. Whole genome bisulfite sequencing (WGBS) captures the greatest breadth of DNA methylation alterations at single-base resolution, however, bulk tumor methylomes are confounded by admixed normal cells and variations in DNA content owing to aneuploidy. We developed copy number-aware methylation deconvolution analysis of cancers (CAMDAC) to account for the effect of these confounders from bulk WGBS data. CAMDAC infers tumor purity and allele-specific copy number events from WGBS data and uses these quantities to deconvolve tumor methylation rates. We apply CAMDAC to 10 tumor samples with matched normal WGBS data from the Stand Up To Cancer (metastatic prostate) study and validate the method using methylation on reads phased to SNVs. Furthermore, we investigate the utility of aggregate panels as a reference for tumor methylation deconvolution in the absence of matched normal methylomes. We show that purified methylation rates offer new avenues for investigating DNA methylation dynamics and the interplay between copy number alterations and methylation in cancer evolution. Enriching for true tumor-normal differential methylation events, CAMDAC offers a lens through which we may better understand the epigenetic component of cancer evolution.

Genomes

Deciphering the etiology and role in oncogenic transformation of the CpG island methylator phenotype: a pan-cancer analysis

Josephine Yates (ETH Zürich) and Valentina Boeva (Institut Cochin/INSERM/CNRS).

Abstract:

Numerous cancer types have shown to present hypermethylation of CpG islands, also known as a CpG island methylator phenotype (CIMP), often associated with survival variation. Despite extensive research on CIMP, the etiology of this variability remains elusive, possibly due to lack of consistency in defining CIMP. In this work, we utilize a pan-cancer approach to further explore CIMP, focusing on 26 cancer types profiled in the Cancer Genome Atlas (TCGA). We defined CIMP systematically and agnostically, discarding any effects associated with age, gender or tumor purity. We then clustered samples based on their most variable DNA methylation values and analyzed resulting patient groups. Our results confirmed the existence of CIMP in 19 cancers, including gliomas and colorectal cancer. We further showed that CIMP was associated with survival differences in eight cancer types and, in five, represented a prognostic biomarker independent of clinical factors. By analyzing genetic and transcriptomic data, we further uncovered potential drivers of CIMP and classified them in four categories: mutations in genes directly involved in DNA demethylation; mutations in histone methyltransferases; mutations in genes not involved in methylation turnover, such as KRAS and BRAF; and microsatellite instability. Among the 19 CIMP-positive cancers, very few shared potential driver events, and those drivers were only IDH1 and SETD2 mutations. Finally, we found that CIMP was strongly correlated with tumor microenvironment characteristics, such as lymphocyte infiltration. Overall, our results indicate that CIMP does not exhibit a pan-cancer manifestation; rather, general dysregulation of CpG DNA methylation is caused by heterogeneous mechanisms.

Genomes

Detection and evaluation of variable number tandem repeats

Maryam Ghareghani (Max Planck institute for molecular genetics), Hossein Moeinzadeh (Max Planck institute for molecular genetics) and Martin Vingron (Max Planck institute for molecular genetics).

Abstract:

Variable number tandem repeats (VNTRs), among other types of genomic variants, e.g. SNVs, indels, are variants consisting of consecutively repeated units with polymorphic copy numbers and mutations in repeat units. VNTRs have been used in DNA fingerprinting for their hyper variability in the population. They have been also associated with genetic disorders and complex traits. The functional impact of VNTRs has been still largely undiscovered due to the repetitive nature of VNTRs, high variability in copy numbers, and their abundance in a genome.

Alignment ambiguity in VNTR regions makes their genotyping complicated, especially in regions with heterozygous copy numbers. We developed a method for discovery and evaluation of VNTRs using Illumina short reads and PacBio long reads. Our method performs several steps including extracting PacBio reads from the annotated VNTR regions, clustering reads by haplotype per region, haplotype assembly of phased reads, discovery of tandem repeat units and and genotyping their copy numbers (VNTR-typing).

We analyzed a cohort of healthy samples and patient samples with limb malformation. For this study, we focused on the VNTR regions overlapping with the genomic intervals associated with limb development including genes and other functional elements. We applied our method on these cohorts and were able to perform haplotype assembly of 80% of these VNTR regions and genotype 52% of them. This study allows us to resolve VNTR-typing and investigate the impact of VNTRs in detection and interpretations of disease causing variants.

Genomes

Development of a new structural variant detection software based on graph clustering algorithms from long reads

Nicolas Gustavo Gaitan Gomez (Universidad de los Andes) and Jorge Duitama (Universidad de los Andes).

Abstract:

Structural variants (SV) are polymorphisms defined by their length (>50 bp). The usual types of SVs are deletions, insertions, translocations, inversions, and copy number variants. SV detection and genotyping is fundamental given the role of SVs in phenomena such as phenotypic variation and evolutionary events. Thus, methods to identify SVs using long-read sequencing data have been recently developed. We present an accurate and efficient software to predict SVs from long-read sequencing data. This tool is implemented as a new functionality of the Next generation Sequencing Experience Platform (NGSEP) which facilitates the integration with other functionalities for genomics analysis. The algorithm starts collecting evidence (Signatures) of SVs from read alignments. Then, signatures are clustered based on a Euclidean graph with coordinates calculated from lengths and genomic positions. Clustering is performed by the DBSCAN algorithm, which provides the advantage of delimiting clusters with a high resolution. Clusters are transformed into SVs and a Bayesian model allows to precisely genotype SVs based on their supporting evidence. For benchmarking, our algorithm is compared against different tools using VISOR for simulation and the GIAB SV dataset for real data. For indel calls in a 20x depth Nanopore simulated dataset, the DBSCAN algorithm performed better, achieving an F-score of 98%, compared to 97.8 for Dysgu, 97.8 for SVIM, 97.7 for CuteSV, 96.8 for Sniffles. Additionally, the DBSCAN algorithm presented a better differentiation of close events.

Genomes

Differential methylation analysis signature evaluation using Hobotnica metric

Anna Budkina (MIPT), Alexey Stupnikov (MIPT, Research Center of Biotechnology RAS) and Yulia Medvedeva (MIPT, Research Center of Biotechnology RAS, VIGG RAS).

Abstract:

A signature as a list of differentially methylated CpG sites is a common result of various differential methylation analysis models. Due to the absence of a gold standard differential methylation analysis method, assessing the quality of a signature is essential. The ability of a signature to reflect general differences between groups and to be relevant both to original data and to data from other sources is a significant quality criterion. The objectives of the study were to validate the applicability of the earlier published Hobotnica metric to methylation signature assessment and to demonstrate its use for differential methylation analysis model evaluation.

The Hobotnica metric was validated using an existing signature and datasets from other studies. Hobotnica was then applied for the evaluation of six models for differential methylation analysis using four WGBS and two RRBS datasets.

It was shown that the Hobotnica metric value reflects the ability of the signature to separate comparison groups on a given dataset. The evaluation highlighted the models that outperformed others on all datasets. This study demonstrates that the quality of a differential methylation signature can be properly assessed using the Hobotnica metric.

Genomes

Discovering Significant Evolutionary Trajectories in Cancer Phylogenies

Leonardo Pellegrina (University of Padova) and Fabio Vandin (University of Padova).

Abstract:

Tumors are the result of a somatic evolutionary process leading to substantial intra-tumor heterogeneity. Single-cell and multi-region sequencing enable the detailed characterization of the clonal architecture of tumors, and have highlighted its extensive diversity across tumors. These insights on the mechanisms of tumor evolution have shown that, while there are inherent stochastic forces driving cancer, there are some features that are shared by the progression of certain tumors, such as some constraints in the order with which alterations arise. While several computational methods have been developed to characterize the clonal composition and the evolutionary history of tumors, the identification of significantly conserved evolutionary trajectories across tumors is still a major challenge.

We present a new algorithm, MASTRO, to discover significantly conserved evolutionary trajectories in cancer. MASTRO discovers all conserved trajectories in a collection of phylogenetic trees describing the evolution of a cohort of tumors, allowing the discovery of conserved complex relations between alterations. MASTRO assesses the significance of the trajectories using a conditional statistical test that captures the coherence in the order in which alterations are observed in different tumors.

On simulated data, MASTRO rigorously controls false discoveries, and it is effective in retrieving known ground truths. In contrast, previous works may be confounded by the co-occurence of alterations rather than their order. We apply MASTRO to data from non-small-cell lung cancer bulk sequencing and to acute myeloid leukemia data from single-cell panel sequencing, and find significant evolutionary trajectories recapitulating and extending the results reported in the original studies.

Genomes

DNA-DDA predicts genome wide 3D chromatin structure directly from the reference sequence

Xenia Lainscsek (Graz University of Technology - Institute for Biomedical Informatics) and Leila Taher (Graz University of Technology - Institute for Biomedical Informatics).

Abstract:

The hierarchical 3D conformation of the genome plays an essential role in gene regulation and proper cell function. With development of the high chromosome conformation capturing technique (Hi-C), researches have revealed the structure of chromatin in various cell types. The Hi-C assay results in a contact map, i.e. a matrix representative of the 3D contacts along the linear chain of DNA, from which various scale-dependent characteristic structures such as chromosomal compartments (100kbp to Mbp) can be deduced. Hi-C has advanced our understanding of genome architecture tremendously. However due to its high cost and time expenditure, increasing effort is being made to derive predictive computational models. Motivated by the field of nonlinear dynamics, in particular chaos theory, we present DNA-DDA, an approach adapted from a time analysis technique, delay differential analysis (DDA), that predicts genome wide contact maps and associated A/B compartments solely from the reference sequence. We hypothesize that sequences close in 3D space will share certain nonlinear dynamical properties to which we can gain access to by mapping the sequences into whats known as an embedding space. Specifically, we used a mere 20Mbp long region on chromosome 22 to generate sparse models for 4 different cell types and achieved exceptional classification performance across all remaining human autosomes. Our approach has the potential to become a powerful tool for studying the biological mechanisms underlying genome folding as well as for modeling the impacts of genetic variation on 3D structure which has been associated with a wide range of diseases.

Genomes

Evidence-driven annotation of the Trichechus manatus latirostris genome using long-reads

Alejandro Paniagua (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia, Spain.), Francisco J. Pardo-Palacios (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia, Spain.) and Ana Conesa (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia, Spain.).

Abstract:

While the production of a draft genome has become more accessible due to third-generation sequencing (TGS), the structural annotation of these new genomes has not been developed at the same pace. TGS of mRNA (lrRNA-seq) could play a key role in the improvement of gene annotation.

Previous studies have not assessed the effect of different sequencing technologies or long-reads processing pipelines in combination with genome annotation pipelines. In this study, we evaluated the accuracy of the lrRNA-seq supported gene prediction as a function of the sequencing platform PacBio or Nanopore, pipelines applied to the sequencing data, and the approach used to incorporate this data into the popular gen prediction tool AUGUSTUS.

For benchmarking purposes, we used a human dataset. Nanopore and PacBio sequencing data from WTC11 human cell line were processed independently and in combination using IsoSeq3 or FLAIR pipelines. The performance of ab initio gene prediction on the human genome did not improve when using long reads-derived genes to train AUGUSTUS instead of BUSCO genes. However, the gene prediction accuracy greatly improve when lrRNA-seq data was incorporated during the gene prediction step, especially with transcript models generated from PacBio long-reads.

We applied this strategy to the annotation of the Trichechus manatus latirostris genome, annotating 25% and 12.3% more BUSCO genes than using only experimental data or ab initio predictions, respectively. Our analysis indicates that lrRNA-seq is a valuable source of experimental data to support gene annotation in mammalian species.

Genomes

FrameRate: predicting coding frames direct from unassembled DNA reads

Wang Liu-Wei (Robert Koch Institute - International Max-Planck Research School for Biology and Computing), Robert Hoehndorf (King Abdullah University of Science and Technology), Wayne Aubrey (Aberystwyth University), Christopher Creevey (Queen's University Belfast), Amanda Clare (Aberystwyth University) and Nicholas Dimonaco (Aberystwyth University - McMaster University).

Abstract:

Genome assembly is a slow and computationally intensive process, needing iterative rounds for improvement and completeness. Most importantly, an assembly often fails to incorporate many of the reads from a sequencing run. Additionally, further complications such as reduced read-depth and chimeric assembly are especially prominent in the assembly of metagenomic datasets.

Many of these limitations could potentially be overcome by exploiting the information content stored in the reads directly and thus eliminating the need for assembly in a number of situations. In this study, we explore the prediction of coding potential of DNA reads by training a machine learning model on existing protein sequences. Named `FrameRate', this model can predict the coding frame(s) from unassembled DNA sequencing reads directly, and thus greatly reduces the computational resources required for genome assembly, homology-based inference or pre-computed databases.

Using the eggNOG-mapper function annotation tool, the predicted coding frames were functionally compared to full-length protein sequences identified through an established metagenome assembly and gene prediction pipeline, produced from the same metagenomic sample. FrameRate captured comparable functional profiles from the coding frames while reducing the required storage and time resources significantly. Interestingly, FrameRate was also able to annotate the previously unassembled reads, essentially profiling the entirety of the dataset. As an ultra-fast read-level assembly-free coding profiler, FrameRate enables us to quickly characterise almost every sequencing read directly, whether it can be assembled or not, and thus circumvent many of the problems caused by contemporary assembly tools.

Genomes

Genetic diversity in Amaranthaceae crops

Felix Leopold Wascher (University of Natural Resources and Life Sciences, Vienna, Institute of Computational Biology), Nancy Stralis-Pavese (University of Natural Resources and Life Sciences, Vienna, Institute of Computational Biology), Heinz Himmelbauer (University of Natural Resources and Life Sciences, Vienna, Institute of Computational Biology) and Juliane C. Dohm (University of Natural Resources and Life Sciences, Vienna, Institute of Computational Biology).

Abstract:

The plant family Amaranthaceae (order Caryophylalles) comprises several important crop plants like quinoa, spinach and sugar beet. Of special interest for our research are the species Chenopodium quinoa (quinoa) and Beta vulgaris subsp. vulgaris (cultivated beets, e.g. sugar beet) that rank among the agriculturally most important crops world-wide. Within our group we have generated reference genomes for quinoa, spinach, sugar beet as well as wild beet species using short-read and long-read sequencing data. Additionally, we have performed a massive re-sequencing effort resulting in the availability of whole genome sequencing data for hundreds of Chenopodium and Beta accessions. We are in the course of analysing the genomic diversity in these plants making use of published and unpublished data, and employing state-of-the-art machine learning algorithms. Our goals are to characterise the genetic diversity within the species, to identify regions within the genomes that play important roles in controlling phenotypic traits, and to search for genomic regions that were involved in the domestication of crops and in the diversification of varieties. To that end we compare a large number of cultivated accessions to each other and to their wild relatives. Analysing the genetic diversity of cultivated plants and their wild relatives is important to adapt modern crops to the agricultural challenges of the future. We will present the status of our work and outline future perspectives.

Genomes

Genome sequencing and comparative genomic analysis of turmeric plant provide insights into adaptive evolution of its medicinal properties

Abhisek Chakraborty (Indian Institute of Science Education and Research (IISER), Bhopal), Shruti Mahajan (Indian Institute of Science Education and Research (IISER), Bhopal), Shubham K. Jaiswal (Indian Institute of Science Education and Research (IISER), Bhopal) and Vineet K. Sharma (Indian Institute of Science Education and Research (IISER), Bhopal).

Abstract:

Plant genomes offer wide range of prospects because of the medicinal properties and the contribution in agricultural science conferred by most of the species. Availability of various sequencing technologies and hybrid genome assembly approaches has aided us in characterization of complex plant genomes, and in understanding the genomic basis of evolutionary adaptation. Curcuma longa (turmeric) is traditionally known for its enormous therapeutic applications. However, the unavailability of the reference genome sequence was a restraining factor in understanding the genomic basis of origin of its medicinal properties. We constructed the draft genome of C. longa using 10X Genomics linked-read (~82x coverage) and Oxford Nanopore long-read (~41x coverage) technologies. C. longa genome is triploid, and the draft genome assembly had a size of 1.02 Gb with high heterozygosity (4.83%), containing 70% repetitive sequences and 50,401 coding gene sequences. The phylogenetic position of C. longa species was resolved through a comprehensive genome-wide analysis including 16 other Angiosperm species. Analyses of the key enzymes involved in curcuminoid biosynthesis pathway revealed the gene structures, and showed that the enzymes have evolved from their ancestor homologous genes in distant species to play key role in this pathway. Comparative evolutionary analyses across 17 Angiosperm species revealed evolutionary signatures in C. longa genes involved in secondary metabolism, phytohormones signaling, and stress tolerance responses. These mechanisms are vital for perennial and rhizomatous plants like C. longa for its defense responses via secondary metabolites production, which are linked with the wide range of therapeutic potential of C. longa.

Genomes

Germline genetics correlates with aberrant signaling pathways in cancer

Davide Dalfovo (Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento), Riccardo Scandino (Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento), Marta Paoli (Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento), Samuel Valentini (Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento) and Alessandro Romanel (Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento).

Abstract:

Cancer is a complex disease influenced by a heterogeneous landscape of both germline genetic variants and somatic aberrations. Although increasing evidences suggest an interplay between germline and somatic variants and a large number of somatic aberrations in specific pathways are now used as hallmarks in many well-known forms of cancer, the interaction landscape between germline variants and the aberration of those pathways in cancer is still largely unexplored.

We conducted a set of genome-wide association studies (GWAS) using >8,500 human samples across 33 cancer types characterized by TCGA, considering 40 traits defined using a large collection of somatic aberration types across 10 well-known oncogenic signaling pathways.

Functional links were established between associated variants and the corresponding oncogenic signaling pathways using genomic and transcriptomic data. Polygenic somatic scores (PSS) were built, using an ad-hoc approach. PSS demostrated association against cancer subtypes, cancer-specific clinical variables and patient's survival data and were validated using PCAWG and CCLE data.

We identified 264 genome-wide ($p<5e-8$) and suggestive ($p<1e-6$) significant associations between 239 SNPs and 34 traits. 119 associated SNPs revealed cis-associations with genes functionally correlated with the corresponding traits oncogenic pathways.

TCGA-based PSS of 23 traits individually demonstrated an average AUC>0.55 and 5, among 13 that could be tested, demonstrated consistent statistically significant shifts across PCAWG patients; three were further confirmed in CCLE cell lines samples.

We showed that germline genetics shapes susceptibility to somatic aberrations in oncogenic signaling pathways and that polygenic scores can describe patients' genetic liability to develop specific cancer molecular profiles.

Genomes

GPU-Accelerated Pairwise Sequence Alignment using the Wavefront Algorithm

Quim Aguado-Puig (Universitat Autònoma de Barcelona), Santiago Marco-Sola (Barcelona Supercomputing Center), Juan Carlos Moure (Universitat Autònoma de Barcelona), Christos Matzoros (Barcelona Supercomputing Center), David Castells-Rufas (Universitat Autònoma de Barcelona), Antonio Espinosa (Universitat Autònoma de Barcelona) and Miquel Moreto (Barcelona Supercomputing Center).

Abstract:

Advances in genomics and sequencing technologies require faster and scalable analysis methods that can process longer sequences with higher accuracy. However, classical pairwise alignment methods based on dynamic programming impose impractical computational requirements to align long and noisy sequences like those produced by PacBio and Nanopore technologies. The recently proposed WFA algorithm introduces an efficient method for pairwise alignment, improving time and memory complexity over classical methods. Notwithstanding, modern high-performance computing (HPC) platforms rely on accelerator-based architectures that exploit parallel computing resources to improve over classical computing CPUs. We present WFA-GPU, a GPU-accelerated solution based on the WFA algorithm for gap-affine pairwise alignment. We use a combination of inter and intra-sequence parallelization combined with an algorithmic adaptation of the WFA to exploit the massively parallel capabilities of GPUs by reducing the memory requirements. As a result, our implementation outperforms the original WFA CPU implementation between 1.5-7.7X. Compared to other tools and libraries, WFA-GPU is up to 175X faster than other GPU solutions and up to four orders of magnitude faster than other CPU implementations.

Genomes

GWAS of vessel diameter, number of bifurcations, and main temporal angles identifies genetic variants in OCA2, HERC2 and other genes to modulate these traits

Sofia Ortin Vela (Dept. of Computational Biology, University of Lausanne), Michael Johannes Beyeler (Dept. of Computational Biology, University of Lausanne), Mattia Tomasoni (Jules-Gonin Eye Hospital, Lausanne, Switzerland), Olga Trofimova (Dept. of Computational Biology, University of Lausanne), Florence Hoogewoud (Jules-Gonin Eye Hospital, Lausanne, Switzerland) and Sven Bergmann (Dept. of Computational Biology, University of Lausanne).

Abstract:

Fundus images of the eye allow for non-invasive inspection of the microvasculature of the retina, while characterising the vasculatures of other body parts is far more challenging. For this reason, retinal images have great potential for better diagnosis and risk assessment not only for ocular but also many vascular diseases.

Here we present results for nine image-derived vascular phenotypes, including vessel diameter, number of bifurcations, and main temporal angles, extracted from around 120 000 colour fundus images from the UK Biobank. We then carried out genome-wide association studies (GWAS) of these phenotypes. Their heritability ranges between 5 and 21%. We evaluated the phenotypic and genetic similarity between different phenotypes, revealing substantial variability across trait pairs. Linking genetic associations to genes, we identified several genes significantly associated with multiple phenotypic traits, in particular OCA2 and HERC2, which played a role for most of the traits we analysed. We show that several retinal phenotypes are associated with systemic and ophthalmological diseases not only at the phenotypic level, as previously reported, but also at the genetic level.

Our analysis highlights the potential of retinal vascular traits as intermediate phenotypes that share genetic loci with common diseases and may help reveal their molecular underpinning.

Genomes

Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data

Rebecca Serra Mari (Heinrich Heine University Düsseldorf), Sven Schrinner (Heinrich Heine University Düsseldorf), Richard Finkers (Gennovation B.V. Wageningen), Paul Arens (Wageningen University & Research), Maximilian H.-W. Schmidt (Forschungszentrum Jülich), Björn Usadel (Heinrich Heine University Düsseldorf), Gunnar W. Klau (Heinrich Heine University Düsseldorf) and Tobias Marschall (Heinrich Heine University Düsseldorf).

Abstract:

Potato is one of the world's major staple crops and like many important crop plants it has a polyploid genome. Because of its high agronomic value, efforts to assemble potato genomes, especially to haplotype resolution, are of crucial importance. However, polyploid haplotype assembly poses a major computational challenge, hindering the use of genomic data in breeding strategies.

Here, we introduce a novel strategy for the de novo assembly of polyploid genomes and present an assembly of the autotetraploid potato cultivar Altus. Our method uses low-depth sequencing data from an offspring population, which is available in many plant breeding settings, to achieve chromosomal clustering and haplotype phasing directly on the assembly graph. This involves two novel strategies for (i) the analysis of k-mers unique to specific graph nodes and (ii) finding resulting graph traversals to identify and assemble the four haplotypes.

We combined accurate PacBio HiFi reads of the cultivar Altus sequenced to 96x coverage with cost-effective low-coverage short reads from 193 offspring of the two cultivars Altus and Colomba. Our approach generates haplotype-resolved assemblies of individual chromosomes with phased haplotig N50 values of up to 13 Mb and haplotig lengths of up to 31 Mb. We show that our assembly maps well to the current monoploid potato reference genome.

This major advance provides high-quality assemblies with haplotype-specific sequence resolution of whole chromosome arms and is immediately accessible in common breeding and research scenarios where collections of offspring are available.

Genomes

Human glycosylation gene variants as possible contenders to explain differences in SARS-CoV-2 disease severity in Indian sub populations.

Bithika Chatterjee (National Centre for Cell Science) and Shekhar Mande (National Centre for Cell Science).

Abstract:

Human glycosylation gene variants as possible contenders to explain differences in SARS-CoV-2 disease severity in Indian sub populations.

The Covid-19 pandemic has exhibited a variable pattern of disease susceptibility amongst individuals. Under such circumstances it is imperative to understand the genetic basis of such variation. While genome-wide association studies of Covid-19 has discovered a plethora disease causing variants, the biological explanation as to which variant may be holding the key for disease susceptibility remains challenging.

Here we focused our research on host glycosylation genes since it plays an important machinery in the viral binding, entry and attachment to the host cell as well as the immune response/escape mechanism of the infection. We searched for common coding region variants in O and N linked glycosylation genes and compared their polymorphism frequencies amongst Indian, European and African populations to get significantly different variants. We used these filtered variants to superimpose with genome-wide association studies on critical covid patients to identify disease associated variants. After confirming similar pattern of Linkage disequilibrium between the patient's population and the Indian population genomes we proceed to compare the polymorphism frequencies in Indian sub populations. We found several candidates displaying variable allele frequencies that we further investigated for their structure, expression and functional role in causing disease susceptibility in different ethnicity. To strengthen our glycosylation genes hypothesis we also analyzed SARS-CoV-2 viral sequences to measure the proportion of amino acids mutations, finding the glycosylated related amino acids mutations frequencies higher than the others.

Genomes

Identification and characterization of the microbiome, resistome, and mobilome in the wastewater treatment plant

Stephanie Pillay (Delft University of Technology), Ramin Shirali Hossein Zade (Delft University of Technology) and Thomas Abeel (Delft University of Technology).

Abstract:

Illnesses caused by antibiotic-resistant (AMR) bacteria have led to 1.2-5 million people dying worldwide. This number is continuously rising as bacteria are gaining resistance to more antibiotics over time. AMR bacteria contain mobile genetic elements (MGEs) which facilitate their spread. The wastewater treatment plant (WWTP) provides a favourable environment for bacteria to gain resistance by MGEs which helps the spread of AMR to humans and animals via water usage/consumption. This is the first study to use metagenomic datasets from all sectors of the WWTP; upstream, influent, activated sludge, effluent, and downstream to identify and characterize the microbiome, resistome, and mobilome. The abundance of pathogenic bacteria resistant to commonly prescribed antibiotics increased in the influent and activated sludge with no significant decrease in the effluent. MGEs, i.e., integrative chromosomal elements, integrons, and plasmids, were found in all sectors and were associated with common antibiotics. Plasmids carrying AMR genes had a total of 51 different resistance profiles. 46 different multi-drug resistance (MDR) profiles were found, the plasmids of which were released at a high abundance with the effluent after treatment. This indicates that AMR bacteria are carrying MGEs such as plasmids with multiple resistance genes that spread to humans and animals. This results in more illnesses that cannot be treated with antibiotics and an increasing morbidity and mortality rate worldwide. This detailed characterization of AMR on plasmids and other MGEs in the context of the WWTP informs us on how to reduce the spread and effects of AMR in the future.

Genomes

Identification of epivariations in rare diseases from a single patient perspective.

Robin Grolaux (Université Libre de Bruxelles), Alexis Hardy (Université Libre de Bruxelles) and Matthieu Defrance (Université Libre de Bruxelles).

Abstract:

DNA methylation is being widely recognized as a surrogate marker of genetic variants or primary marker that can be used in the diagnosis of rare neurodevelopmental and imprinting disorders. In addition, identification of variations in DNA methylation plays an important role for understanding the etiology of those diseases. Canonical pipelines for the detection of variations in methylation levels (i.e. epivariants) based on methylation-array technologies (such as the 450k and EPIC Infinium arrays) rely on case-control groups comparisons. However, in the context of rare diseases and multi-locus imprinting disturbances, small cohorts and inter-patients' heterogeneity prevent the use of these tools. Therefore, there is a need for a comprehensible and statistically robust pipeline that perform analyses at the single patient level. This poster describes a statistical method to detect differentially methylated regions in correlated datasets based on the z-score and the empirical Brown aggregation method from a single patient perspective. It further provides a characterization of how the chosen parameters may influence epivariants detection. We generated semi-simulated data based on a public control population of 521 samples. This enabled us to evaluate how control population, effect and region size affect the performance of epivariants detection, in order to define the optimal parameters of the method. Finally, we validated the detection of pathological methylation events in patients suffering from rare multi-locus imprinting disturbances and showed how this method is complementary to the validation of clinical diagnosis.

Genomes

Identification of genes under positive selection in E. coli: Focusing on antibiotic resistance

Negin Malekian Boroujeni (TU Dresden).

Abstract:

Background: Failures in antibiotic therapy result in hundreds of thousands of deaths each year. Having a good knowledge of the evolution of antibiotic resistance and the genomics behind it assists in the development of successful drugs that are less prone to evolving resistance.

Here, we investigate how genome-wide positive selection screening in 92 diverse E. coli genomes from wastewater may identify genes that are critical in the evolution of this species under all potential selective pressures and, in particular, antibiotic resistance.

Description: We identified 75 genes under positive selection by normalizing non-synonymous mutations against an overall mutation burden. We found that some of these genes have known functions relevant to antibiotic resistance, such as biofilm formation, reduction of outer membrane permeability, efflux pumps, and bacterial persistence. Moreover, we correlated the mutations in these genes with resistance to the 20 most commonly prescribed antibiotics. Mutations in two genes, the porin ompC and the bacterial persistence gene hipA, were correlated with antibiotic resistance. Also, we found that these mutations are on the protein's surface, and therefore they may directly impact structure and function.

Conclusion: The positive selection analysis of a large number of genomes has found both known and novel genes and mutations that contribute to the evolution of antibiotic resistance.

Genomes

Identification of genome edited cells using CRISPRnano

Thach Nguyen (IUF – Leibniz Research Institute for Environmental Medicine), Haribaskar Ramachandran (IUF – Leibniz Research Institute for Environmental Medicine), Soraia Martins (IUF – Leibniz Research Institute for Environmental Medicine), Jean Krutmann (IUF – Leibniz Research Institute for Environmental Medicine) and Andrea Rossi (IUF – Leibniz Research Institute for Environmental Medicine).

Abstract:

Motivation: Genome engineering-induced cleavage sites can be resolved by non-homologous end joining (NHEJ) or homology-directed repair (HDR). Identifying genetically modified clones at the target locus remains an intensive and laborious task. Different workflows and software that rely on deep sequencing data have been developed to detect and quantify targeted mutagenesis. Nevertheless, these pipelines require high-quality reads generated by Next Generation Sequencing (NGS) platforms.

Results: We have developed a robust, versatile, and easy-to-use computational webserver named CRISPRnano that analyzes low-quality reads generated by affordable and portable sequencers, including Oxford Nanopore Technologies (ONT) devices. CRISPRnano allows fast and accurate identification, quantification, and visualization of genetically modified cell lines. It is compatible with NGS and ONT sequencing reads and can be used without an internet connection.

Availability:

CRISPRnano website is free and open to all users at www.CRISPRnano.de

The source code of the website is available at:

https://github.com/thachnguyen/CRISPRnano

Genomes

Identification of regions with high evolutionary variability on the SARS-CoV-2 S1 and S2 surface glycoprotein.

Katrina Norwood (Computational Biology of Infection Research, Helmholtz Centre for Infection Research), Alice McHardy (Computational Biology of Infection Research, Helmholtz Centre for Infection Research), Susanne Reimering (Computational Biology of Infection Research, Helmholtz Centre for Infection Research) and Thorsten Klingen (Formerly at Computational Biology of Infection Research, Helmholtz Centre for Infection Research).

Abstract:

Over the course of the pandemic, SARS-CoV-2 has demonstrated its propensity to evolve adaptively, particularly in the major surface glycoprotein S1 subunit, thus optimizing its capability to spread efficiently in the human population. Some of these mutations have an important bearing on the alteration of the virus' antigenic phenotype. Identification of such sites on the SARS-CoV-2 spike glycoprotein is therefore important for the monitoring and detection of variants with altered antigenic phenotype as well as improving vaccine efficacy through the development of more targeted vaccines.

Using German GISAID viral genomic sequences from December of 2019 to April of 2021, a total of 13GB of sequencing data, we identified seven patches of evolutionary variable sites in the SARS-CoV-2 spike glycoprotein based on recurrent substitutions inferred from a phylogenetic tree that grouped into spatially distinct clusters on the protein structure. These patches fall into regions known to impact viral fitness, ie. by facilitating immune escape or binding to the host angiotensin-converting enzyme 2 receptors (ACE2).

The largest patch resides on the N-terminal domain, and another in the receptor binding domain, a region responsible for the binding of ACE2. Interestingly, a patch was also identified in the Heptad repeat 1, which has been shown to mediate viral fusion and entry into the host cell along with the Heptad repeat 2. Overall, these results confirm previous findings as well as highlight regions of relevance for viral adaptation which have not been characterized so far. These regions offer new avenues for more detailed explorations.

Genomes

Identification of viral miRNAs

Alexandra Schubö (LMU Munich), Armin Hadziahmetovic (LMU Munich), Leonie Pohl (LMU Munich), Markus Joppich (LMU Munich) and Ralf Zimmer (LMU Munich).

Abstract:

miRNAs are important post-transcriptional regulators in eukaryotes and previous research revealed their involvement in viral infection. Viral miRNAs have been shown to be encoded in the genomes of DNA-viruses and, more recently, RNA viruses like SARS-CoV-2, where they partake in the interplay between virus and host.

In previously published research, we collected a comprehensive set of 574 viral miRNA candidates from 21 publications for SARS-CoV-2, which are encoded in the SARS-CoV-2 genome.

Using 86 publicly available small RNA-seq sequencing runs, spanning various samples, conditions, and cell lines, we employ two approaches to identify supporting evidence for miRNA candidates: (1) Given the set of candidates we check the small RNA-seq reads for matches. We assume that the reads capture expressed miRNA and, therefore, exact matching the collection of viral miRNA candidates provides evidence of miRNA expression in the respective transcriptome. (2) Given processed and mapped read data, a sliding window expression difference approach determines novel miRNA-like short sequences. Regions containing miRNA candidates with significantly higher counts as compared to the surrounding region indicate potential expressed miRNA genes. Reconstructed hairpin structures from these potential pre-miRNAs are further experimental evidence for real and expressed viral miRNAs (svRNAs).

In this poster, we identify expression data evidence for a set of potential SARS-CoV-2 encoded miRNAs. The approach presented here can be used to identify miRNAs for other viruses which is the basis to compare the miRNA arsenal of e.g. SARS-CoV2 and SARS-CoV1.

Genomes

Impacts of COVID-19 on the Microbiome: A Bioinformatics and Machine Learning Study

Bertalan Takács (Biological Research Centre, Institute of Genetics, HCEMM-BRC Mutagenesis and Carcinogenesis Research Group, Szeged), Zoltán Gyuris (Delta Bio 2000 Ltd., Szeged), Lajos Pintér (Delta Bio 2000 Ltd., Szeged), Ádám Visnyovszki (University of Szeged, Faculty of Medicine, Division of Infectiology, Szeged), Nóra Éva Nagy (University of Szeged, Faculty of Medicine, Division of Infectiology, Szeged), Márton Zsolt Enyedi (Delta Bio 2000 Ltd., Szeged), Edit Hajdú (University of Szeged, Faculty of Medicine, Division of Infectiology, Szeged) and Lajos Haracska (Biological Research Centre, Institute of Genetics, HCEMM-BRC Mutagenesis and Carcinogenesis Research Group, Szeged).

Abstract:

SARS-Cov-2, despite primarily attacking the respiratory system, can also affect the gut and its microbiome, thus, exploring its effects on the latter has become a hot topic in microbial research during the ongoing pandemic. To gain insight into the impact of the virus on the microbiota, we analyzed the microbial composition of patients during and after COVID-19, followed by a comparison to healthy controls. For each patient, samples were taken from three areas of the body: nose, pharynx, and the gut. Changes in the microbial composition caused by the infection at both the taxonomic and genetic levels were analyzed utilizing bioinformatics and machine learning methods. The following results are preliminary and part of a larger study involving more participants and a diverse set of clinical and immunological parameters.

We found that COVID-19 causes an observable change in the microbial composition of all three niches, which persists after the viral particles are gone. The Campylobacter and Escherichia genera showed the most significant differences in both infected and recovered patients.

With functional metagenomics, we demonstrated that these changes can also be observed at the genetic level. Infection caused underrepresentation of genetic pathways associated with healthy microflora, whereas dysbiosis-related pathways were overrepresented.

These results help us gain a better understanding of the impact of viral infections on the body and the recovery of microbiota after such an event, which can be beneficial for designing therapeutic methods for the recovery from viral respiratory infections with probiotics and bacterial microbiota transplantation.

Genomes

Improving multiplicity assignments in de Bruijn graphs using Loopy Belief Propagation inference on Conditional Random Fields.

Aranka Steyaert (Ghent University - imec), Pieter Audenaert (Ghent University - imec) and Jan Fostier (Ghent University - imec).

Abstract:

De Bruijn graphs are used to determine the overlap between Illumina reads for genome assembly. Nodes represent all k-mers present in the reads, while an arc represents a k+1-mer present in the reads such that its source node's last k-1 characters overlap with its sink node's first k-1 characters. The multiplicity of a node/arc, i.e., the number of times its corresponding sequence is present in the genome, is unknown. Accurately estimating the multiplicity of nodes and arcs in a de Bruijn graph is crucial when performing error correction (based on nodes/arcs with multiplicity 0) or repeat resolution (using multiplicities higher than 1); both are important to obtain an accurate and contiguous assembly.

Steyaert et al. (BMC Bioinformatics 21, 402, 2020) presented a Conditional Random Field Model that improves multiplicity assignment accuracy compared to methods that existing error correction tools use. Now we expand this method with approximate inference computations using Loopy Belief Propagation (LBP). We determine all multiplicities within feasible runtimes and achieve higher accuracy than the previous model. Additionally, we empirically show how LBP can be successfully implemented as an alternative to the more frequently used sampling-based approximate inference for Conditional Random Fields with higher order interactions. The successful convergence of LBP depends on the order in which computations are performed. However, little theoretical guarantees exist to guide the choice of that order of computations. We believe our empirical evaluation may guide others in their design choices when using LBP as an alternative to sampling-based methods.

Genomes

Inferring mutational signatures from distinct copy-number events

Tom L Kaufmann (Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany.) and Roland F Schwarz (Cancer Research Center Cologne Essen (CCCE), University Hospital and University of Cologne.).

Abstract:

Somatic copy number alterations (SCNA) include large-scale events, such as chromosome arm-level gains and losses as well as focal amplifications and deletions and play a key role in the evolutionary processes that shape cancer genomes. In the case of small-scale events such as point mutations and indels, there exists a list of established mutational signatures that can be linked to distinct exogenous or endogenous exposures such as tobacco use. Despite previous efforts, accurate and meaningful copy-number signatures are still elusive. The biggest obstacle in creating copy-number signatures is that due to their cascading nature, traditional segment-based representations of copy number do not reveal individual evolutionary events and the order in which they occurred.

Here we introduce a new method for deriving copy-number signatures that explicitly models evolutionary copy-number events using our phylogenetic copy-number model MEDICC2 (Kaufmann 2021, bioRxiv). By combining the evolutionary events identified by MEDICC2 with structural variant data and the timing of point mutations, we derive event-based copy-number signatures that include the order and timing of events. We demonstrate the power of our approach on an independent simulation of mutational processes and real world data from 2,778 tumors from the Pan Cancer Analysis of Whole Genomes (PCAWG) and demonstrate how the extracted copy-number signatures reveal novel insights into the nature of the mutational processes shaping cancer genomes.

Genomes

Influence of motif interactions on post-hoc attribution methods in genomic CNNs

Marta S. Lemanczyk (Hasso Plattner Institute for Digital Engineering, University of Potsdam), Jakub M. Bartoszewicz (Hasso Plattner Institute for Digital Engineering, University of Potsdam) and Bernhard Y. Renard (Hasso Plattner Institute for Digital Engineering, University of Potsdam).

Abstract:

Background: Convolutional neural networks (CNN) are capable of learning patterns and complex interactions between features. This makes them a useful tool for biological sequence-based tasks. Post-hoc interpretability methods, like Integrated Gradients or DeepSHAP, are applied on trained CNNs to identify regions in the input sequence that contributed to the model's decision and indicate biologically relevant motifs. To verify the interpretations, a common approach is to compare identified motifs with known motifs from task-specific databases. However, this approach does not ensure completeness of all motifs contributing to a given outcome. Feature interactions can affect those methods and result in misleading interpretations.

Results: In this work, we investigate post-hoc interpretability of models trained on interacting motif data. First, we define two groups of interactions that can occur in sequence-based tasks for CNNs: (1) Biological interactions that represent the different logical relations of motifs and their effect on the outcome, and (2) interactions resulting from the influence of design and data selection on the model. Following these definitions, we generate genomic sequence data containing homogenous and heterogenous motif subsets based on real motifs to ensure a suitable ground truth for the comparison between models trained on interactive and non-interactive sequences. We show that the performance of post-hoc interpretability methods decreases when motif interactions are introduced to the data.

Genomes

Interpretable deep learning for phage life cycle prediction

Melania Nowicka (Hasso Plattner Institute), Jakub M Bartoszewicz (Hasso Plattner Institute) and Bernhard Y Renard (Hasso Plattner Institute).

Abstract:

Background: Rapidly emerging antibiotic-resistant bacteria pose a threat to the efficacy of antibiotics, endangering public health worldwide. Phage therapy, known for decades but significantly under-researched, provides a promising alternative. Phages are natural antibacterial agents that can be applied to treat bacterial infections. However, selecting appropriate phages is notoriously challenging. One of the optimality criteria for the design of phage-based therapeutics is the life cycle type that a phage follows after entering the host cell. Here, lytic phages that replicate immediately after entering the host cell, causing lysis and resulting in a cell's death, are the most suitable candidates.

Results: We repurpose the DeePaC framework, initially developed to predict the pathogenic potential of novel species, to indicate the life cycle of phages as lytic or non-lytic. We train deep residual networks (ResNets) on labelled data acquired from publicly available databases, focusing on new phages belonging to either known or entirely novel phage clusters. Hence, we perform the training-test split by single phage or phage cluster, respectively. Further, we visualize the lytic potential of genomes to identify virulence-associated genes. The models achieve 99.4% and 91.7% accuracy for the phage-wise and cluster-wise classification. However, we demonstrate that only the cluster-wise models provide an insight into regions of genomes potentially influencing the life cycle, stressing the importance of avoiding information leakage to ensure proper generalization.

Genomes

Intrinsic linking of chromatin in human cells

Maciej Borodzik (Institute of Mathematics, University of Warsaw, ul. Banacha 2, 02-097 Warsaw, Poland), Michał Denkiewicz (Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland), Krzysztof Spaliński (Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland), Kamila Winnicka (Centre of New Technologies, University of Warsaw, ul. Banacha 2c, 02-097 Warsaw, Poland), Sevastianos Korsak (Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland), Kaustav Sengupta (Centre of New Technologies, University of Warsaw, ul. Banacha 2c, 02-097 Warsaw, Poland), Marcin Pilipczuk (Institute of Informatics, University of Warsaw, ul. Banacha 2, 02-097 Warsaw, Poland), Michał Pilipczuk (Institute of Informatics, University of Warsaw, ul. Banacha 2, 02-097 Warsaw, Poland), Yijun Ruan (The Jackson Laboratory for Genomic Medicine, USA) and Dariusz Plewczynski (Centre of New Technologies, University of Warsaw, Warsaw, Poland).

Abstract:

We propose a practical algorithm based on graph theory, with the purpose of identifying CTCF-mediated chromatin loops that are linked in 3D space. Our method is based on finding certain graph structures, K6 minors, in graphs constructed from pairwise chromatin interaction data obtained from ChIA-PET experiments. We show that such a graph structure, representing a particular arrangement of loops, mathematically necessitates linking, if co-occurring in an individual cell.

We apply our method to graphs created from in situ ChIA-PET data for GM128787, H1ESC, HFFC6 and WTC11 cell lines, and from long-read ChIA-PET data for GM12878. We find numerous regions with minors, indicating the presence of links and study their charcteristics and location with respect to chromatin compartments.

Genomes

JLOH: Extracting Loss of Heterozygosity Blocks from Short-Read Sequencing Data

Matteo Schiavinato (Barcelona Supercomputing Center (BSC-CNS)).

Abstract:

Loss of heterozygosity (LOH) happens when a heterozygous genome loses one of the two alleles at a locus. This may have an evolutionary advantage in highly unstable genomes such as those of hybrids. By extracting LOH from a hybrid we understand which alleles were beneficial in its evolution, which is relevant in wild, clinical, and industrial settings. The genomic properties of hybrids are still, however, poorly understood. LOH are studied with reliable short-read sequencing data, but the downstream analysis is often done with custom scripts that reduce reproducibility and do not to leverage the power of a computing cluster. Here we present a program called “JLOH” that streamlines LOH extraction from sequencing data maximizing parallel computing. We simulate a series of divergent genomes from the S. cerevisiae genome, introducing LOH blocks in variable amounts in them. We then attempt to retrieve these blocks from common SNP data with JLOH. We form simulated hybrid genomes by concatenating the simulated divergent genomes and a copy of the S. cerevisiae genome. We simulate reads from these hybrid genomes and call crossmapping SNPs between their subgenomes. We identify SNP-depleted regions as candidate LOH blocks, assess them in terms of length and coverage, and compare them against the LOH blocks originally introduced in the simulated genomes. We conclude that, in most configurations, JLOH successfully finds most of the LOH blocks that were artificially introduced, without finding many false positives. We also conclude that subgenomic divergence in a hybrid may limit JLOH’s precision.

Genomes

k-mer and GWAS approaches to identify host-specific genomic determinants in Klebsiella pneumoniae

Konstantinos A. Kyritsis (Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece), Georgios-Nikolaos Kartanos (Laboratory of Microbiology and Infectious Diseases, School of Veterinary Medicine, Aristotle University of Thessaloniki), Victoria Siarkou (Laboratory of Microbiology and Infectious Diseases, School of Veterinary Medicine, Aristotle University of Thessaloniki) and Fotis Psomopoulos (Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece).

Abstract:

Klebsiella pneumoniae is an important Gram-negative opportunistic bacterial pathogen that causes a variety of community and healthcare-associated infections in human and animals. The emergence and spread of multidrug resistant K. pneumoniae strains is now recognized as an urgent threat to public health worldwide. However, the epidemiology of K. pneumoniae has not been extensively studied and reservoirs of the organism have been rarely investigated. Hence, understanding and monitoring of K. pneumoniae transmission across host species is of paramount importance. In this study, we aimed to identify host-associated genomic differences across various K. pneumoniae strains originated from human and animals. Machine Learning (ML) classification models were trained for host prediction on pre-processed 9-mer count data from 706 publicly available whole genomes (European Nucleotide Archive, ENA). Model performance reached over 85% (accuracy and f1-score), showcasing the presence of host-associated differences. Pangenome and genome-wide association analyses (GWAS) were further employed for interpretation. Several acquired accessory genes, implicated in pathways such as carbohydrate metabolism, iron binding, and antibiotic resistance, were identified in agreement with recent studies. Interestingly, we also detected novel host-associated acquired genes related to stress-response. Our results support the application of ML algorithms and k-mers in parallel with GWAS workflow for the epidemiological surveillance of human and zoonotic transmission throughout K. pneumoniae outbreaks. Furthermore, we validate recently reported K. pneumoniae accessory genome variations and present novel ones that could associate with host specificity and/or reflect the selective pressure exerted on commensal and pathogenic bacteria by the excessive use antibiotics.

Genomes

Leveraging patient-level metatranscriptomics data and phase variation predictions to prioritize bacterial antigens

Bart Cuypers (Univeristy Of Antwerp), Pieter Meysman (University of Antwerp), Alessandro Brozzi (GlaxoSmithKline Biologicals (GSK), Siena, Italy), Normand Blais (GlaxoSmithKline Biologicals (GSK), Rixensart, Belgium), Christophe Lambert (GlaxoSmithKline Biologicals (GSK), Rixensart, Belgium) and Kris Laukens (University of Antwerp).

Abstract:

In 2020, the WHO warned that the clinical pipeline and recently approved antibiotics are insufficient to tackle the emerging antimicrobial resistance. Vaccines provide a promising and much-needed alternative strategy; however, many challenges remain in selecting potent antigens for bacterial vaccine development. A vital property of a candidate antigen is its (preferentially highly) expression in vivo, yet to our knowledge, no standardized reverse vaccinology pipelines currently include expression data. Therefore, as a proof-of-concept we developed an antigen-prioritization pipeline that leverages metatranscriptomics data to estimate pathogen gene expression levels in vivo. Our workflow uses Centrifuge and Kallisto to rapidly identify and map relevant pathogenic reads to the pathogen pangenome. Antigens can then be ranked on a range of gene expression parameters that our workflow provides (maximal/median expression, % of patients with expression). Antigens can additionally be selected using several standard reverse vaccinology criteria, such as subcellular compartment, degree of protein sequence conservation, MHC presentability and protein domain. Our workflow also annotates antigens that are prone to undergo phase variation (PV) by using PhaseFinder (Inverted Repeats) and Phasomeit (Simple Sequence Repeats). PV is an adaptive process by which a gene's expression can be switched entirely on or off, hence providing crucial information for antigen selection. Preliminary antigen expression-ranking validation was done by checking the rankings of experimentally validated immunogenic antigens from VaxiJen and Protegen databases. Supervised machine learning and their derived feature-importances were used to understand the predictive importance of expression- and other parameters.

Genomes

LIVE-DREAM: Live analysis of NGS data using an optimized hierarchical inter-leaved Bloom Filter index

Ferdous Nasri (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Tobias Loka (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam) and Bernhard Y Renard (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam).

Abstract:

As Next Generation Sequencing (NGS) tools become the gold-standard for sequencing clinical probes, it is vital to speed-up the analysis of the large amount of produced data. Illumina is established as the main workhorse due to its unmatched read quality and throughput. With LiveTools, we pioneered the first software to produce real-time sequence analysis results while NGS reads are being sequenced. Their results are crucial for quick pathogen detection or personalized medicine reporting. However, with an increasing number of pathogens and available genomic data, the underlying algorithms need to support large database indexing using modern data structures. For example, the DREAM index has been shown to reduce the size of database indexes and allow for quick query-time by an innovative use of approximate searching with Bloom Filters.

We focus on adapting the DREAM index to be able to include large databases, such as the full RefSeq, into the underlying hierarchical inter-leaved Bloom Filter and use it for fast querying and real-time result reporting. Our approach allows for a hierarchical separation of genomes based on taxonomy and sequence similarity, leading to quick approximate querying results and lower false discovery rates for each taxonomic level. At the same time, the integrated alignment procedure allows for nucleotide-level resolution and more comprehensive follow-up analysis. Overall, our adaption and integration of the DREAM index into the live analysis procedure enables a significant decrease of the time-to-result while still providing high-quality results based on large databases.

Genomes

Long non-coding RNAs and novel transcripts as the principal source of potential novel neoantigens in hepatocellular carcinoma

Marta E. Camarena (Hospital del Mar Research Institute (IMIM)), Júlia Perera-Bel (MARGenomics, Hospital del Mar Research Institute (IMIM)) and M. Mar Albà (Hospital del Mar Medical Research Institute (IMIM)).

Abstract:

Immunotherapy is a promising alternative to conventional therapies in cancer. However, the prediction of patients' response remains highly inaccurate, presumably because only a part of the factors that drive immunogenicity have been identified. Neoantigens derived from tumor-specific proteins, together with mutated peptides, can potentially trigger a response of T cells against the tumor. A potentially important class of neoantigens that remains poorly characterized is non-canonical peptides derived from the translation of open reading frames (ORFs) in long non-coding RNAs and novel transcripts. We hypothesize that this type of neoantigens, which are completely unrelated to self-proteins, could be highly immunogenic and contribute to explain the patients' differential response to immunotherapy.

We have developed an approach to identify and classify novel tumor-specific neoantigens derived from the translation of ORFs in non-coding and novel transcripts, and to estimate the affinity of the derived peptides for MHC I receptors, using RNA sequencing data from tumors and adjacent tissue. We have found that, in four hepatocellular carcinoma (HCC) patient cohorts, the number of putative tumor-specific neoantigens derived from non-canonical peptides is comparable or even higher to that of canonical peptides. Moreover, it is clearly much larger than the number originating from mutations in annotated coding sequences. This shows that non-canonical peptides might have a more important role in mediating cancer immunogenicity than previously anticipated, opening new avenues to develop new anti-cancer treatments such as vaccines.

Genomes

Looking at the overall picture: development of a metagenomic analyses pipeline for highly diverse microbial communities in heritage science

Monika Waldherr (Department of Applied Life Sciences, University of Applied Sciences, FH Campus Wien, Austria), Laura Rabbachin (Institute of Natural Sciences and Technology in the Art, Academy of Fine Arts Vienna, Austria), Johannes Tichy (Institute of Natural Sciences and Technology in the Art, Academy of Fine Arts Vienna, Austria), Guadalupe Piñar (Institute of Natural Sciences and Technology in the Art, Academy of Fine Arts Vienna, Austria), Martin Ortbauer (Institute for Conservation and Restoration, Academy of Fine Arts Vienna, Austria), Beate Sipek (Institute for Conservation and Restoration, Academy of Fine Arts Vienna, Austria) and Alexandra Graf (Department of Applied Life Sciences, University of Applied Sciences, FH Campus Wien, Austria).

Abstract:

Microorganisms are found almost anywhere in the world, including extreme environments providing high temperatures, salinity or pH. This high diversity makes cultivation very difficult and therefore the vast majority of microorganisms is still unclassified and their role unknown. But, the understanding of the microbial communities, containing bacteria, archaea and fungi, is tightly linked to the understanding of our environment and the impact of its changes.

Historical pieces of art and architectural surfaces suffer from continuos exposure to weathering processes, leading to irreversible damages. These processes are predicted to increase through climatic changes, challenging existing restauration methods. To invesitgate the role of microorganisms in that degradation pattern, we used long-read sequencing of whole-genome as well as 16S rRNA gene samples and developed a metagenomic analyses pipeline suitable for routine application on highly diverse microbial communities in heritage science. We reviewed the most promising available bioinformatic tools used for assembly of metagenomes and taxonomic classification and evaluated their performance on different use cases. While the high diversity in environmental samples definitely opens up the possibility of new findings, it also proved to make

assembly and classification not a trivial task and we found most tools as well as long-read sequencing are still in need of further development.

Nevertheless, with our approach we are able to identify part of the microbial composition found on the studied objects, providing the basis for further characterizations and functional analyses. The results contribute to the evaluation and development of more effective conservation methods for our cultural heritage.

Genomes

Lossless Approximate Pattern Matching on Pan-genome de Bruijn Graphs

Lore Depuydt (Ghent University - imec), Luca Renders (Ghent University - imec), Thomas Abeel (Delft University of Technology) and Jan Fostier (Ghent University - imec).

Abstract:

Pan-genome graphs are gaining importance in the field of bioinformatics as data structures to represent multiple genomes or natural variation within a population. In 2016, Beller and Ohlebusch proposed a memory-efficient, pan-genome de Bruijn graph representation that builds upon the FM-index. As such, it inherits the functionality to efficiently locate exact occurrences of a search pattern.

We extend the pan-genome de Bruijn graph data structure to support bidirectional search functionality and show that this enables efficient lossless approximate pattern matching using search schemes. Two classes of approximate pattern matching are considered: strain-fixed and strain-free pattern matching. Using strain-fixed pattern matching, only approximate occurrences that occur as a substring in one of the underlying pan-genome strains are reported. In contrast, strain-free pattern matching allows to freely navigate the nodes of the de Bruijn graph, and occurrences can be reported along any path of connected nodes in the graph. We demonstrate a proof-of-concept implementation that can locate all strain-free approximate occurrences of 100 000 Illumina reads, within an edit distance of 4, on a pan-genome de Bruijn graph (k=77) that consists of two human genomes (GRCh37 and GRCh38), in about 9 minutes.

We propose a memory-efficient, pan-genome graph data structure that supports fast, lossless approximate pattern matching using search schemes. The C++ source code of our software, called Nexus, is available at https://github.com/biointec/nexus under AGPL-3.0 license.

Genomes

Metagenetics versus metagenomics, a dual approach to reach strain-level resolution for starter culture monitoring

Cristian Díaz-Muñoz (Vrije Universiteit Brussel), Hannes Decadt (Vrije Universiteit Brussel), Luc De Vuyst (Vrije Universiteit Brussel) and Stefan Weckx (Vrije Universiteit Brussel).

Abstract:

A wide range of microbiological techniques have been applied to monitor the growth and prevalence of starter culture strains used in the production of fermented foods. Differentiating between inoculated starter culture strains and background strains from the same species spontaneously present in the fermenting matrices has been challenging. However, the intra-species diversity present in the microbiota of fermented food ecosytems plays an important role in the strain fitness for fermentation in terms of substrate utilization and stress resistance and is of major importance for their contribution to the formation of flavour compounds. The aim of the current study was to monitor starter culture strains inoculated to steer cocoa fermentation processes performed in Costa Rica at strain-level resolution. First, whole-community DNA-based, high-throughput, full-length 16S rRNA gene sequencing was performed using PacBio circular consensus sequencing and amplicon sequence variants (ASVs) were inferred using DADA2. Further, two shotgun metagenomics-based approaches were followed, namely one relying on species-specific marker genes using StrainPhlAn, and one relying on metagenome-assembled genome (MAG) reconstruction followed by average nucleotide identity (ANI) estimation. Albeit that the ASV-based strategy was time- and cost-effective to study microbial community dynamics at strain-level resolution, the application of novel metagenomic strategies provided more information and slightly higher resolution to recover the inoculated starter culture strains from complex ecosystems. Further, resolving different strains of the same species into independent MAGs allowed the study of intra-species differences at gene level that could be of interest to guide future starter culture trials for any fermented food ecosystem.

Genomes

metagWGS: a workflow to analyse short and long HiFi metagenomic reads

Joanna Fourquet (INRAE), Jean Mainguy (INRAE), Maïna Vienne (INRAE), Céline Noirot (INRAE), Vincent Darbot (INRAE), Pierre Martin (INRAE Occitanie), Olivier Bouchez (INRAE), Adrien Castinel (INRAE), Sylvie Combes (INRAE), Carole Iampietro (Inrae), Christine Gaspin (INRAE), Denis Milan (INRAE Toulouse/ GenPhySE / GeT facility), Cécile Donnadieu (INRAE), Geraldine Pascal (INRAE) and Claire Hoede (INRAE).

Abstract:

We are developing a complete, scalable, easy-to-use and reproducible workflow (with nextflow and singularity containers), metagWGS, able to process short Illumina and long HiFi PacBio reads (new feature still rare in this type of workflow) from shotgun metagenomics data. Using state of the art tools, it provides (i) contig assemblies, (ii) syntactic and functional annotations of genes, (iii) taxonomic affiliations of reads and contigs, (iv) a counting table of reads per non redundant gene and (v) contigs binning to obtain Metagenome-Assembled Genomes*.

The workflow begins by preprocessing steps that clean raw data from adapters, low quality reads and the host reads. The assembly is made by metaSPAdes or megahit for short reads and Hifiasm or metaFlye for long reads to generate contigs for each sample. This assembly can be realized per sample or as a co-assembly of several samples*.

Resulting contigs are annotated with Prokka. ORFs are clustered with CD-HIT using a 95% sequence identity cutoff to remove redundancy and generate a unique gene catalog between samples. Reads are mapped back to contigs and featureCounts is used to count the reads overlapping annotated genes. DIAMOND is used for the taxonomic affiliation of contigs versus nr database.

MetagWGS is available on https://forgemia.inra.fr/genotoul-bioinfo/metagwgs with a complete and up to date documentation.

* in current development

Acknowledgements

SeqOccIn project funded by FEDER (Programme Opérationnel FEDER-FSE_Midi-Pyrénées et Garonne 2014-2020), ATB_Biofilm project funded by PNREST Anses, 2020/01/142, Antiselfish project funded by LabEx ECOFECT, Université de Lyon, ExpoMycoPig project funded by France Futur Elevage.

Genomes

MetaProFi: An ultrafast chunked Bloom filter for storing and querying protein and nucleotide sequence data for accurate identification of functionally relevant genetic variants

Sanjay Kumar Srikakulam (Helmholtz Institute for Pharmaceutical Research Saarland), Sebastian Keller (Helmholtz Institute for Pharmaceutical Research Saarland), Fawaz Dabbaghie (Institute for Medical Biometry and Bioinformatics), Robert Bals (Department of Internal Medicine V - Pulmonology, Allergology, Intensive Care Medicine, University Hospital Saarland) and Olga Kalinina (Helmholtz Institute for Pharmaceutical Research Saarland).

Abstract:

Bloom filters are a popular data structure that allows rapid searches in large sequence datasets. So far, all tools work with nucleotide sequences; however, protein sequences are conserved over longer evolutionary distances, and only mutations on the protein level may have any functional significance. We present MetaProFi, a Bloom filter-based tool that, for the first time, offers the functionality to build indexes of amino acid sequences and query them with both amino acid and nucleotide sequences, thus bringing sequence comparison to the biologically relevant protein level. MetaProFi implements additional efficient engineering solutions, such as a shared memory system, chunked data storage, and efficient compression. In addition to its conceptual novelty, MetaProFi demonstrates state-of-the-art performance and excellent memory consumption-to-speed ratio when applied to various large datasets.

Genomes

Modeling dynamic interactions within human gut microbiome using statistical approach

Zuzanna Karwowska (Malopolska Center of Biotechnology, Jagiellonia University), Marcin Możejko (Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw), Paweł Szczerbiak (Malopolska Center of Biotechnology, Jagiellonia University), Ewa Szczurek (Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw) and Tomasz Kościółek (Malopolska Center of Biotechnology, Jagiellonia University).

Abstract:

Bacteria in the gut microbiome form a dynamic milieu due to their interactions with the host, the environment and most importantly with each other. However, most of the research on the gut microbiome uses cross-sectional data to find differences in microbiome composition between groups. This ignores the importance of the dynamic changes in the microbiome and makes the analysis of causality and dynamic interactions between bacteria impossible.

Here, we used a statistical framework in order to understand the dynamic interactions within the healthy human gut microbiome and to what extent we are able to predict it.

For our analysis we used publicly available datasets containing time series of the gut microbiome collected from four healthy individuals. To overcome limitations of microbiome time series data, i.e. compositionality, sparsity and missing values, we performed necessary transformations. Transformed data was used as input for linear regression based models and their performance was evaluated using statistical as well as ecological metrics.

In this project we answer questions such as: how bacteria interact with each other in time; are interactions between bacteria unique for a host or similar between subjects; to what extent gut microbiome is self-explainable; are interactions within a phylogenetic/taxonomic group stronger; which bacteria are the main drivers of the gut microbiome community?

We believe that our results and subsequent work will help us to better understand the dynamic interactions within the human gut microbiome. In future, they can be used in microbiome-based personalised therapy where we can plan in advance how to model microbiome changes and anticipate diseases originating from dysbiosis, or to manipulate microbiome composition.

Genomes

Multi-scale phase separation by explosive percolation with single chromatin loop resolution

Kaustav Sengupta (Center of New Technologies, University of Warsaw, Warsaw, Poland), Michał Denkiewicz (Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland), Mateusz Chiliński (Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland), Sevastianos Korsak (Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland), Teresa Szczepinska (Centre for Advanced Materials and Technologies, Warsaw Technical University, Warsaw, Poland), Ayatullah Faruk Mollah (Department of Computer Science and Engineering, Aliah University, Kolkata, West Bengal, India), Raissa D'Souza (Department of Computer Science, University of California, Davis, USA), Yijun Ruan (The Jackson Laboratory for Genomic Medicine, USA) and Dariusz Plewczynski (Centre of New Technologies, University of Warsaw).

Abstract:

The 2m-long human DNA is tightly intertwined into the cell nucleus of the size of 10µm. The DNA packing is explained by folding of chromatin fiber. This folding leads to the formation of such hierarchical structures as: chromosomal territories, compartments; densely packed genomic regions known as Chromatin Contact Domains (CCDs), and loops. We propose models of dynamical genome folding into hierarchical components in human lymphoblastoid, stem cell, and fibroblast cell lines. Our models are based on explosive percolation theory. The chromosomes are modeled as graphs where CTCF chromatin loops are represented as edges. The folding trajectory is simulated by gradually introducing loops to the graph following various edge addition strategies that are based on topological network properties, chromatin loop frequencies, compartmentalization, or epigenomic features. Finally, we propose the genome folding model - a biophysical pseudo-time process guided by a single scalar order parameter. The parameter is calculated by Linear Discriminant Analysis. We simulate the loop formation by using Loop Extrusion Model (LEM) while adding them to the system. The chromatin phase separation, where fiber folds into topological domains and compartments, is observed when the critical number of contacts is reached. We also observe that 80% of the loops are needed for chromatin fiber to condense in 3D space, and this is constant through various cell lines. Overall, our in- silico model integrates the high-throughput 3D genome interaction experimental data with the novel theoretical concept of phase separation, which allows us to model event-based time dynamics of chromatin loop formation and folding trajectories.

Genomes

Mutation Clonality, Neoantigens, and Immune Biomarkers relate to Immunotherapy Response in Bladder Cancer patients

Lilian Marie Boll (Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM)), Júlia Perera-Bel (MARGenomics and Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM)), Oriol Arpí (Medical Oncology Department, Hospital del Mar), Ana Rovira (Medical Oncology Department, Hospital del Mar), Núria Juanpere Rodero (Department of Pathology, PSMAR-IMIM Research Institute, Barcelona), Silvia Hernández-Llodrà (Department of Health and Experimental Sciences, Universitat Pompeu Fabra (UPF)), Josep Lloreta (Department of Pathology, PSMAR-IMIM Research Institute, Barcelona), Alejo Rodriguez-Vida (Hospital del Mar Medical Research Institute (IMIM)), M. Mar Albà (Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM)) and Joaquim Bellmunt (Beth Israel Deaconess Medical Center, PSMAR-IMIM Lab, Harvard Medical School, Boston, MA, USA).

Abstract:

The approval of checkpoint inhibitors (CPIs) states a major advance in immunotherapy. Yet, a majority of cancer patients does not profit from CPI treatment. Biomarkers suggested as potential predictors fail in most cases and evidence on immunogenic neoantigens remains scarce. We analysed an extensive list of biomarkers and related them to CPI treatment response in 29 advanced bladder cancer patients using raw tumor and germline WES as well as RNA-Seq data. We present evidence of a significantly higher tumor mutational burden in responders than in non-responders. Clonal tumor mutational burden increased the difference between treatment groups. In addition, we observed the mutational signature APOBEC in clonal mutations to be higher in responders. In regards to neoantigen quality, complete responders were found to have stronger neoantigen presentation than non-complete responders. In summary, results from this study contribute to a better understanding of the response to immunotherapy and help decipher the complex role of biomarkers in predicting CPI treatment response for bladder cancer patients.

Genomes

Mutational signatures are markers of drug sensitivity of cancer cells

Fran Supek (Institute for Research in Biomedicine (IRB Barcelona)), Jurica Levatić (Institute Jožef Stefan), Marina Salvadores (Institute for Research in Biomedicine (IRB Barcelona)) and Francisco Fuster-Tormo (Josep Carreras Institute).

Abstract:

Genomic analyses have revealed mutational footprints associated with DNA maintenance gone awry, or with mutagen exposures. Because cancer therapeutics often target DNA synthesis or repair, we asked if mutational signatures make useful markers of drug sensitivity. We detect mutational signatures in cancer cell line exomes -- where matched healthy tissues are not available -- by adjusting for the confounding germline mutation spectra across ancestries. We identify robust associations between various mutational signatures and drug activity across cancer cell lines; these are as numerous as associations with established genetic markers such as driver gene alterations. Signatures of prior exposures to DNA damaging agents – including chemotherapy – tend to associate with drug resistance, while signatures of deficiencies in DNA repair tend to predict sensitivity towards particular therapeutics. Replication analyses across independent drug and CRISPR genetic screening data sets reveal hundreds of robust associations, which are provided as a resource for drug repurposing guided by mutational signature markers. [this study was recently published as Levatic et al. (2022) Nature Communications]

Genomes

NETWORK ANALYSIS OF MULTI-OMICS MICROBIOME DATA TO IDENTIFY KEYSTONE UNKNOWN TAXA

Rocío Amorín de Hegedüs (University of Florida), Ana Conesa (Spanish National Research Council) and Jamie Foster (University of Florida).

Abstract:

Microorganisms have shaped Earth's biochemical and physical landscapes by inhabiting diverse metabolic niches. However, most microbial species remain unknown and are referred to as "microbial dark matter", highlighting a gap in our understanding of structured complex ecosystems. There are different omics methodologies to study microbial communities and combining these omics can provide a better look into the distinct biological layers in the ecosystem. However, there are no standard methodologies for meta-omics integration of microbiome data. We evaluate the role of 'microbial dark matter' in structured communities, using microbialites as a model ecosystem, and different types of data: metagenomics, metatranscriptomics and amplicon sequencing. Metagenomics (n=56) and metatranscriptomics (n=34) samples and were input into SortMeRNA to extract 16S reads. The output, as well as amplicon sequencing samples (n=14), were processed through QIIME2 for taxonomy analysis. Afterwards, R package MDMnets was utilized to build co-occurrence networks. Most hubs presented unknown classifications, sometimes up to the phyla level. A lack of classification at higher taxonomy levels suggests unknown microbes represent poorly characterized microbial lines. Future comparison of the highest scoring hubs of each data type using sequence similarity networks can yield the most similar important hubs occurring across the different data types, allowing the identification of the most relevant hubs. This work highlights the importance of unknown taxa in community structure and proposes microbiome data-integration and ecosystem network construction to identify significant unknown taxa for further characterization.

Genomes

NGSEP 4: EFFICIENT AND ACCURATE IDENTIFICATION OF ORTHOGROUPS AND WHOLE-GENOME ALIGNMENT

Laura Natalia González García (Universidad de los Andes), Daniel Tello (Universidad de los Andes), Juan Camilo Zuluaga-Monares (Universidad de los Andes), Ricardo Angel (Universidad de los Andes), Daniel Mahecha (Universidad de los Andes), Nicolas Cardozo (Universidad de los Andes), Camilo Escobar (Universidad de los Andes), Jorge Gomez (Universidad de los Andes), Mario Linares-Vasquez (Universidad de los Andes) and Jorge Duitama (Universidad de los Andes).

Abstract:

Whole-genome alignment is a basic task to understand genome evolution and diversity through comparative genomics. Approaches based on direct pairwise comparisons of DNA sequences require large computational capacities. As a consequence, pipelines combining tools for orthologous gene identification and synteny have been developed. Here we present an efficient and accurate algorithm for identification of orthogroups, combined with new implementations of the HalSynteny and DAGchainer algorithms to align and compare complete genomes. These developments are released as new functionalities of NGSEP 4. Our results analyzing the Orthobench dataset show that the NGSEP algorithm for ortholog identification has competitive accuracy and better computational efficiency compared to Orthofinder, being at least 3x faster than Orthofinder. Furthermore, the accuracy of NGSEP is superior if only the subset of mammalian genomes is taken into account. This suggests that the clustering of the NGSEP algorithm is more robust to changes in the complexity of the analyzed phylogeny and it. Additionally, the implementation includes a reconstruction of pangenomic datasets based on frequencies of the orthogroups among the genomes. Pangenome results over subsets of 10 to 100 bacterial genomes show that the soft core-genome is stable when increasing the number of genomes. Finally, we implemented an interactive visualization of the whole genome alignment based on synteny of the orthogroups. We expect that these new developments will be very useful for several studies in evolutionary biology and population genomics.

Genomes

OMArk: Quality assessment of protein-coding gene repertoires

Yannis Nevers (University of Lausanne), Victor Rossier (University of Lausanne) and Christophe Dessimoz (University of Lausanne).

Abstract:

Assessing the quality of protein-coding gene repertoires inferred from genome annotations has become critical in an era of increasingly abundant genome sequences for a widening diversity of species . State-of-the-art quality assessment tools like BUSCO can be used to measure the completeness of a gene repertoire - using a limited set of conserved genes - but they are blind to other types of errors in genome annotation: the over-prediction of protein-coding genes from non-coding or contaminant genomic regions.

To overcome these limitations, we present OMArk, a software relying on fast alignment-free sequence comparisons with precomputed evolutionary relationships to quickly assess the quality of a proteome (hereby, gene repertoire represented by one protein sequence per gene). OMArk estimates the completeness and consistency of the proteome by performing comparisons to the expected ancestral gene content of the corresponding species' lineage (gene families with at least one representative in extant species of this lineage). Completeness is measured as the proportion of the conserved gene families that are found in the proteome in one or multiple copies and consistency as the proportion of genes with clear homologs in gene families known to exist in the target lineage. Finally, OMArk evaluates contamination based on the taxonomic distribution of gene families with homologs in the proteome.

We validate OMArk with simulated data, then perform a global analysis of a publicly available dataset of 1805 eukaryotic proteomes—identifying examples of data quality issues.

Genomes

Omics approaches and artificial intelligence strategies applied to industrial Saccharomyces cerevisiae yeasts for corn and second-generation ethanol production processes

Marcelo Falsarella Carazzolle (University of Campinas), Beatriz Vargas (University of Campinas), Larissa Escalfi Tristao (University of Campinas), Thais Oliveira Secches (University of Campinas), Jade Ribeiro dos Santos (University of Campinas), Juliana Jose (University of Campinas), Fellipe da Silveira Bezerra Mello (University of Campinas) and Gonçalo Amarante Guimarães Pereira (University of Campinas).

Abstract:

Ethanol is one of the most important biofuels to mitigate climate change in a short period. Brazil is one of the pioneers in ethanol production (since 1970) and is currently the second-largest producer with 31% of world production. Brazilian production is based on sugarcane juice (or molasse), but corn ethanol has been growing rapidly. In addition, second-generation (2G) ethanol emerges as a biofuel with great potential in the world, as it does not compete with food and used biomass residues as feedstock. 2G technology is still under development in the world, and in Brazil, there are already two industrial plants operating and two under construction. The development of high-performance fermentative yeasts is a challenge to increase the productivity and yield, mainly focusing on the corn- and 2G-ethanol processes, in which genetically modified yeasts are often used. This work used the combination of multi-omic approaches and artificial intelligence strategies applied to industrial yeasts Saccharomyces cerevisiae to develop genetically modified yeasts with great potential for application in corn and 2G ethanol industries. The final strains were successfully evaluated in standardized bench-scale fermentation to mimic the industrial processes. For the corn ethanol-producing yeast, the focus was 1) prospection of ethanol tolerant genes using QTL (Quantitative Trait Locus) analysis and 2) in-silico prospection of secreted alpha- and gluco-amylases that have activity at fermentation temperature (32 C) and are patent-free. For the 2G yeast, the focus was on xylose consumption through a combination of metagenomics data and machine learning analysis.

Genomes

PerSVade: Personalized Structural Variation detection in your species of interest

Miquel Àngel Schikora Tamarit (Barcelona Supercomputing Centre; Institute for Research in Biomedicine (IRB Barcelona)) and Toni Gabaldón (Barcelona Supercomputing Centre; Institute for Research in Biomedicine (IRB Barcelona); ICREA).

Abstract:

Structural variants (SVs) like translocations, deletions, and other rearrangements underlie genetic and phenotypic variation. SVs are often overlooked due to difficult detection from short-read sequencing. Most algorithms yield low recall on humans, but the performance in other organisms is unclear. Similarly, despite remarkable differences across species' genomes, most approaches use parameters optimized for humans. To overcome this and enable species-tailored approaches, we developed perSVade (personalized Structural Variation Detection), a pipeline that identifies SVs in a way that is optimized for any input sample. Starting from short reads, perSVade uses simulations on the reference genome to choose the best SV calling parameters. The output includes the optimally-called SVs and the accuracy, useful to assess the confidence in the results. In addition, perSVade can call small variants and copy-number variations. In summary, perSVade automatically identifies several types of genomic variation from short reads using sample-optimized parameters. We validated that perSVade increases the SV calling accuracy on simulated variants for six diverse eukaryotes, and on datasets of validated human variants. Importantly, we found no universal set of "optimal" parameters, which underscores the need for species-specific parameter optimization. PerSVade will improve our understanding about the role of SVs in non-human organisms.

Genomes

PhyloCloud: an online platform for making sense of phylogenomic data

Ziqi Deng (Centro de Biotecnología y Genómica de Plantas, Madrid), Jorge Botas (Centro de Biotecnología y Genómica de Plantas, Madrid), Carlos P Cantalapiedra (Centro de Biotecnología y Genómica de Plantas, Madrid), Ana Hernández-Plaza (Centro de Biotecnología y Genómica de Plantas, Madrid), Jordi Burguet-Castell (Centro de Biotecnología y Genómica de Plantas, Madrid) and Jaime Huerta-Cepas (Centro de Biotecnología y Genómica de Plantas, Madrid).

Abstract:

Rapid growth of genome data over the last decades generates large amounts of phylogenetic trees and multiple sequence alignments, which may be enormous in size, providing opportunities to the evolutionary history of species. However, the analysis and interpretation of such data still rely on custom bioinformatic and visualisation workflows that are not entirely user-friendly for researchers without prior programming background. Besides, there is increasing demand for fast exploration of large trees and comprehensively managing a large number of phylogenomics data.

Here we present PhyloCloud, an online platform aimed provide a one-stop solution of hosting, managing and exploring large phylogenetic tree collections, providing also various options of analysis and operations, such as taxonomic annotation, searching, topology editing, automatic tree rooting, orthology detection, evolutionary events detection, etc. Besides, PhyloCloud provides a handful of phylogenetic tools such as allowing users to reconstruct their own phylogenies using predefined workflows, graphically compare tree topologies, and query taxonomic databases such as NBCI or GTDB. It is worth mentioning that PhyloCloud utilized a novel tree visualisation system empowered by ETE Toolkit v4.0, which can be used to explore very large trees up to one million tree nodes and enhance them with custom annotations and multiple sequence alignments. The platform allows for sharing tree collections and specific tree views via private links, or make them fully public, serving also as a repository of phylogenomic data. PhyloCloud is available at https://phylocloud.cgmlab.org

Genomes

PhylomeDB v5: An updated site to browse and mine genome-wide catalogs of gene phylogenies

Diego Fuentes Palacios (Barcelona Supercomputing Center BSC & IRB Barcelona - Institute for Research in Biomedicine), Manu Molina (Barcelona Supercomputing Center BSC & IRB Barcelona - Institute for Research in Biomedicine), Uciel Chorostecki (Barcelona Supercomputing Center BSC & IRB Barcelona - Institute for Research in Biomedicine), Salvador Capella-Gutiérrez (Barcelona Supercomputing Center BSC), Marina Marcet-Houben (Barcelona Supercomputing Center BSC & IRB Barcelona - Institute for Research in Biomedicine) and Toni Gabaldón (Barcelona Supercomputing Center BSC & IRB Barcelona - Institute for Research in Biomedicine).

Abstract:

Gene phylogenies represent the evolutionary relationships across genes in different species. These phylogenetic trees are commonly used to aid in the inference of homology relationships (i.e. orthology and paralogy) as well as of evolutionary relevant events such as family expansions, recombination and horizontal gene transfer. The plurality of evolutionary histories of genes encoded by an organism's genome is best represented by a genome-wide collection of phylogenetic trees (i.e. phylome).

Phylomes are defined as complete sets of evolutionary histories of the genes encoded in a given organism. They provide at the evolutionary history of the organism through the perspective of all of its individual gene histories. Phylomes are powerful tools to study genome dynamics and how they relate to important phenotypic innovations. They provide at the evolutionary history of the organism through the perspective of all of its individual gene histories.

PhylomeDB is a free, accessible and comprehensive web-server of genome-wide collections of gene phylogenies. First described in 2006, it has been regularly updated and expanded over the years. It is the largest public database for pre-computed gene family trees, currently storing more than 6 million phylogenetic trees and alignments including genes for up to 803 species. Trees are built using a sophisticated automated pipeline that includes, homology searches, alignment reconstruction and trimming, evolutionary model selection and maximum-likelihood inference. PhylomeDB allows the visualization of phylogenetic trees, multiple alignments and orthology and paralogy predictions.

Genomes

Predicting Virulence of Listeria Monocytogenes using Whole Genome Sequencing and Machine Learning

Alexander Gmeiner (Research Group for Genomic Epidemiology, Technical University of Denmark), Patrick Murigu Kamau Njage (Research Group for Food Microbiology and Hygiene, Technical University of Denmark), Lisbeth Truelstrup Hansen (Research Group for Food Microbiology and Hygiene, Technical University of Denmark) and Pimlapas Leekitcharoenphon (Research Group for Genomic Epidemiology, Technical University of Denmark).

Abstract:

Listeria monocytogenes (LM) is a concerning food-borne pathogen that poses a substantial threat to public health. Hence, many countries have rigorous regulations for LM in food products. Currently, these regulations do not consider the heterogeneity of LM virulence, even though research shows major differences in LM virulence on a sub-species level. Whole Genome Sequencing (WGS) has widely become the standard for pathogen surveillance, and thorough screening networks have been implemented which combine clinical and food industry data. In this study, we aim to harness these networks and unravel LM virulence further. To do this, we combine Machine Learning (ML) techniques with WGS data to predict LM virulence on a sub-species level.

The data used in this study was obtained from two exhaustive surveillance systems of LM conducted by Danish and French authorities and consists of 169 data points. As an estimate for virulence, we used the clinical frequency (number of clinical isolates/ (number of clinical isolates + number of food isolates)). This study compares the two-layer cross-validation performance of three different genomic levels (i.e., virulence genes, pangenome genes, kmers). The preliminary results suggest that a broader genomic level (i.e., pangenome) yields better predictive performances (F1-score: 0.88; 95%-CI: 0.87, 0.91).

In conclusion, the results suggest that exhaustive WGS surveillance data can be used to predict LM virulence on a sub-species level. As the sparsity of well-suited data limits our findings, we are also exploring the possibility of using other publicly available datasets for our methodology.

Genomes

Prevalence and Specificity of Chemoreceptor Profiles in Plant-Associated Bacteria

Claudia Sanchis-López (Universidad Politécnica de Madrid), Jean Paul Cerna-Vargas (Centro de Biotecnología y Genómica de Plantas (CBGP). Universidad Politécnica de Madrid (UPM)), Saray Santamaría-Hernando (Centro de Biotecnología y Genómica de Plantas (CBGP). Universidad Politécnica de Madrid (UPM)), Cayo Ramos (Universidad de Málaga-Consejo Superior de Investigaciones Científicas (IHSM-UMA-CSIC)), Tino Krell (Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas), Pablo Rodríguez-Palenzuela (Universidad Politécnica de Madrid (UPM)), Emilia López-Solanilla (Universidad Politécnica de Madrid (UPM)), Jaime Huerta (CBGP (UPM-INIA)) and José J. Rodríguez-Herva (Centro de Biotecnología y Genómica de Plantas (CBGP). Universidad Politécnica de Madrid (UPM)).

Abstract:

Chemoreceptors (CRs) are proteins able to detect environmental changes by binding specific signals to their ligand binding domain (LBD), initiating a signal transduction cascade that controls a variety of cell processes. However, although CRs play a central role for microbiome interactions, little is known about their phylogenetic and ecological specificity. Considering the enormous variety of LBDs at sensor proteins, an important question resides in establishing the forces that have driven their evolution and selection.

Here, I will present our recent work on the systematic analysis and classification of plant-related CRs, covering more than 82,000 chemosensing sequences extracted from 11,806 representative microbial species. To do so, we classified CRs according to their LBD type using a de novo homology clustering method. Through phylogenomic analysis, we identified hundreds of LBDs that are present predominantly in plant-associated bacteria, finding that the taxonomic distribution of these LBD types is only partially explained by phylogeny. Our results show that the profile of LBD types in a given genome is related to the lifestyle specialization, with plant symbionts and phytopathogens showing the highest number of niche-specific LBDs.

These findings offer a number of research opportunities in the field of signal transduction, such as the exploration of similar relationships in CRs of bacteria with a different lifestyle (e.g. human intestine). In this talk, I will also present our follow-up work on the topic, where we extended our analyses to global metagenomics data, covering CR and LBD distributions across docens of environments and different hosts.

Genomes

Proton and alpha radiation-induced mutational profiles in human cells.

Tiffany Delhomme (IRB Barcelona), Manuela Buonanno (Radiological Research Accelerator Facility (RARAF), Columbia University), Grilj Veljko (Radiological Research Accelerator Facility (RARAF), Columbia University), Josep Biayna (IRB Barcelona) and Fran Supek (IRB Barcelona).

Abstract:

Ionizing radiation (IR) is known to be DNA damaging and mutagenic, however less is known about which mutational footprints result from exposures of human cells to different types of IR. We were interested in the mutagenic effects of galactic cosmic radiation (GCR) exposure on genomes of various human cell types, in order to gauge the genotoxic risks of space travel, and of certain types of tumor radiotherapy. To this end, we exposed cultured cell lines from the blood, breast and lung to intermittent proton and alpha (helium nuclei) rays at doses sufficient to affect cell survival. Whole-genome sequencing revealed that mutation rates were not overall markedly increased upon GCR exposures. However there were changes in mutation spectra and distributions, such as the increases in clustered mutations and of certain types of indels and structural variants. The spectrum of mutagenic effects of GCR exposures may often be cell-type and/or genetic background specific. Overall, the mutational effects of recurrent exposures to proton and alpha radiation on human cells appear subtle, however further work is warranted to understand effects of chronic, long-term exposures on various human tissues.

Genomes

Quantifying microbial dark matter and its impact on metagenomic analyses

Elizabeth Yuu (Hasso-Plattner-Institut), Vitor Piro (Hasso-Plattner-Institut) and Bernhard Renard (Hasso_Plattner-Institut).

Abstract:

Alterations in the microbiome are known to cause severe health problems which can lead to infectious and chronic diseases. Analyzing the microbiome provides a better understanding of what depicts healthy versus unhealthy microbial compositions. For example, closely related bacterial species that have similar genomes can present large phenotypic differences. We therefore focus on how similar species differ and how these deviations affect the microbiome. In metagenomics, mapping reads to reference genomes allows for insights into taxonomic compositions and variations between microbial communities. We previously introduced DiTASiC (Differential Taxa Abundance including Similarity Correction) for shared read ambiguity resolution based on a regularized, generalized linear model (GLM) framework. This, and similar approaches, does not address the remaining unmapped reads, or “microbial dark matter”. We extend DiTASiC by introducing a LASSO GLM filtering step and a new classification tool, Ganon. When supplying more reference genomes than actual genomes in the sample, DiTASiC produces negative estimates. The filtering step corrects this by removing the unnecessary reference genomes, thus only focussing on the genomes that are most likely in the sample. Ganon indexes using interleaved bloom filters and applies a lowest common ancestor step which provides more accurate k-mer counts and taxa abundance estimates specifically when the sample contains unknown genomes. These alterations have already presented promising results for addressing excessive reference genomes and improved taxa quantification. From six simulated datasets, our filtering step successfully removed all excessive reference genomes, while Ganon outperformed our previous method under the scenario when the sample had unknown genomes.

Genomes

Read-level GC bias correction for improved cell-free DNA signal processing

Sebastian Röner (Berlin Institute of Health (BIH) at Charité – Universitätsmedizin Berlin), Benjamin Spiegl (Institute of Human Genetics, Diagnostic and Research Center for Molecular BioMedicine, Medical University of Graz), Michael R. Speicher (Institute of Human Genetics, Diagnostic and Research Center for Molecular BioMedicine, Medical University of Graz) and Martin Kircher (Berlin Institute of Health (BIH) at Charité – Universitätsmedizin Berlin).

Abstract:

Cell-free DNA (cfDNA) is found in many bodily fluids and is believed to derive primarily from apoptosis of hematopoietic cells. In the context of certain physiological conditions or disease processes, the proportion of tissues contributing to cfDNA changes. These observations led to an increased research interest in cfDNA and liquid biopsies.

In recent years, many approaches were developed to extract information from cfDNA samples for disease monitoring or early diagnostics. Methods range from identification of allelic differences at known disease markers, coverage alterations indicative of structural changes in tumor tissues, fragmentation differences to measuring methylation state. These methods measure different signals, but all rely on the quantification of read distributions and slight changes in read recovery affect their results. A common bias in sequencing and DNA library preparation is related to the Guanine+Cytosine (GC) content of fragments, leading to under- or overrepresentation of certain read populations. To address this, methods were proposed, but none specifically addresses read-level correction for cfDNA.

Here, we describe an easy-to-use tool to determine and correct GC biases in cfDNA samples. Based on fragment length and GC distributions, we calculate read-level correction values and make them available as tags in SAM/BAM format. We can use this additional information during signal extraction in various metrics (e.g., allelic balance, coverage, read ends/midpoints, WPS), while preserving the original read coverage patterns for specific analyses. Our tool is available on GitHub (https://github.com/kircherlab/cfDNA_GCcorrection.git), implemented in Python and uses multiprocessing to efficiently process large sequencing datasets.

Genomes

Recurrent positive selection of lipid trafficking genes in Clupeiformes

Jorge Langa (University of the Basque Country), Yuri Rueda (University of the Basque Country UPV/EHU), Aitor Albaina (University of the Basque Country UPV/EHU), Martin Huret (DECOD (Ecosystem Dynamics and Sustainability), IFREMER), Darrell Conklin (University of the Basque Country) and Andone Estonba (University of the Basque Country UPV/EHU).

Abstract:

Clupeiformes is one of the most important orders of fishes due to its ecological and economic importance: it comprises emblematic species such as the European anchovy, the Atlantic herring and the European anchovy. The plasticity that their genomes show explains the ubiquity of these species across the globe: they occupy tropical and temperate latitudes, they are typically marine but adaptable to freshwater. However, little is known about their concrete genetic makeover and evolutive strategies, or how these species have become one of the richest sources of omega-3 long-chain polyunsaturated fatty acids (ω-3 LC-PUFAs). Here we report the discovery of genes and families under positive selection in the Clupeiformes order. Through RNA-Seq, we assembled the transcriptome of 12 clupeids, predicted protein coding sequences, clustered them into genes, and analyzed them to detect signals of positive selection. In general, we found positively selected portions of the genome related to mitochondria, ribosomes, lysosomes, caveolae, extracellular proteins, and CD molecules. Furthermore, we observe positively selected genes associated mainly with apolipoproteins and caveolae, among others, indicating that these fishes have adapted their molecular machinery to efficiently store and transport fats across tissues. The herein applied methodology and the obtained results pave the way for further research into the evolutionary history of Clupeiformes, while also streamlining the study of another set of species, from raw RNA-Seq reads to tabular results.

Genomes

Redefining Promoter DNA Methylation Change in Cancer

Richard Heery (European Institute of Oncology) and Martin Schaefer (European Institute of Oncology).

Abstract:

Promoter DNA methylation has been recognized for decades as one of the major mechanisms of epigenetic regulation of gene expression and has consistently been found to be perturbed early during tumorigenesis.

While promoter methylation has generally been considered to lead to downregulation of transcription, we and others have observed that DNA methylation proximal to the TSS is often associated with increased transcription.

Moreover, there remains little consensus on the actual genomic extent of promoters relative to the location of the TSS. In particular, it is poorly understood how far upstream promoters extend from the TSS, whether or not to also consider the sequence immediately downstream of the TSS as part of the promoter and it is unknown if the size of promoters varies from gene to gene. Additionally, the relative importance of methylation of different parts of the promoter to transcriptional activity of the associated TSS has yet to be thoroughly explored.

We are applying machine learning approaches to a large dataset of human healthy prostate samples with RNA-seq and whole genome bisulfite sequencing (WGBS) data to create for the first time a detailed map of the CpG sites most important to transcriptional activity.

We previously developed a computational tool MethylDriver, which identifies regions under selection for methylation changes in cancer. Using a matching set of prostate cancer samples, we will use this revised knowledge of promoter methylation along with MethylDriver to detect the subregions of promoters most relevant to transcriptional activity that may be under selection in cancer

Genomes

Relation between genome mutability, variant pathogenicity and vertical ionization potential of nucleobase motifs

Cyril Karamaoun (Université libre de Bruxelles), Pauline Hermans (Université libre de Bruxelles), Fabrizio Pucci (Université libre de Bruxelles) and Marianne Rooman (Université libre de Bruxelles).

Abstract:

Single-base substitutions (SBS) in genomes are known to be non-random and influenced by the type of nucleobase and flanking sequence. They are potentially responsible for a range of diseases such as cancer and various neurodevelopmental disorders. The biophysical mechanisms that are at the basis of mutagenesis are still unclear but have been related to the electron hole transport along the DNA base stacks. Indeed, cell exposure to high-energy radiations or to reactive oxygen species, for example, can lead to DNA ionization through the creation of electron holes. These holes usually migrate along the DNA stack until they remain localized in regions of low vertical ionization potential (vIP). To advance this issue, we estimated the vIP of every possible sequence of 1 to 4 successive nucleobases using MP2/6-31G* ab initio quantum chemistry calculations, and analyzed in parallel several SBS datasets that contain (germline or somatic) variants observed in rare diseases, cancer, normal tissues, and centenarians. We computed the correlation between the normalized frequency of SBSs in the different datasets and the vIP values of the flanking sequence motifs. We found a statistically significant overall anticorrelation between these two quantities: the lower the vIP value, the more probable the substitution. The value of the anticorrelation coefficient is shown to depend on the genome regions (introns, exons, intergenic), variant type (missense, synonymous), and the pathogenic effect. Interestingly, it is significantly better for somatic cancer variants and worse for centenarians.

Genomes

Resolution of deep nodes and new solid backbone phylogeny in Ophioglossaceae ferns

Darina Koubínová (University of Neuchâtel), Li-Yaung Kuo (National Tsing Hua University) and Jason Grant (University of Neuchâtel).

Abstract:

The Ophioglossaceae are an ancient, eusporangiate fern family distributed worldwide in temperate and tropical regions. Many of the species are widespread, others are locally endemic. There are currently more than 100 species, about 12-14 genera and 4 subfamilies recognized. Some of the main lineages are speciose, but some are monotypic, such as the subfamilies Helminthostachyoideae and Mankyuoideae. Former attempts to infer phylogenetic tree structures of the Ophioglossaceae included only short plastid regions and some of the used datasets even contained a great portion of missing data. In two recent phylogenomic analyses, whole plastome sequences were used but scarce representatives were sampled. Furthermore, rather simplified substitution models were applied.

To resolve the deepest nodes and obtain a solid backbone of Ophioglossaceae, we adopted a phylogenomic approach using the majority of the currently recognized genera and analyzed datasets from not only the plastome but also the mitogenome. We used genome skimming data to assemble these organelle genomes and from the resulting assemblies, we extracted the coding sequences (CDS) for the phylogenomic inferences. We tested different partition and substitution models, including finer ones in order to better account for rate heterogeneity among loci and codon positions. Our phylogenomic results overall supported a novel, previously uncovered topology which presented the most solid infra-family backbone for Ophioglossaceae. Finally, based on this infra-family backbone, we traced phylogenetic origins of the hypothesized horizontal gene transfer (HGT) in organellar genomes, ancient whole-genome duplication (WGD) events, and key morphological innovations in Ophioglossaceae.

Genomes

Sample demultiplexing and doublets removal from single-nuclei RNA sequencing data

Kwong Leong Wong (German cancer research center), Lena Jassowicz (German cancer research center), Peter Lichter (German cancer research center), Christel Herold-Mende (University of Heidelberg), Martina Seiffert (German cancer research center) and Marc Zapatka (German cancer research center).

Abstract:

Single-nuclei sequencing is gaining popularity in biomedical research by its capability in identifying tissue heterogeneity. The need of scaling up cohort sizes for statistical power is increasing. The capability of 10X genomics platforms to sequence thousands of cells in a single run opened an avenue for sample multiplexing in each library.

Here we describe a pipeline using vireo, a demultiplexing platform using single nucleotide polymorphisms (SNPs) to inform sample identity on cell level. The pipeline was applied to a cohort of 28 breast cancer brain metastasis specimens. To match tumor identities with demultiplexed sample IDs, Illumina OncoArrays were performed on 24 of the multiplexed specimens.

Our cohort of breast cancer brain metastasis is aiming to study the immune and stromal compartment to explore markers and interactions that could improve immunotherapeutic outcomes. Tissues were selected to best represent all the cell populations in the tumor immune microenvironment. During experiment we analyzed 28 tumors demultiplexed using 12 single-nuclei libraries. Our experience suggests a hybrid-mode utilizing both SNP arrays and SNPs from RNA-Sequencing improves assigning cell-identities. Finally we developed a two-step approach using vireo to recover initially unassigned cells.

We also identified doublets defined as cell-barcodes that can be linked to more than one individual based on the identified SNPs. Doublets were unexpectedly being most abundant in the low-read counts cell population instead of high-read counts cell populations. The doublet cell-populations present with a gene-signature enriched for ribosomal genes and some mitochondrial genes. The signature was applied to non-multiplexed pilot runs to remove potential doublets.

Genomes

scTAM-seq enables targeted high-confidence analysis of DNA methylation in single cells

Agostina Bianchi (Centre for Genomic Regulation), Michael Scherer (Centre for Genomic Regulation), Roser Zaurin (Centre for Genomic Regulation), Kimberly Quililan (Centre for Genomic Regulation), Lars Velten (Centre for Genomic Regulation) and Renee Beekman (Centre for Genomic Regulation).

Abstract:

Profiling DNA methylation at the single-cell level remains challenging due to excessive noise inherent in the assays, as well as limited cellular throughput.

To overcome these two issues, we developed a targeted bisulfite-free method termed scTAM-seq, which profiles up to 650 CpGs in up to 10,000 cells per experiment. ScTAM-seq has a dropout rate of less than seven percent, since sequencing coverage is focused on informative, variably methylated CpGs. Epigenomic consortia have profiled DNA methylation at the bulk level in many tissues and uncovered that cell-type markers make up a minor fraction of all CpGs. We developed an accompanying software suite that leverages bulk data to determine putatively informative CpGs for investigating cellular differentiation and states.

To analyze DNA methylation data generated by scTAM-seq, we extended existing software applications and implemented new analysis workflows. Subsequently, we applied scTAM-seq to B cells from blood and bone marrow and demonstrated that it can resolve DNA methylation dynamics across B-cell differentiation at unprecedented resolution. Within the population of non-switch memory B cells, we identified intermediate differentiation states that were previously masked in bulk DNA methylation data. Since scTAM-seq leverages the Mission Bio tapestri platform, it additionally queries surface protein expression, thus enabling integration of single-cell DNA methylation information with cell atlas data. Additionally, it can investigate somatic mutations thus allowing applications in tumor profiling. In summary, scTAM-seq is the first high-throughput, high-confidence method for analyzing DNA methylation at single-CpG resolution across thousands of single cells.

Genomes

Shared genetic architecture between tobacco smoking and iron concentration in the brain's dorsal striatum

Olga Trofimova (Dept. of Computational Biology, University of Lausanne) and Sven Bergmann (Dept. of Computational Biology, University of Lausanne).

Abstract:

Tobacco smoking is a major modifiable risk factor for cardiovascular and lung diseases. Better understanding its neurobiological underpinnings will benefit the prevention of smoking-related illnesses and mortality. Recent neuroimaging studies have identified a correlation between smoking and iron concentration in the dorsal striatum, a brain region involved in habit formation and compulsive behaviour, and a central node of dopamine activity.

Here we investigated the genetic correlation and possible causal relationships between smoking initiation (ever smoked regularly) and T2* – a magnetic resonance imaging marker of iron content – in the bilateral putamen and caudate nuclei. We performed linkage disequilibrium score regression and Mendelian randomization, using genome-wide association studies summary statistics. We found a genetic correlation between smoking and iron concentration in the four tested brain regions ($r \in [0.066,0.08]$, $p \in [0.002,0.033]$) but no evidence of causal relationship in either direction.

Our results suggest a common biological mechanism between tobacco smoking and iron concentration in the dorsal striatum, that could function as a reinforcing feedback loop rather than a one-way causal effect. This mechanism could involve the dopaminergic system, as it was previously shown that iron binds to dopamine D2 receptors in the dorsal striatum.

Genomes

Single-cell guided deconvolution of bulk AML transcriptomics recapitulates FAB landscape and CD14+ Monocyte percentage predicts Venetoclax resistance

Emin Onur Karakaslar (LUMC), Jeppe Severens (LUMC), Elena Sanchez Lopez (LUMC), Peter van Balen (LUMC), Heendrik Veelken (LUMC), Marcel Jt Reinders (TU Delft), Marieke Griffioen (LUMC) and Erik van den Akker (LUMC).

Abstract:

The diagnostics landscape for AML patients have shifted towards genetic abnormalities in the last ~10 years due to their prognostic insights, however treatment options are still lagging. Therefore, information regarding the maturational arrest has started to regain its popularity over its predictive power on drug use. Here, we deconvolute 1350 bulk RNA-seq samples from five independent AML cohorts via a single-cell healthy BM reference and demonstrate that FAB landscape (M0-M7) for these patients could be reconstituted using these immune compositions (ICs). Then using these ICs, we predict in-vitro drug resistances from BEAT-AML study, and we showcase Venetoclax (ABT-199), a BCL-2 inhibitor, has resistance specifically for patients with CD14+ Monocyte phenotype. Furthermore, using in-house proteomics data, we show that BCL-2 protein abundance is split into two distinct clusters for NPM1 patients at the extremes of CD14+ Monocyte percentages, which could be crucial for the Venetoclax dosage and administration for these patients. We also believe that these decisions might be extended for patients without genetic abnormalities (NOS), and possibly to MDS and sAML. Lastly, we also think that the proposed framework could be used as a blueprint for testing new drugs' resistances to distinct cell types in future. Our pipeline is accessible under GPL-3 license at https://github.com/eonurk/seAMLess.

Genomes

Single-cell resolution unravels spatial alterations in metabolism, transcriptome and epigenome of ageing liver

Chrysa Nikopoulou (Max Planck Research Group 'Chromatin and Ageing', Max Planck Institute for Biology of Ageing), Niklas Kleinenkuhnen (Max Planck Research Group 'Chromatin and Ageing', Max Planck Institute for Biology of Ageing), Swati Parekh (Max Planck Research Group 'Chromatin and Ageing', Max Planck Institute for Biology of Ageing), Tonantzi Sandoval (Max Planck Research Group 'Chromatin and Ageing', Max Planck Institute for Biology of Ageing), Farina Schneider (Institute for Pathology, University Hospital Cologne), Patrick Giavalisco (Metabolic Core Facility, Max Planck Institute for Biology of Ageing), Mihaela Bozukova (Max Planck Research Group 'Chromatin and Ageing', Max Planck Institute for Biology of Ageing), Anna Juliane Vesting (Max Planck Institute for Metabolism Research), Janine Altmüller (Cologne Center for Genomics, University of Cologne), Thomas Wunderlich (Max Planck Institute for Metabolism Research), Vangelis Kondylis (Institute for Pathology, University Hospital Cologne), Achim Tresch (Institute of Medical Statistics and Computational Biology, Faculty of Medicine, University of Cologne) and Peter Tessarz (Max Planck Research Group 'Chromatin and Ageing', Max Planck Institute for Biology of Ageing).

Abstract:

The liver lies in the center of metabolic regulation of the organism and is essential for metabolic homeostasis. Importantly, metabolic work in the tissue is spatially organized.

Nevertheless, ageing leads – among others – to metabolic rearrangements, most notably a substantial accumulation of large lipid droplets in the pericentral area of the tissue, which increases the risk for age-related liver diseases, such as NAFLD or NASH. We generated one of the first ageing liver cell chromatin atlas to assess liver ageing on a chromatin architecture level.

Further, we tried to answer whether hepatocytes age differently depending on the location within the tissue by (sc)spatial sequencing. We identify age-related chromatin architectural changes by scATAC-seq. that manifest as a zonation-specific expression of genes. These genes have an impact on lipid droplet size and composition, linking chromatin changes with phenotypic output. We could verify our findings by an analysis of the ageing liver lipidome.

We observed a decoupling from chromatin alterations on a gene expressional level that indicates a considerable role of post-transcriptional processes on the ageing liver phenotype.

Genomes

SonicParanoid enhanced by machine learning allows fast de novo orthology inference of huge MAG datasets

Salvatore Cosentino (Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Japan) and Wataru Iwasaki (Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Japan).

Abstract:

Accurate inference of orthologous genes constitutes a prerequisite for genomic and evolutionary studies. SonicParanoid is one of the fastest methods for orthology inference and comparably accurate to well-established methods despite being orders of magnitude faster. Nevertheless, its scalability is hampered by the lengthy all-vs-all alignments, and sequence-similarity search alone is not enough to predict very distant orthologs. In this work we try to tackle these two limitations using machine learning.

We substantially reduced the all-versus-all alignment execution time using an AdaBoost model which exploits the properties of the Bidirectional-Best-Hit and the factors affecting the computational time in local sequence alignment. Evaluation based on multiple datasets showed reductions in execution time up to 50% without negative effects on the accuracy of the orthology inference.

To address the second limitation we trained a doc2vec model with domain-architectures extracted from the input proteins, and we used it to infer orthologs based on domain-architecture similarities, resulting in an increase of one-third in the number of predicted orthologs.

Lastly, we evaluated the scalability using a dataset of 2000 MAGs. SonicParanoid was able to infer the orthologs in about 30 hours using 128 CPUs, where other well-established orthology inference tools were still running after a week.

The way we reduced all-vs-all execution time could be used by other graph-based methods, while the domain-based approach could in the future, thanks to its scalability, eliminate the need for all-vs-all alignments in orthology inference.

Documentation is available at: http://iwasakilab.bs.s.u-tokyo.ac.jp/sonicparanoid.

PyPI: https://pypi.org/project/sonicparanoid/

Genomes

Sounds trivial - but it's not! Automatic reference detection and web service independent reconstruction of influenza A and B genomes

Katja Winter (Robert Koch-Institut), Oliver Drechsel (Robert Koch-Institut), Marianne Wedde (Robert Koch-Institut), Ralf Dürrwald (Robert Koch-Institut), Thorsten Wolff (Robert Koch-Institut) and Stephan Fuchs (Robert Koch-Institut).

Abstract:

The Influenza virus caused multiple pandemics over the past centuries, such as in 1918 and in 2009 that have claimed millions of lives. Hence, the seasonal occurrence and spread of influenza is closely monitored in many countries under coordination of WHO in the Global Influenza Surveillance and Response System.

However, with continuous molecular surveillance and increase of sequenced virus genomes, the time required for in-depth analyses of the resulting data increases, necessitating scalable, reproducible and automated bioinformatics workflows.

Available online web services, such as INSaFLU provide a convenient way to perform bioinformatics analysis to assemble whole genome consensus sequences. However, these services are not suitable for many institutions because they do not comply with privacy regulations for sequencing data potentially containing human reads.

In this project, we present FluPipe, a Snakemake-based pipeline for influenza genome assembly based on Illumina data. The pipeline supports automated quality control and cleaning of raw data from contaminating e.g. human reads as well as reassortment-aware selection of reference sequences from a curated sequence database to deal with the genetic diversity of the virus. A BWA mapping is followed by precise variant calling and filtering using Lofreq. Special attention was paid when creating the consensus sequences in order to accurately identify ambiguous bases, deletions or insertions. Different quality and genomic metrics are clearly recorded and output in an intuitive report.

Thereby, FluPipe minimizes the manual input time required for successful influenza genome assembly and subsequent analysis to enable near real-time monitoring.

Genomes

Statistical and data mining analysis of stop codon triplets in introns in view of the stop-to-stop ORF definition

Valentin Wesp (Friedrich-Schiller-University Jena) and Stefan Schuster (Friedrich-Schiller-University Jena).

Abstract:

In the basic definition, ORFs are nucleotide sequences enclosed by a start and a stop codon with no other stop in between and whose lengths (in nt) are divisible by three. While this definition is sufficient as a first step for gene finding in prokaryotes, it usually fails in eukaryotes due to the presence of introns. An alternative ORF definition is often used, especially in gene finding software, while it has not yet attracted much attention from a theoretical point of view. It says that an ORF is delimited by two consecutive stop codons. By data mining in two plant and seven vertebrate genomes, we examine the relative frequency of stop codons in all three reading frames (for a given strand) for all introns to find whether ORFs are appropriately delimited by stop codons near splice sites. It is found that stops are particularly enriched at position 2 downstream of the 5' splice site, which corresponds to splicing phase 1. We compare the data mining results with the theoretically predicted frequency for random sequences with a given GC content. The calculations include the prediction of the average distance between stops and the minimum length so that at least one stop codon occurs with a 95 % probability. At 50 % GC content and with the standard genetic code, a length of 189 nt is obtained, which is much shorter than the median intron lengths of almost all organisms investigated here. Our results support the applicability of the alternative ORF definition.

Genomes

Strengths and weaknesses of metabarcoding long read.

Jean Mainguy (INRAE), Adrien Castinel (INRAE), Olivier Bouchez (INRAE), Sylvie Combes (INRAE), Carole Iampietro (INRAE), Christine Gaspin (INRAE), Denis Milan (INRAE), Cécile Donnadieu (INRAE), Claire Hoede (INRAE) and Geraldine Pascal (INRAE).

Abstract:

Metabarcoding is the large-scale taxonomic identification of complex environmental samples via analysis of DNA reads of one marker gene. Different marker genes are used, the 16S rRNA gene is mainly used to identify bacteria. We conducted a comparative study to understand the limitations and strengths of PacBio HiFi long read sequencing technology for metabarcoding analyses. We performed the same analyses on two types of datasets. A commercial ZymoBIOMICS mock community consisting of eight bacteria and 32 samples containing pig fecal microbiota from the ExpomycoPig project. We focused on the 16S-23S rRNA gene operon and sequenced 3 types of amplicons by metabarcoding: the V3-V4 region of the 16S rRNA, the full-length 16S rRNA gene and the full-length 16S-23S gene operon. We observe that by using HIFI long reads we increase the specificity of sample characterization. But when we analyze the 16S-23S amplicons, the specificity increases within the limit of the availability of the sequence in the databases that are much less comprehensive than the databases containing only 16S data. A strategy of affiliating 16S-23S amplicons on complementary databases (Silva 16S and 23S and a custom 16S-23S operon database) improves microbial identifications.

Acknowledgements

This work has been carried out within the framework of the SeqOccIn project (Sequencing Occitanie Innovation), supported by Get-PlaGe and Genotoul Bioinfo core facilities and financed by FEDER funds (Programme Opérationnel FEDER-FSE_Midi-Pyrénées et Garonne 2014-2020).

Genomes

Structural comparison of chromatin interaction networks generated from Hi-C data

Gatis Melkus (Institute of Mathematics and Computer Science, University of Latvia), Lelde Lace (Institute of Mathematics and Computer Science, University of Latvia), Peteris Rucevskis (Institute of Mathematics and Computer Science, University of Latvia), Sandra Silina (Institute of Mathematics and Computer Science, University of Latvia), Andrejs Sizovs (Institute of Mathematics and Computer Science, University of Latvia), Edgars Celms (Institute of Mathematics and Computer Science, University of Latvia) and Juris Viksna (Institute of Mathematics and Computer Science, University of Latvia).

Abstract:

Among the numerous mechanisms regulating gene expression in eukaryotes, one of the lesser understood systems is the chromatin architecture of the genome. The spatial organization of genes, cis-regulatory elements and other structural features of the genome has substantial implications on transcriptional regulation, and the processes underlying this organization are both complex and difficult to study. Hi-C is a chromatin conformation capture-based method that provides an incomplete set of chromatin contacts in a population of cells that serve as an indicator of genome-wide chromatin architecture.

In order to better understand the structure of Hi-C datasets and search for useful commonalities indicative of chromatin architecture, we have constructed networks out of several publicly available Hi-C and pcHi-C datasets for various human tissues and cell types. Comparing these networks, we find notable structural consistencies useful for a systematic approach to Hi-C analysis with graph-based methods, and test a variety of graph metrics for utility in distinguishing between tissue types and discovering specific genomic features.

We compare our network-based metrics with RNA-seq gene expression measurements, tissue similarity measures also calculated from RNA-seq data, chromatin activity states inferred from histone marks and other biological data. Our results show that Hi-C datasets and pcHi-C datasets, while variable in detail and structure, can be analyzed in an integrated fashion with a graph-based methodology, and we propose several metrics by which such an analysis can be conducted.

Genomes

Synggen: fast and data-driven generation of synthetic heterogeneous NGS cancer data

Riccardo Scandino (Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento), Federico Calabrese (Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento) and Alessandro Romanel (Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento).

Abstract:

Whole-exome (WES) and targeted (TS) next-generation sequencing (NGS) are widely utilized both in translational cancer genomics and in the setting of precision medicine. The benchmarking of computational methods and tools that are in continuous development is fundamental for the correct interpretation of somatic genomic profiling results. To this aim we developed synggen, a tool that enables fast and scalable generation of realistic and heterogeneous cancer sequencing synthetic WES and TS datasets, characterized by the incorporation of phased germline polymorphisms, complex allele-specific somatic copy number aberration and point mutations, together with clonality of somatic events and overall sample tumor content.

Synggen is written in C programming language and exploits a set of control (non-cancer) NGS sequencing files (BAM format) to generate reference models capturing a collection of data summary statistics. Reference models are used in conjunction with user-specified input lists of germline and somatic variants, together with clonality and tumor content information, to directly generate platform-specific cancer and matched control NGS files (FASTQ format).

The time required to generate WES reference models exploiting one control sample using 4 or 16 cores takes approximately 5 and 2.5 minutes, respectively. The generation of a FASTQ file with 100 million reads using the same number of cores requires about 10 and 4 minutes, respectively.

Overall, synggen allows to easily emulate varied and realistic cancer- and patient-specific data across different multi-subclones composition, tumor purity, aneuploidy and tumor evolution scenarios, enabling the simulation of large-scale synthetic cancer datasets for benchmarking studies.

Genomes

Taking the prediction of pathogenic variant-combinations to the next level with VarCoPP2.0

Nassim Versbraegen (Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles, Vrije Universiteit Brussel), Barbara Gravel (Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles, Vrije Universiteit Brussel), Charlotte Nachtegael (Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles, Vrije Universiteit Brussel), Alexandre Renaux (Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles, Vrije Universiteit Brussel), Emma Verkinderen (Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles, Vrije Universiteit Brussel), Ann Nowé (Artificial Intelligence Laboratory, Vrije Universiteit Brussel, 1050 Brussels, Belgium;), Tom Lenaerts (Universite Libre de Bruxelles) and Sofia Papadimitriou (Université Libre de Bruxelles).

Abstract:

Genetic sequencing enabled the development of new approaches to identify the genetic causes of rare diseases.

However, non-monogenic causes remain difficult to identify. A first step that addressed this was VarCoPP [1], An ensemble model of 500 Random Forests that was trained on disease-causing variant combinations that were curated from literature and gathered in DIDA [2] and neutral combinations stemming from the 1000 genomes project (1KGP).

While VarCoPP was an important milestone towards predictive oligogenic methods, it suffered from a high false positive rate and it required significant computational resources.

In order to address these issues, we created VarCoPP2.0: a simplified, faster and more accurate model to classify potentially pathogenic variant combinations.

Novel features have been identified via an original wrapper method and newly curated training data stemming from OLIDA [3] and 1KGP were used to train the model.

Additionally, model complexity has been significantly reduced through the use of a single balanced random forest.

Results on a new independent data set reveal that VarCoPP2.0 brings the analysis of potentially disease-associated variant combinations to a new level.

[1] Papadimitriou, Sofia, et al. "Predicting disease-causing variant combinations." Proceedings of the National Academy of Sciences 116.24 (2019): 11878-11887.

[2] Gazzo, Andrea M., et al. "DIDA: A curated and annotated digenic diseases database." Nucleic acids research 44.D1 (2016): D900-D907.

[3] Nachtegael, Charlotte, et al. "Scaling up oligogenic diseases research with OLIDA: the Oligogenic Diseases Database." Database 2022 (2022).

Genomes

Target-Oriented miRNA Discovery (TOMiD) Analysis of qCLASH Ribonomics Experiments

Daniel Stribling (University of Florida, Department of Molecular Genetics & Microbiology), Nicholas Hiers (University of Florida, Department of Biochemistry & Molecular Biology), Mingyi Xie (University of Florida, Department of Biochemistry & Molecular Biology) and Rolf Renne (University of Florida, Department of Molecular Genetics & Microbiology).

Abstract:

MicroRNAs (miRNAs) are short, noncoding RNAs that play a major role in cellular regulation and disease in many plant and animal species by binding and repressing target transcripts. The Quick Crosslinking, Ligation, and Sequencing of Hybrids (qCLASH) ribonomics technique has recently provided significant insight on miRNA function by high-throughput identification of miRNA targets via RNA ligation and sequencing of chimeric (hybrid) reads. Previous analysis methods for identification of qCLASH hybrids require a well-annotated reference transcriptome, limiting this technique to use with known miRNAs and in species with well-characterized genomes. In order to expand the scope of qCLASH, we developed the Target-Oriented miRNA Discovery (TOMiD) method to enable identification of novel miRNAs using qCLASH experiments. TOMiD uses an annotation-naïve approach for identification of miRNA / target hybrids by first identifying the target portion of a potential hybrid, then evaluating the remaining region for miRNA-associated characteristics, including region length, coverage, predicted physical properties, and other selected criteria. In addition to recapturing a high proportion of known hybrids compared to traditional qCLASH analysis, TOMiD has successfully identified a novel human miRNA: miR-snaR found to promote cell migration in human breast cancers (1). We are currently implementing TOMiD in an open-source pipeline in the Nextflow programming language to enable flexible and reproducible use of this analysis across a variety of model and non-model species, expanding the scope of the qCLASH method.

(1) Stribling et al. RNA (2021). A non-canonical microRNA derived from the snaR-A non-coding RNA targets a metastasis inhibitor.

Genomes

The additive effects of Next Generation Sequencing error sources on the quality of de novo genome assembly

Alex Váradi (University of Debrecen), Zoltán Rádai (University of Debrecen), Nikoletta A Nagy (University of Debrecen), Péter Takács (University of Debrecen), Gábor Kardos (University of Debrecen) and Levente Laczkó (University of Debrecen).

Abstract:

Next Generation Sequencing methods (NGS) revolutionized the study of prokaryotes providing a cost-efficient method for determining their whole genome sequence. Illumina's technology has become the most widespread NGS platform of DNA sequencing used by scientists across the globe to sequence whole genomes. Although the effect of Illumina's error rate on de novo genome assembly is known, the similar effect of PCR duplication artifacts and optical duplicate reads has never been investigated comprehensively. We used genomes of 13 species covering a range of genome sizes (1-6.43 Mbp) and GC ratios (36-66%) to simulate paired-end short read sequencing libraries with 150 base pair long reads. Read simulation and de novo assembly were performed in five replicates by varying coverage (25×-150×), error rate (0, 0.01, 0.025, 0.05), and the rate of optical and PCR duplicate production (0, 0.01, 0.05, 0.15, 0.3). We assessed the effects of these input parameters on the contiguity metrics of the assemblies using generalized linear regression modelling (GLM), then the results were evaluated against individual genomic characteristics (genome size, GC %). We observed that PCR- and optical duplication artifacts affect assembly contiguity and the effect size depends on the unique genome characteristics. Genomes of low complexity produced more assembly biases when the error rate was larger than zero, which was enhanced by deep coverage. Thus, our results show not only that PCR and optical duplications significantly affect de novo assembly quality, but that high coverage can exacerbate their effects.

Genomes

The genome of Belgium: whole genome sequencing of the Belgian population to enable public health genomics

Thomas Delcourt (sciensano), Charlotte De Vogelaere (sciensano), Emilie Cauët (sciensano), Nina Van Goethem (sciensano), Johan Van der Heyden (sciensano), Stefaan Demarest (sciensano), Karin De Ridder (sciensano), Nancy Roosens (sciensano), Marc Van Den Bulcke (sciensano) and Kevin Vanneste (sciensano).

Abstract:

To enable public health genomics in Belgium, the Belgian Institute for Health (Sciensano) is setting up infrastructure and is developing expertise to process hundreds of Whole Genome Sequencings (WGS) of human blood samples from the Belgian population. A germline short variant discovery pipeline for WGS data based on the GATK best practices was implemented, validated using nine extensively characterized datasets (seven from the GiaB consortium, one from the Platinum Genomes project and a synthetic dataset from Li et al. (2018)) and benchmarked against two FDA precision challenges. As a pilot set-up, we analyzed one hundred WGS datasets obtained during the 2018 Health Examination Survey (HES). Nineteen pathogenic variants were identified in genes implicated in cancer and fourteen variants were found in medically actionable genes (ACMG SF v3.0). Ancestry of samples was assessed with Principal Component Analysis (PCA) against the 1000 Genomes Project dataset, which showed that the majority of Belgian samples grouped with the European super-population. Future work will involve sequencing additional samples from the HES cohort and new Belgian population-based studies, additional evaluation of the benefit of population scale WGS for various areas of research and medical fields (cancer genomics, rare diseases, personalized prevention), comparisons of the Belgian population against close European populations based on large-scale sequencing efforts (e.g. Genome of the Netherlands), linkage with Belgian public health-related databases (HES, cancer registry…) and identification of population-specific variants and their added value to the quality of imputation of rare variants in the Belgian population.

Genomes

The Spatial Organization Of Enhancers Around Promoter Regions Within Chromatin Contact Domains For Selected Human Cell Lines: Structural Regulatory Landscape

Abhishek Agarwal (Centre of New Technologies, University of Warsaw) and Dariusz Plewczyński (Centre of New Technologies, University of Warsaw).

Abstract:

Our research aims to develop and test the concept of the structural epigenomic landscape (SEL) of regulatory elements around promoter regions for selected cell types and the different individuals of the human population. We will propose a biophysical method to construct probabilistic ensembles of three-dimensional conformations at genomic domains scale (i.e. for chromatin contact domains - CCDs, or topologically associating domains - TADs), compartments, chromosomal territories and finally at the whole genome-scale. The chromatin loops mediated by cohesin were used to identify the set of genomic domains, chromatin contact domains (CCDs). Further, we defined reference sets of enhancers using epigenomics datasets collected by ENCODE across a compendium of cell and tissue types. In our analysis, we focused on specific histone modification (H3K27ac ChIP-seq), and chromatin accessibility data based on DNase I hypersensitive sites (DHS) identified with DNase-seq. In parallel, we are processing the RNA-Seq datasets for the corresponding individuals to find the common and unique differentially expressed genes across family members. Our plan is to construct a comprehensive set of spatial networks across different individuals representing the enhancer-promoter (EP) pairs for lymphoblastoid cell lines. We are also planning to compare the cohesin, CTCF and RNAPOL2-mediated chromatin interactions in order to understand the spatial mechanism of the transcriptional process.

Genomes

Towards an atlas of copy-number conditional selection on somatic mutations in cancer driver genes

Elizaveta Besedina (IRB Barcelona) and Fran Supek (IRB Barcelona).

Abstract:

Many statistical methods have been developed for discovering the genes under selection in tumor genomes, modeling background mutation rates using covariates, such as DNA replication timing or transcription activity. However, these approaches were not designed to detect condition-specific selection (i.e., a change in selection intensity appearing under certain circumstances). We have developed a statistical methodology that aims to do so while stringently controlling for the mutation rate heterogeneity across the genome and for trinucleotide mutation signatures. Our log-linear based dN/dES methodology estimates a baseline mutation rate from neighboring genes ("Neighbors" method) or from sites in the gene where mutations have a low impact ("CADD" method). Optionally, our methodology can draw on the whole-genome sequencing data by deriving a mutational baseline from introns and flanking non-coding regions.

To study the interplay between selection on cancer genes and copy number changes, such as same locus deletion or amplification, we applied the CADD and Neighbors methods to somatic mutation datasets (panel sequencing, and exome/genome sequencing) separately for different mutation classes: nonsense, missense and synonymous. The estimates for selection in the diploid state, selection change upon deletion, and amplification in various cancer types were summarized using an NMF algorithm to create an atlas of copy-number dependent selection in the human soma. Selection patterns reflect the mechanisms of driver events in cancer genes, for example, two-hit loss driver, or one-hit driver. The latter category, for tumor suppressor genes, could be further split into haploinsufficient genes or genes with dominant-negative acting mutations.

Genomes

TP53-dependent toxicity of CRISPR/Cas9 cuts is differential across genomic loci and can confound genetic screening

Miguel M Álvarez (Institute for Research in Biomedicine (IRB Barcelona)), Josep Biayna (Department of General, Visceral, Transplant, Vascular and Pediatric Surgery, University Hospital, Würzburg) and Fran Supek (Institute for Research in Biomedicine (IRB Barcelona)).

Abstract:

CRISPR/Cas9 gene editing can inactivate genes in a precise manner, however this process involves DNA double-strand breaks (DSB) and repair thereof, which may incur a loss of fitness. We hypothesized DSB toxicity may be variable depending on the chromatin environment in the targeted locus. By using a previously described genome-wide Cas9 library, we detected an excess of genes whose knockout decreases fitness in wild-type lung cancer cells, but not in their TP53-/- isogenic counterpart, independently of the gene function. By analyzing isogenic pair experiments jointly with previous CRISPR screening data from across ~900 cell lines, we found that p53-associated break toxicity is higher in genomic regions that harbor active chromatin, such as gene regulatory elements or transcription elongation histone marks, while it is lower in heterochromatin. In addition, activity of homologous recombination repair and microhomology-mediated end joining pathways, as well as specific sequence patterns in the vicinity of the target site also associate with the fitness effects across genomic loci. Due to noise introduced by differential toxicity of sgRNA-targeted sites, the power of genetic screens to detect conditional essentiality is reduced in TP53 wild-type cells, which we demonstrate using an example of an ATR inhibitor screen. Understanding the determinants of Cas9 cut toxicity will help improve design of CRISPR reagents to avoid incidental selection of TP53-deficient and/or DNA repair deficient cells, and to improve genome-wide screening for identifying conditionally essential genes.

Genomes

Training DeepSignal models to call CpG methylation in pig and quail ONT reads

Paul Terzian (INRAE), Céline Vandecasteele (INRAE), Christine Gaspin (INRAE), Cécile Donnadieu (INRAE), Denis Milan (INRAE), Rémi-Félix Serre (INRAE) and Christophe Klopp (INRAE).

Abstract:

DNA methylation calling from ONT (Oxford Nanopore Technologies) reads relies on prediction models. As today, the most accurate caller tools for ONT reads (DeepSignal, Megalodon) provide calling models specific to a wide range of conditions. Models can be designed for pore types such as R9.4.1 or R10.3 or for modifications types such as 5mC, 6mA or even more specific like in CpG context or GATC motif in bacteria. This range of conditions raises issues such as how specific to our DNA sample should a model be or can a model trained with human reads be accurate on other species. These issues have no commonly accepted answer yet as very few open access datasets and scientific studies are available.

In the SeqOccIn project, we produced novel DeepSignal models for pig (Sus scrofa) and quail (Coturnix japonica) based on native datasets. We then compared their accuracy with the reference human CpG models.

Our preliminary results show that DeepDignal and Megalodon CpG models have different accuracies when calling modifications from pig and quail native samples. After an overview of the training and evaluation processes, we present preliminary results obtained by using different training strategies and sequencing technologies applied to pig and quail datasets. Interestingly, the most effective strategy to improve DeepSignal model accuracy is to train the model with a pool of reads from both species. Such a model achieves a better accuracy than both DeepSignal and Megalodon available human CpG models and all our mono-species pig or quail de novo models.

Genomes

Transformer Language Models for Genomic Sequences

Vlastimil Martinek (CEITEC MU), David Cechak (CEITEC MU), Petr Simecek (CEITEC MU) and Panagiotis Alexiou (CEITEC MU).

Abstract:

Since long short-term memory (LSTM) architecture, neural networks have been proven helpful for natural language processing (NLP) tasks. But what about the language of genomic sequences written in a four-letter alphabet? In the last few years, researchers have demonstrated that neural networks can be used to identify functional elements in genomic sequences and that the resulting models can predict the function of previously uncharacterized genomic regions.

The "revolution" had come in 2018 with the transformer architecture having the ability to detect complex dependencies between elements of a series thanks to a mechanism of attention or self-attention. Our approach explores pre-training these models on genomes using language models and masked language modeling objective. The pre-trained language models are then fine-tuned to specific downstream classification tasks. For this purpose, we explore various popular transformer architectures (Bert, DistillBest, Deberta, MobileBert, Albert, Perceiver) and genomes of multiple organisms (human, mouse, drosophila, c'elegans, zebrafish). Finally, we demonstrate the importance of organism-specific pre-training and improved downstream task performance on multiple datasets using pre-trained language models.

Genomes

Transgenerational epigenetics in quail: whole genome DNA methylation analysis

Chloé Cerutti (INRAE), Sophie Leroux (INRAE), Paul Terzian (INRAE), David Gourichon (INRAE), Frederique Pitel (INRAE) and Guillaume Devailly (INRAE).

Abstract:

The influence of the prenatal environment on the adult phenotype development is partially mediated by epigenetics phenomena. One study highlighted significant effects of the in-ovo injection of endocrine disruptors or DNA methyltransferase inhibitor on quail development, significantly reducing their weight [1]. Recently, an increasing number of studies highlighted the transmission of epigenetics marks between generations following an environmental exposure. However, there is much debate about their acquired transmission beyond the exposed individuals. Recent studies revealed that non-genetics inheritance was probably present in avian species. In one of them [2], fertilized eggs were divided into two groups: one group injected with an endocrine disruptor, Genistein, and a non-injected control group. After three generations without any other injection, several traits were impacted by the ancestor treatment such as the reproduction and the behavior. This pilot study highlighted the potential existence of transgenerational transmission of environmental effects in quails. To better understand the transgenerational transmission of these environmental effects, DNA methylation data are available from blood samples from the third generation (WGBS and ONT). To analyse the DNA methylation state, we developed a bioinformatics pipeline in order to detect differential methylated cytosines (DMCs) affected by these transgenerational phenomena. We detected thousands of DMCs between both groups. In addition to these analyses, these data allowed us to perform a comparative analysis between the WGBS and ONT sequencing technologies.

References:

1. Cerutti, C., et al., Animal, 2021. 16(3). 1751-7311.

2. Leroux, S., et al., Genet Sel Evol, 2017. 49(1). 14.

Genomes

Unbiased discovery of diversity-generating mechanisms and mobile genetic elements

Jordi Abante (Stanford University) and Julia Salzman (Stanford University).

Abstract:

Current genomic databases contain petabytes of sequencing data, providing an excellent resource for computational methods to jumpstart ground-breaking discoveries. For example, CRISPR was hypothesized to be a genome-modifying system as early as 2000 using purely computational methods. In addition, diversity-generating mechanisms (DGMs), such as CRISPR, and mobile genetic elements (MGEs) actively contribute to the greater virulence and antibiotic resistance of several pathogenic bacteria, which remain a critical cause of morbidity and mortality worldwide.

Unfortunately, the current paradigm to discover DGMs and MGEs is arguably exhausted. State-of-the-art algorithms rely on (meta)genome assemblers, reference genomes, and substantial heuristics, resulting in computational and time bottlenecks. Furthermore, given the immense diversity and rapid evolution inherent to bacteria and viruses, assemblies and reference genomes introduce substantial biases in the analysis. As a result, the current paradigm has significant shortcomings that limit the breadth of potential biological discoveries.

Here, we present a novel statistical algorithm that enables the discovery of MGEs and DGMs directly from sequencing data alone. The potential of the proposed approach is incredibly vast, ranging from the discovery of new CRISPR-like mechanisms, both in bacteria and viruses, to the discovery of novel MGEs and enzymes. As a result, our work can lead to ground-breaking discoveries with potentially transformative implications to genetics, cell biology, and medicine. The application of our algorithm to E. coli sequencing data resulted in the discovery of over 3,000 known MGEs and several CRISPR systems in under two hours.

Genomes

Uncovering signatures of mutational processes in SARS-CoV-2

Kieran Lamb (University of Glasgow).

Abstract:

Mutational signatures are patterns of mutations that can be attributed to a mutational process. Uncovering mutational processes using signature extraction is a task that has so far been pioneered in cancer studies. Over the last decade, expanding datasets and the development of new methods have increased our understanding of the effects of mutational processes on cancer. There is now a pan-cancer library of verified mutational signatures for both cellular (e.g DNA mismatch repair) and non-cellular (e.g UV light exposure) mutational processes. Successfully extracting these signatures required a large set of mutation rich sequence data to analyse. Given the vast repository of viral sequences produced as a result of the SARS-CoV-2 pandemic, signature extraction could be attempted in order to understand the mutational landscape affecting the virus (particularly with regards to the host immune response). Here we demonstrate how our computational framework can successfully extract mutational signatures from viral sequences. These signatures represent the mutational processes operating on the virus throughout the pandemic, and are accompanied by the exposures that show their changing activity. The framework takes into account key differences between cancer and viral sequence data without which signature extraction would prove challenging. The framework shows the mutational process dynamics as they have changed through the pandemic, and identifies signatures for the processes that have enacted changes on SARS-CoV-2.

Genomes

Untangling Costa Rican cocoa bean fermentation processes using a combined shotgun metagenomics, metatranscriptomics, and meta-metabolomics approach

Stefan Weckx (Vrije Universiteit Brussel), Marko Verce (Vrije Universiteit Brussel) and Luc De Vuyst (Vrije Universiteit Brussel).

Abstract:

Cocoa fermentation is the first step in the transformation of raw cocoa beans into chocolate. The fermentation process is important for the removal of the mucilaginous cocoa pulp and the development of flavour and colour precursors within the cocoa beans. Shotgun metagenomic and metatranscriptomic sequencing were applied to Costa Rican cocoa fermentation processes to unravel the microbial diversity and assess the function and transcription of their genes. Meta-metabolomics was applied to obtain data accompanying the metatranscriptomics data. Among 82 genera found in these fermentation processes, the major ones were Lactobacillus-derived genera, Acetobacter, Hanseniaspora, Komagataeibacter, Leuconostoc, Saccharomyces, and Pectobacterium. The most abundant species were Limosilactobacillus fermentum, Liquorilactobacillus cacaonum, and Lactiplantibacillus plantarum among the lactic acid bacteria, Acetobacter pasteurianus and Acetobacter ghanensis among the acetic acid bacteria, and Hanseniaspora opuntiae and Saccharomyces cerevisiae among the yeasts. Consumption of glucose, fructose, and citric acid, and the production of ethanol, lactic acid, acetic acid, and mannitol were linked to the major species mentioned above through metagenomic binning and the application of metatranscriptomic sequencing. By using this approach, new insights were obtained not revealed by previous research approaches, such as the fact that Lacp. plantarum consumed mannitol and oxidised lactic acid, that A. pasteurianus degraded oxalate, and that bacterial species, such as Cellvibrio sp., Pectobacterium spp., and Paucilactobacillus vaccinostercus, could contribute to pectin degradation. The knowledge obtained will enhance the selection and development of appropriate starter cultures to transform the spontaneous cocoa fermentation process into a controlled one.

Genomes

Unveiling epigenetic profiles of transposable elements in ChIP-seq data using T3E

Michelle Almeida da Paz (Institute of Biomedical Informatics, Graz University of Technology, Graz, Austria) and Leila Taher (Institute of Biomedical Informatics, Graz University of Technology, Graz, Austria).

Abstract:

Transposable elements (TEs) comprise almost half of the human genome and the study of their epigenetic profiles has become a hot topic. Chromatin Immunoprecipitation Sequencing (ChIP-seq) technologies have helped reveal the cis-regulatory roles of TEs. However, several technical challenges are faced. The most prominent difficulty concerns the ambiguous mapping of reads derived from TEs, especially those that were recently proliferating. Another difficulty is the definition of an appropriate background for enrichment analysis. The standard approach randomly permutes the location of the read mappings in the genome, assuming a uniform distribution that does not reflect biases in library preparation or TE insertion patterns, and often leads to false positive or negative TE enrichments. To tackle these issues we developed the Transposable Element Enrichment Estimator (T3E), a framework for the functional analysis of TEs. T3E compares the epigenetic profiles of TE families/subfamilies to a background profile constructed based on the structure of the ChIP-seq control experiment. Another innovation of T3E is how it estimates the coverage of TE families/subfamilies. Specifically, acknowledging the ambiguity of the data, T3E weights the number of reads mapping to a TE family/subfamily copy by the overall number of loci to which the reads map in the genome, and this is done at single-nucleotide resolution. We applied T3E to confirm or refute previous results, and we found evidence supporting the implication of TEs in important cellular processes.

Genomes

Uplifting trimAl for handling thousands of sequences

Nicolás Díaz Roussel (Barcelona Supercomputing Center) and Salvador Capella Gutiérrez (Barcelona Supercomputing Center).

Abstract:

Motivation: Multiple sequence alignments are the basis of many downstream analysis such as phylogenetic trees reconstruction. The appearance of NGS has fueled many large-scale studies that imply the generation of an incredible number of alignments, making the use of automated filtering methods a requirement.

Results: We show how the use of machine-learning techniques allows us to generate new automated filtering algorithms based on the most relevant MSAs features. This work is complemented by additional investigations on how to extend the benchmark to measure the impact of trimming on (very) large alignments in terms of global accuracy. Finally, we demonstrate the technical capabilities of the newest trimAl version when there are no limitations in terms of computational resources.

Genomes

Using AlphaFold2-generated structural information to improve detection of evolutionary adaptations in proteins

Sophie-Luise Heidig (Interuniversity Institute Of Bioinformatics Brussels), Ravy Leon Foun Lin (Universite Claude Bernard - Lyon 1), Danny Ionescu (Leibniz-Institut für Gewässerökologie und Binnenfischerei (IGB)), Jean-François Flot (Université libre de Bruxelles) and Wim Vranken (Vrije Universiteit Brussel).

Abstract:

A common issue when analyzing the evolution of proteins is low-quality multiple sequence alignments (MSAs) causing artifacts. Assessing the ratio of non-synonymous over synonymous substitution rates (dN/dS) particularly suffers from this. It is difficult to unravel whether sequence diversification associated with high dN/dS values results in poor alignments or poor alignments result in artefactually increased values. MSA quality can be improved by including structural information but resolving protein structure is a tedious experimental process. However, with the release of AlphaFold2 it is possible to generate high-confidence structures directly from protein sequences. We hypothesized that this predicted structural data could enhance protein evolution analyses by improving the MSA step. Therefore, we created a pipeline to streamline the processing of information from the acquisition of gene sequences from online databases, parsing of ortholog groups, clustering, selection of representative sequences for structure prediction with AlphaFold2, generation of structure informed MSAs, construction of phylogenetic trees and finally calculation of dN/dS. We tested this pipeline on predicted proteins from published genomes of Synechococcus, a cyanobacterial genus abundant in sea surface waters. Our results indicate that using structure-informed alignments as basis of dN/dS calculations leads to less frequent, more extended gaps and lower dN/dS values at gaps in otherwise dispersed alignments. The remaining occurrences of high dN/dS can be used as a proxy to detect selective pressure. The availability of this new pipeline paves the way for efficient large-scale analysis of protein evolution across the tree of life.

Genomes

Variable DNA methylation underlies mutation rate variability at the mesoscale in human somatic cells

David Mas-Ponte (Institute for Research in Biomedicine (IRB Barcelona)) and Fran Supek (Institute for Research in Biomedicine (IRB Barcelona)).

Abstract:

The cytosine methylation in CpG dinucleotides is pervasive in mammalian genomes and its variability across regions can regulate gene expression and define cell differentiation. Although the role of DNA methylation in gene regulation is well understood, how the local variation in DNA methylation shapes somatic mutation rates is less well explored. Here, we show that hypomethylated (UMR) regions are also generally hypomutated in a wide range of human tumors and healthy somatic tissues.

Remarkably, the exposure of the tissue to various mutational processes shapes its predisposition to this effect: while there is depletion in the mutation rates resulting from signatures of deamination of methylated cytosines, UV light, POLE and MMR deficiency, there is an increase in mutation rates from signatures of AID/APOBEC cytosine deaminase enzymes in the UMRs. Therefore, hypomethylated DNA loci can be either mutational coldspots or hotspots, depending on the mutagen exposure history of a particular cell.

In addition to these genome-wide distributed UMRs we also identify several kilobases at the 5' ends of gene bodies as commonly hypomethylated and thus hypomutated. Clustering genes by methylation profiles also yielded variability in their mutation rate gradients along the gene body. Interestingly, lowly expressed genes have a less steep gradient due to a higher relative methylation of their 5' end, and polycomb repressed genes also show no relative hypomutation due to the lack of methylation at their gene body.

Overall, we suggest DNA methylation is an important determinant of mesoscale, sub-genic, resolution mutation rate variability in human somatic tissues.

Genomes

VARIANT PRIORITIZATION IN PLASMA WHOLE-EXOME SEQUENCING FOR THE IDENTIFICATION OF POTENTIAL THERAPEUTIC TARGETS IN RELAPSING COLON CANCER PATIENTS

Jorge Martín-Arana (Department of Medical Oncology, INCLIVA Biomedical Research Institute, University of Valencia, Valencia, Spain), Francisco Gimeno-Valiente (Cancer Evolution and Genomic Instability Laboratory, University College London Cancer Institute, London, UK), Roberto Tébar-Martínez (Department of Medical Oncology, INCLIVA Biomedical Research Institute, University of Valencia, Valencia, Spain), Blanca García-Micó (Department of Medical Oncology, INCLIVA Biomedical Research Institute, University of Valencia, Valencia, Spain), Valentina Gambardella (Department of Medical Oncology, INCLIVA, Valencia, Spain. Health Institute Carlos III, CIBERONC, Madrid, Spain), Marisol Huerta (Department of Medical Oncology, INCLIVA Biomedical Research Institute, University of Valencia, Valencia, Spain), Carolina Martínez-Ciarpaglini (Department of Pathology, INCLIVA Biomedical Research Institute, Valencia, Spain), Juan Antonio Carbonell-Asins (Bioinformatics and Biostatistics Unit, INCLIVA Biomedical Research Institute, Valencia, Spain), Manuel Cabeza-Segura (Department of Medical Oncology, INCLIVA Biomedical Research Institute, University of Valencia, Valencia, Spain), José Martín-Arévalo (Department of Surgery, INCLIVA Biomedical Research Institute, Valencia, Spain), David Casado (Department of Surgery, INCLIVA Biomedical Research Institute, Valencia, Spain), Vicente Pla (Department of Surgery, INCLIVA Biomedical Research Institute, Valencia, Spain), Leticia Pérez (Department of Surgery, INCLIVA Biomedical Research Institute, Valencia, Spain), Pilar Rentero-Garrido (Precision Medicine Unit, INCLIVA Biomedical Research Institute, Valencia, Spain), Sheila Zúñiga-Trejos (Precision Medicine Unit, INCLIVA Biomedical Research Institute, Valencia, Spain), Susana Roselló (Department of Medical Oncology, INCLIVA Biomedical Research Institute, University of Valencia, Valencia, Spain), Tania Fleitas (Department of Medical Oncology, INCLIVA Biomedical Research Institute, University of Valencia, Valencia, Spain), Josefa Castillo (Department of Biochemistry and Molecular Biology, Universitat de Valencia, Valencia, Spain), Desamparados Roda (Department of Medical Oncology, INCLIVA, Valencia, Spain. Health Institute Carlos III, CIBERONC, Madrid, Spain), Andrés Cervantes (Department of Medical Oncology, INCLIVA, Valencia, Spain. Health Institute Carlos III, CIBERONC, Madrid, Spain) and Noelia Tarazona (Department of Medical Oncology, INCLIVA, Valencia, Spain. Health Institute Carlos III, CIBERONC, Madrid, Spain).

Abstract:

INTRODUCTION: Circulating tumor DNA (ctDNA) detection in plasma at the postoperative period indicates relapse in colon cancer (CC) patients. Nevertheless, about 80% of patients with positive ctDNA do not become negative after conventional therapy (ACT). Whole-exome sequencing (WES) of ctDNA may help molecularly characterize the tumor and identify potentially actionable mutations that guide targeted treatments with better response than ACT. However, a large number of variants are identified and a mutational prioritization process is necessary to point the oncogenic mutations.

METHODS: WES of paired plasma at baseline and relapse and peripheral blood of 25 CC patients was performed including Unique Molecular Identifiers technology. Oncogenic somatic variants were identified by annotating with Ensembl, COSMIC, dbSNP and matched with OncoKB levels

of evidence for the choice of specific targeted therapies. Functional enrichment analysis was performed by selecting cancer hallmarks with FDR < 0.05.

RESULTS: The variant prioritization process significantly reduced the number of variants detected (p-value=5.96e-08), being able to identify at least one potential therapeutic target in 80% of patients. Comparing the mutational profile of baseline and relapse stages, 26.5% of mutations were exclusively identified at relapse related to cellular stress response and WNT signaling pathways. Furthermore, a significant mutational enrichment was detected in genes related to immune evasion.

CONCLUSIONS: The variant prioritization process helps to clarify the mutational information obtained from plasma WES and may aid in the detection of alterations with therapeutic value. Relapse and baseline mutational comparison helps to identify the tumor progression mechanism that leads to the relapse.

Genomes

Vessel type-specific GWAS of Retinal Vessel Tortuosity Identifies 173 Novel Loci Revealing Genes and Pathways Associated with Vascular Pathomechanisms and Diseases

Michael Beyeler (University Lausanne), Sofia Ortin Vela (University of Lausanne), Mattia Tomasoni (University of Lausanne) and Sven Bergmann (University of Lausanne - Department of Computational Biology).

Abstract:

Fundus images allow for non-invasive assessment of the retinal vasculature whose features provide important information on health.

We analysed fundus images of 62 751 participants in the UK Biobank. We built a fully automated image processing pipeline to annotate vessels, using a deep learning algorithm to segment the images and determine the vessel type, characterising participants in terms of their median retinal vessel tortuosity specific to arteries and to veins. Tortuosity was measured by the length of a vessel segment over its chord length, as well as measures that integrate over vessel curvature. Higher tortuosity was significantly associated with higher incidence of angina, myocardial infarction, stroke, deep vein thrombosis, and hypertension. Our GWAS identified 175 significantly associated genetic loci in the UK Biobank; 173 of these were novel and 4 replicated in our second, much smaller, meta-cohort. We estimated heritability at ~25% using linkage disequilibrium score regression. Vessel type specific GWAS revealed 114 loci for arteries and 63 for veins. Genes with significant association signals included COL4A2, ACTN4, LGALS4, TNS1, MAP4K1, EIF3K, CAPN12, ECH1, and SYNPO2, and were overexpressed in arteries and heart muscle. Pathway analysis with PascalX provided links to structural properties of the vasculature. Several alleles associated with retinal vessel tortuosity pointed to a common genetic architecture of this trait with cardiovascular diseases and metabolic syndrome.

Our results shed new light on the genetics of vascular diseases and their pathomechanisms and highlight how GWASs and heritability can be used to improve phenotype extraction from high-dimensional data, such as images.

Genomes

What is the reality? – Influence of the sequencing technology and the approach on microbial community composition estimation

Dedan Githae (Malopolska center of Biotechnology- Jagiellonian University), Agata Jarosz (Malopolska center of Biotechnology- Jagiellonian University), Kinga Herda (Malopolska center of Biotechnology- Jagiellonian University), Kamila Marszałek (Malopolska center of Biotechnology- Jagiellonian University), Wojciech Branicki (Malopolska center of Biotechnology- Jagiellonian University) and Paweł Łabaj (Malopolska center of Biotechnology- Jagiellonian University).

Abstract:

Advancements in sequencing technologies has made uncovering the entire microbial communities present in an environment not only possible, but achievable at a faster rate than ever before. Characterisation of the metagenomic composition not only gives taxonomic profile to determine species richness and diversity that describes an ecological niche, but also allows for the discovery of new and novel species. Although it was already previously possible with 16S sequencing, now the Whole Metagenome Sequencing is providing resolution not seen before. Identification of stable microbiomes can be used as a baseline to describe their environment, having applications such as surveillance and monitoring the spread of pathogens of interest, forensic intelligence on identifying the source of the microbial specimen, among other related uses.

Here, we present the findings from studying samples from multiple different ecological niches, sequenced using different sequencing technologies (long-reads vs short reads) as well as exploiting different sequencing approaches (probe-based targeted sequencing vs whole metagenomic sequencing). We do show the consistency and reliability in obtaining the environmental microbial profiles by different strategies. We also do show the power and limitations of each for specific applications.

Genomes

Whole genome of a biparental beetle species, Anoplotrupes stercorosus

Nikoletta A Nagy (University of Debrecen), Levente Laczkó (University of Debrecen) and Zoltán Barta (University of Debrecen).

Abstract:

One of the most complex social behaviour is parental care which occurs in diverse forms across the animal kingdom. Despite its prevalence, the evolution and molecular mechanisms regulating this behaviour are still not fully uncovered, especially in invertebrates including insects. One of the most important tool to understand these processes is to identify the key genes involved in forming parental care through comparative genomics. The first step in finding such genes is to determine the genome sequences of as many species with analogous parental care as possible.

The beetle species Anoplotrupes stercorosus belongs to the family Geotrupidae. This beetle is common in Europe and has biparental care during which the parents create an underground nest and provision food for the offspring in advance of their hatching. In this study, we collected male and female individuals from a natural population of A. stercorosus for high molecular weight DNA isolation. We combined Illumina and Oxford Nanopore sequencing data to reconstruct the whole genome sequence of the species. Based on the assembly, the size of the genome is ca. 840 Mbp with 34.47% GC content and N50 of 0.5 Mbp. The genome has a high completeness value with 96.5% of complete sequences from the Endopterygota BUSCO collection, however, a relatively high proportion of duplicated sequences (12%) was found. After structural and functional annotation, this genome can contribute to the comparative analyses aiming to discover the evolution and regulation of parental care.

Proteins

3DBionotes COVID-19 Structural Hub: a central resource for validation information and refined models

Jose Ramon Macias Gonzalez (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)), Carolina Simon Guerrero (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)), Erney Ramírez Aportela (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)), Jose Luis Vilas Prieto (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)), Marta Martinez Gonzalez (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)), Carlos Oscar Sanchez Sorzano (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)) and Jose Maria Carazo (Biocomputing Unit (BCU), National Centre of Biotechnology (CNB-CSIC)/Instruct Image Processing Centre (I2PC)).

Abstract:

3DBionotes-WS API offers a set of Web Services that aggregate and distribute multiple functional, genomic, proteomic and structural feature annotations oriented to structural biology analysis serving as a base for a website including a fully interactive 3D viewer for macromolecular models and Cryo-EM derived maps.

During the COVID-19 pandemic outbreak, the urgency of having rapid access to molecular structures as a basis for finding potential drugs and vaccines has led to an unprecedented effort to obtain atomic models of viral proteins and deploy them in common repositories. In line with previous concerns raised, our group has aligned with other initiatives to bring more attention not only to the quantity of data, but also to the quality of the data.

With the COVID-19 Structural Hub we have dedicated a website section to collect and provide easy access to all available structural information related to the topic. New annotation types have recently been added with validation information about cryo-EM maps, atomic models, and map-to-model fitting. In addition to re-refined models by PDB-REDO, CERES and the Coronavirus Structural Task Force, we provide annotations for local resolution validation methods like DeepRes, MonoRes and BlocRes. If an atomic model is available, we also compute MapQ and FSC-Q scores, a quantitative estimation of how much of the model is supported by the signal content of the map. Our goal is to assess the quality of maps and atomic models not only globally, but also from a local perspective.

Proteins

A graph based method for identifying and clustering short tandem repeats

Patryk Jarnot (Silesian University of Technology), Joanna Ziemska-Legięcka (Institute of Biochemistry and Biophysics, Polish Academy of Sciences), Marcin Grynberg (Institute of Biochemistry and Biophysics Polish Academy of Sciences), Aleksandra Gruca (Silesian University of Technology) and Vasilis Promponas (University of Cyprus).

Abstract:

Short Tandem Repeats (STRs) are fragments of protein sequences containing short neighbouring duplications of residue pattern. The importance of these fragments is backed by numerous research about their functional and structural properties. For instance, they frequently play an important role in molecular interactions. To discover new biological properties of STRs, scientists compare them to other similar STRs to conclude about their biological properties, ideally confirming with a wet-lab experiment which can be costly and time consuming. Here we present a method which is designed to identify and cluster STRs where both steps are realised in a single model.

Graph Based on Sequence Clustering (GBSC) method builds De Bruijn-like graphs by scanning the sequence with a given window length. In these graphs transitions connect neighbouring k-mers which represent nodes. Sequences which are represented by cycles are marked as STRs. To find similar STRs clustering is done by grouping sequences represented by similar graphs. GBSC uses the same model for both identifying and clustering which makes it a fast method. To examine its usability, we compare GBSC to other methods for identifying tandem repeats and protein sequence clustering, highlighting their strengths and weaknesses. We provide as test cases (1) the analysis of Pfam families tagged as repetitive and (2) GO enrichment analysis of GBSC clusters for UniprotKB/Swiss-Prot sequences. In conclusion, our method efficiently clusters large sequence datasets by STRs tagged during identification. (Co-financed by the EU grant POWR.03.02.00-00-I029/17)

Proteins

A graph-based algorithm for detecting rigid domains in protein structures

Linh Dang (Universitätmedizin Göttingen), Thach Nguyen (IUF Duesseldorf) and Michael Habeck (University Hospital Jena).

Abstract:

Conformational transitions are implicated in the biological function of many proteins. Structural changes in proteins can be described approximately as the relative movement of rigid domains against each other. A powerful approach to understand structural transitions in proteins is to decompose structures of different states into rigid domains and classify protein movements by hinge and shear motions of these structural domains.

Thus, we develop a new graph-based method for detecting rigid domains in proteins where the number of rigid domains is automatically estimated.

Proteins

Alternative to combat superbugs – dynamic determinants of quorum quenching enzymes and their engineering towards efficient antibacterial potency

Bartlomiej Surpeta (Adam Mickiewicz University in Poznan & International Institute of Molecular and Cell Biology in Warsaw), Michal Grulich (Institute of Microbiology, v.v.i., Academy of Sciences of the Czech Republic), Andrea Palyzová (Institute of Microbiology, v.v.i., Academy of Sciences of the Czech Republic), Helena Marešová (Institute of Microbiology, v.v.i., Academy of Sciences of the Czech Republic) and Jan Brezovsky (Adam Mickiewicz University in Poznan & International Institute of Molecular and Cell Biology in Warsaw).

Abstract:

Rapidly growing bacterial antibiotic resistance is defined by WHO as one of the most alarming global health issues. Due to their lethal effect on the microbiota, antibiotics exert high selective pressure that boosts resistance development. Clearly, effective alternatives to substitute or at least support antibiotics are needed. In this work, we target the bacterial communication process – quorum sensing. Degradation of microbial signaling molecules, commonly known as quorum quenching (QQ), has been shown to reduce the expression of genes controlling virulence factors, limit biofilm formation and due to its less-selective mode of action, is believed to escape from conventional resistance mechanisms. Using state-of-the-art quantum mechanics/molecular mechanics molecular dynamics simulations enriched by experimental verification, we prove QQ activity of two biotechnologically well-established enzymes – E. coli and Achromobacter spp. penicillin G acylases (PGAs).(1) Through multi-scale modeling that accounts for protein dynamics as a crucial factor of catalytic function and by comparing PGAs with prototypical acyl-homoserine lactone acylase possessing native activity towards bacterial signaling molecules, we determined common determinants responsible for QQ activity of N-terminal serine hydrolases and highlighted reasons for relatively low activity of PGAs. Building on top of these findings, we propose rationally-engineered variants of E. coli PGAs with improved activities towards bacterial signaling molecules bringing us closer to potent antimicrobial agents.

This work was supported by the National Science Centre, Poland (2017/25/B/NZ1/01307, 2021/41/N/NZ2/01365). The computations were performed at the Poznan Supercomputing and Networking Center.

1. Surpeta B., et al. ACS Catal., 6359–6374(2022)

Proteins

AmyloGraph: A comprehensive database of amyloid-amyloid interactions

Michal Burdukiewicz (Autonomous University of Barcelona), Dominik Rafacz (Warsaw University of Technology), Agnieszka Barbach (Wrocław University of Science and Technology), Katarzyna Hubicka (Wrocław University of Science and Technology), Laura Bąkała (Warsaw University of Technology), Anna Lassota (Coventry University), Jakub Stecko (Wrocław Medical University), Natalia Szymańska (Wrocław Medical University), Jakub Wojciechowski (Wrocław University of Science and Technology), Dominika Kozakiewicz (Institute of Immunology and Experimental Therapy, Polish Academy of Sciences), Natalia Szulc (Wroclaw University of Science and Technology), Jarosław Chilimoniuk (University of Wrocław), Izabela Jęśkowiak (Wrocław Medical University), Marlena Gąsior-Głogowska (Wroclaw University of Science and Technology) and Malgorzata Kotulska (Wroclaw University of Technology, Department of Biomedical Engineering and Instrumentation).

Abstract:

The interactions between amyloid proteins can induce the self-assembly or, on the opposite, completely inhibit this process. Currently, we are still very far from discovering the rules underlying such interaction, as the knowledge on that topic is scattered among many publications and often contradictory. Moreover, there is a lack of well-defined terminology that could describe such phenomena and the existing definitions do not cover all possible outcomes.

Therefore, we propose AmyloGraph (https://www.amylograph.com/), the first database to collect information on interactions between amyloid proteins. To systematize the information, we have designed a system of three descriptors that catch the most critical features of interactions between two amyloid proteins. The manually curated data was independently validated and consulted with the authors of the original publications. The in-depth description of the validation process, along with its statistics, is available in the documentation (https://kotulskalab.github.io/AmyloGraph/articles/definitions.html).

AmyloGraph represents its contents in the form of an interactive graph and table. Aside from the three descriptors of interaction, it also contains the sequences of both participants. The information is downloadable from the database. Moreover, users can install AmyloGraph as the R package (https://github.com/KotulskaLab/AmyloGraph) and run it locally.

AmyloGraph contains the unique data from 173 manuscripts on 883 interactions between 47 amyloid proteins. Its usefulness applies to the research on the fundaments of amyloid-amyloid interactions, co-occurrence of amyloidosis, amyloid propagation and inhibitors of the self-assembly.

Proteins

Analysis of fragment screening datasets in context with Human genetic variation

Javier Sánchez Utgés (University of Dundee), Callum Ives (University of Dundee), Stuart MacGowan (University of Dundee) and Geoff Barton (University of Dundee).

Abstract:

Fragment screening takes a library of small molecules (fragments) and tests them for binding to a protein by X-ray crystallography. Any bound fragments found may provide a lead in the drug discovery pipeline. While fragments often bind at known active sites, many bind at uncharacterised sites. Here, we seek to understand and rank all binding sites by combining structural, evolutionary and human population data. We introduce a novel approach for binding site definition from protein-ligand interaction fingerprints rather than clustering ligands, and for the first time, examine the distribution of human population variants from large-scale sequencing of healthy individuals across binding sites. Population variants are not randomly distributed along the genome but are constrained by protein structure and function. Fragment screening data were analysed for 35 different experiments, accounting for a total of 1309 protein structures binding to 1601 ligands. The 291 defined binding sites (2664 residues) were characterised by amino acid conservation, composition, enrichment in variation and surface accessibility. Preliminary results indicate, as expected, that the most conserved sites across homologues and depleted in missense variation in human are the known catalytic and active sites of the target proteins. Our analysis also suggests that larger binding sites tend to be more conserved across homologues, less enriched in variants and less accessible to solvent. These findings will be of interest to those studying protein-ligand interactions or developing new drugs.

Proteins

Annotating the regeneration transcriptome of Cloeon dipterum.

Patricia Medina-Burgos (CABD-CSIC), Israel Barrios-Núñez (CABD-CSIC), Fernando Casares (CABD-CSIC) and Ana Rojas (CABD-CSIC).

Abstract:

To recover function, damaged or amputated tissues/organs can regenerate in certain animals. Understanding this process requires investigating the mechanisms controlling regeneration in the widest range of organisms. The mayfly Cloeon dipterum is an emerging model insect to study this phenomenon. Using this model, we analysed the transcriptional response of Cleoon's gills during their regeneration, which is completed in just a few days, by focusing on the annotation of differentially expressed genes at different points of the process. A substantial part of these genes showed no apparent homology after running standard comparative genomics approaches. Here we present a computational pipeline devised to annotate genes which cannot be annotated due to lack of homologues elsewhere. With extensive use of technologies based on Deep Learning, we will describe a novel and reliable procedure to improve the functional annotation of regeneration related genes.

Proteins

Annotation of biologically relevant ligands in UniProtKB using ChEBI

Elisabeth Coudert (SIB Swiss Institute of Bioinformatics), Sebastien Gehant (SIB Swiss Institute of Bioinformatics), Edouard de Castro (SIB Swiss Institute of Bioinformatics), Monica Pozzato (SIB Swiss Institute of Bioinformatics), Christian Sigrist (SIB Swiss Institute of Bioinformatics), Delphine Baratin (SIB Swiss Institute of Bioinformatics), Teresa Neto (SIB Swiss Institute of Bioinformatics), Nicole Redaschi (SIB Swiss Institute of Bioinformatics) and Alan Bridge (SIB Swiss Institute of Bioinformatics).

Abstract:

The UniProt Knowledgebase (UniProtKB, at www.uniprot.org) is a reference resource of protein sequences and functional annotation that covers over 200 million protein sequences from all branches of the tree of life.

UniProtKB provides a wealth of information on protein sequences and their functions, including descriptions of the nature and binding sites of biologically relevant ligands (also known as cognate ligands) such as activators, inhibitors, cofactors, and substrates. UniProtKB captures this information through expert literature curation and from experimentally resolved protein structures in the Protein Data Bank (PDB/PDBe), filtering ligands that are technical artefacts and mapping observed ligands to cognate ligands.

In this work, we describe improvements to the representation of cognate ligands in UniProtKB using the chemical ontology ChEBI (www.ebi.ac.uk/chebi/). We have performed a complete reannotation of all binding sites for cognate ligands in UniProtKB – replacing textual descriptions of defined ligands with stable unique identifiers from the ChEBI ontology, covering over 750 distinct cognate ligands and almost 200,000 UniProtKB/Swiss-Prot entries – and now use ChEBI as the reference vocabulary for all new ligand annotations. This enhanced dataset will provide improved support for efforts to study, and predict, functionally relevant interactions between proteins and their cognate ligands. We will describe new search and query facilities with which users can mine it using the chemical ontology and chemical structure data provided by ChEBI via the UniProt website, REST API, and SPARQL endpoint.

Proteins

APPRIS Principal Isoforms and MANE Select Transcripts Define Reference Splice Variants

Fernando Pozo (CNIO), Laura Martinez Gomez (CNIO), Jose Manuel Rodriguez (CNIC), Jesús Vázquez (CNIC) and Michael Tress (CNIO).

Abstract:

Ensembl/GENCODE and RefSeq have collaborated to produce MANE, a reference transcript set for the human genome. There is a single MANE Select reference transcript per coding gene. These splice variants are annotated in both the RefSeq and Ensembl/GENCODE reference gene sets.

MANE Select transcripts add to the list of tools for selecting a single variant per coding gene (APPRIS principal isoforms, UniProt display variants). But does a single reference transcript per gene make sense from a biological point of view, or is a single reference isoform/transcript per gene an over-simplification, or even “dangerous”, as some believe?

Here we compared MANE Select transcripts to other reference splice variant prediction methods, using data from large-scale proteomics experiments and human genetic variation studies. There is overwhelming support for a single main protein isoform for most coding genes. The best method for determining these reference splice variants? MANE Select transcripts and APPRIS principal isoforms.

The two methods are particularly powerful when their predictions agree (94.2% of coding genes). In fact, in 98.2% of these genes, the main isoform detected in proteomics experiments, the MANE Select transcripts, and the APPRIS principal isoforms all agreed. Germline variant rates showed that exons unique to MANE Select and APPRIS principal transcripts are subject to purifying selection, while exons unique to alternative transcripts evolve neutrally. By contrast, choosing the longest splice variant as the representative is a poor strategy because exons unique to the most extended splice variants are not under selective pressure and are unlikely to be functionally relevant.

Proteins

ATHENA: Analysis of Tumor Heterogeneity from Spatial Omics Measurements

Adriano Martinelli (IBM Research Zurich), Pushpak Pati (IBM Research Zurich) and Maria Anna Rapsomaniki (IBM Research Zurich).

Abstract:

Tumor heterogeneity has emerged as a fundamental property of most human cancers, and its accurate and biologically meaningful quantification has the potential to translate biological complexity into clinically actionable insight. Currently, spatial omics technologies are revolutionizing our understanding of tumor ecosystems, enabling their deep phenotypic profiling at an unprecedented resolution while preserving the tumor topology. Although several spatial omics data analysis tools have started to emerge, a dedicated resource that enables tumor heterogeneity quantification is largely missing. We introduce here ATHENA, a computational framework that brings together a large collection of established and novel heterogeneity scores borrowing ideas from spatial statistics, graph theory and information theory, able to capture the heterogeneity of the tumor ecosystem. At the core of ATHENA resides a graph representation of the tissue that models different levels of cell-cell interactions. ATHENA supports any spatial omic dataset, as well as standard tissue imaging data and implements a new concept of local heterogeneity scores, computed at a single-cell level using the graph topology that capture spatial heterogeneity. Using a publicly available imaging mass cytometry dataset, we show how ATHENA can highlight tumor regions of high spatial heterogeneity and quantify spatial properties, cell interaction and immune infiltration patterns present in the tumor ecosystem. ATHENA is implemented in a highly modular, extendable, and scalable fashion, with emphasis in visualization and interoperability with other popular computational frameworks, and it's available as a Python package under an open-source license here: https://github.com/AI4SCR/ATHENA.

Proteins

ATMision: a web portal for the in silico annotation of ATM missense variants beyond pathogenicity predictions

Natàlia Padilla Sirera (VHIR), Luz Marina Porras (VHIR), Alejandro Moles-Fernandez (VHIO), Lidia Feliubadaló (ICO), Marta Santamariña-Pena (FPGMX), Alysson T. Sánchez (IDIBELL), Anael López-Novo (USC), Ana Blanco (FPGMX), Miguel de la Hoya (IdISSC), Ignacio J. Molina (UGR), Ana Osorio (CNIO), Marta Pineda (ICO), Daniel Rueda (Hospital 12 de Octubre), Clara Ruiz-Ponte (SERGAS), Ana Vega (FPGMX), Conxi Lázaro (IDIBELL), Orland Díez (VHIO), Sara Gutiérrez-Enríquez (VHIO) and Xavier de la Cruz (VHIR).

Abstract:

The ataxia-telangiectasia mutated (ATM) gene encodes a serine/threonine kinase essential in the detection and signaling to repair DNA double-strand breaks. Monoallelic pathogenic germline variants in ATM increase the risk of cancer, particularly breast cancer, but also prostate and pancreatic cancer.

Identification of carriers of ATM disease-causing variants offers patients and families a precise clinical management based on personalized prevention. To have this clinical benefit, it is of paramount relevance to accurately assess the deleterious effect of sequence variants on ATM protein function. Presently, standard in silico tools for pathogenicity prediction can contribute to this goal but they are incompletely accurate and their results are difficult to interpret.

In this work, we present our website ATMision (http://biotoclin.org/ATMision) that provides interested users with a family of competitive predictors for ATM missense variants and a series of visual tools for post-prediction analysis. Apart from their application as trust-generation wrappers for the predictors, these graphical tools have other applications. They can be accessed independently from the predictors, to study the behavior of groups of variants, looking for trends of interest. Here, we illustrate this application in the study and prioritization of ATM VUS. We also describe how these tools allow a graphical comparison of the two ATM-adapted versions of the ACMG/AMP guidelines.

Proteins

Automated System for Mechanistic Analysis and Interpretation of Genetic Variants

Ana C. González-Álvarez (King Abdullah University of Science and Technology), Francisco J. Guzmán-Vega (King Abdullah University of Science and Technology), Kelly J. Cardona-Londoño (King Abdullah University of Science and Technology), Karla A. Peña-Guerra (King Abdullah University of Science and Technology) and Stefan T. Arold (King Abdullah University of Science and Technology).

Abstract:

As sequencing has become cheap and fast, clinicians can rapidly obtain genetic and phenotypic data from their patients. However, the analysis of all data obtained may become a major bottleneck for delivering personalized medicine in time. Genetic variants can affect the function of proteins by various mechanisms including changes in a protein's stability, catalytic activity, autoregulation, subcellular targeting, and associations with other biomolecules or drugs. The correct identification of the disease-causing mechanisms is crucial to identify the best therapeutic approach. To facilitate this step, we are establishing a computational pipeline, implemented in Python, for the automated analysis of patient's genetic mutations. The particular strength of this platform is to analyze patient variants in the context of the protein's three-dimensional structure. For this analysis, the computational platform uses experimental or AlphaFold-predicted structures and combines various in silico tools to infer the molecular mechanism by which a mutation affects the protein's function, taking into account many sources of existing and predicted data including domain and active site information, intra and inter-molecular interactions, amino acid conservation, among others. This methodology has been applied to several cases to evaluate the disease mechanism of patient variants. The approach can be extended for large-scale variant analysis in existing databases, to devise a "structural landscape" for different genetic diseases. Once the full implementation is complete and available to the public, the integrated tool could help clinicians to achieve tailored patient support and contribute to the better understanding of genetic disease.

Proteins

Benchmarking of current conformational B-cell epitope prediction methods

Gabriel Cia (Université Libre de Bruxelles), Fabrizio Pucci (Université Libre de Bruxelles) and Marianne Rooman (Université Libre de Bruxelles).

Abstract:

The accurate prediction of conformational B-cell epitopes, i.e. antigen surface regions bound by antibodies, would lead to major improvements in disease diagnostics, drug design and vaccine development. A variety of computational methods, mainly based on machine learning approaches, have been developed in the last decades to tackle this challenging problem. Here, we rigorously assessed the performances of nine state-of-the-art predictors. Surprisingly, the results of our benchmarking and statistical analyses on a dataset of over 250 antibody-antigen structures show that none of the methods perform significantly better than the trivial procedure of randomly generated patches of surface residues. We further show that commonly used consensus strategies that combine the predictions from multiple webservers are also no better than random. In addition, we found that the scores predicted by each method correlate very poorly with immunodominance (all methods have < 0.1 Spearman correlation). Finally, we applied all the predictors to a case study consisting of the recent SARS-CoV-2 S1 surface protein that largely recapitulates our benchmarking conclusions, with only one method performing slightly better than random. We hope that these results will serve as a wake-up call to identify the biases and issues that limit current methods, favor the adoption of state-of-the-art evaluation methodologies in future publications, and suggest new strategies to improve the performance of conformational B-cell epitope prediction methods.

Proteins

Biases and generalizability in predictions of protein-protein binding affinity changes upon mutations

Matsvei Tsishyn (Université Libre de Bruxelles), Marianne Rooman (Université Libre de Bruxelles) and Fabrizio Pucci (Université Libre de Bruxelles).

Abstract:

In the last decades, bioinformatics tools to predict protein-protein affinity change upon point mutations brought new insight into multiple critical biological topics such as protein rational design, phenotype-genotype relations and genetic diseases by allowing in silico proteome-scale mutagenesis experiments. While all those tools are submitted to strict cross-validation processes, they suffer from biases toward learning datasets. Indeed, the golden standard of affinity-change experimental data, SKEMPI-2, is literature-based and thus is highly unbalanced in terms of protein families, spatial location of mutations, binding-sites and types of mutated amino acid. For example, more than a half of mutations are toward alanine. These imbalances lead to overoptimistic performance measures and raise severe concerns about the generalizability of predictors. We analyzed and quantified the unbalances present in the SKEMPI-2 dataset as well as its intrinsic noise (caused by experimental noise, curation errors and variation in experimental conditions). We showed that the performances of the predictors largely depend on the sub-group of mutation that we consider and that they tend to overestimate destabilizing mutations by violating the symmetry principle (affinity change of A to B is the negative of affinity change of B to A). Finally we proposed cross-validation, architecture and dataset sampling schemes in order to build more generalizable and less biased predictors as well as a highly-interpretable linear predictor based on only three simple sequence- and structure-derived features.

Proteins

CanProSite: Predicting potential residues associated with lung cancer using deep neural network

Medha Pandey (IIT MADRAS) and M. Michael Gromiha (IIT MADRAS).

Abstract:

Lung cancer is a prominent type of cancer, which leads to high mortality rate worldwide. The major lung cancers lung adenocarcinoma (LUAD) and lung squamous carcinoma (LUSC) occur mainly due to somatic driver mutations in proteins and screening of such mutations is often cost and time intensive. Hence, in the present study, we systematically analyzed the preferred residues, residues pairs and motifs of 4172 disease prone sites in 195 proteins and compared with 4137 neutral sites. We observed that the motifs LG, QF and TST are preferred in disease prone sites whereas GK, KA and ISL are predominant in neutral sites. In addition, Gly, Asp, Glu, Gln and Trp are preferred in disease prone sites whereas, Ile, Val, Lys, Asn and Phe are preferred in neutral sites. Further, utilizing deep neural networks, we have developed CanProSite for predicting disease prone sites with amino acid sequence based features such as physicochemical properties, conservation scores, secondary structure and di and tri-peptide motifs. The model is able to predict the disease prone sites at an accuracy of 81 % with sensitivity, specificity and AUC of 82 %, 78 % and 0.91, respectively, on 10-fold cross-validation. When the model was tested with a set of 417 disease-causing and 413 neutral sites, we obtained an accuracy and AUC of 80 % and 0.89, respectively. We suggest that our method can serve as an effective method to identify the disease causing and neutral sites in lung cancer.

Proteins

Characterization of rare and novel AlphaFold structural space

Janani Durairaj (Biozentrum, University of Basel; SIB Swiss Institute of Bioinformatics), Mehmet Akdel (Wageningen University and Research), Pedro Beltrão (ETH Zurich) and Torsten Schwede (Biozentrum, University of Basel; SIB Swiss Institute of Bioinformatics).

Abstract:

The recent and ongoing releases of millions of highly accurate, computationally predicted, structures as part of the AlphaFold Protein Structure Database has greatly expanded structural coverage of proteomes across species and function. We perform a comparison of structural elements between proteins from the AlphaFold database and experimentally characterised structures from the PDB using “shape-mers”, analogous to sequence k-mers, defined by rotation-invariant moment functions. We then apply natural language processing approaches to these shape-mers to explore the global protein structural space.

First, we use topic modelling, an unsupervised approach to discover abstract “topics” in structural space. We focus on high confidence predictions of structural elements (combinations of shape-mers) that are absent or very rare in experimentally characterised structures. We pinpoint areas of protein functional and evolutionary space now expanded by AlphaFold as well as structural regions which could represent novel folds.

We then applied Word2vec embedding to find shape-mer associations across structural and functional domains. This allows for defining semantic similarity between shape-mers, i.e structural fragments that appear in similar contexts, thus pinpointing very remote homology impossible to see from sequence alone.

The shape-mer framework enables fast comparison of millions of structures while also defining connected structural elements spanning across a protein which, analogous to sequence motifs, can then be linked to protein function.

Proteins

Characterizing and explaining impact of disease-associated mutations in proteins without known structures or structural homologues

Neeladri Sen (UCL), Ivan Anishchanka (Institute for Protein Design, UW), Nicola Bordin (UCL), Ian Sillitoe (UCL), Sameer Velankar (EMBL-EBI), David Baker (Institute for Protein Design, UW) and Christine Orengo (UCL).

Abstract:

The structure of proteins can help understand the mechanism of diseases associated with missense mutations and help develop therapeutics. With improved deep learning techniques such as RoseTTAFold and AlphaFold, we can predict the structure of proteins even in the absence of structural homologues. We modelled and extracted the domains from 553 disease-associated human proteins without known protein structures or sequential homologues in the Protein Databank. Domains that could be assigned to CATH superfamilies had higher quality and lower RMSD between AlphaFold and RoseTTAFold models compared to those that could only be assigned to Pfam or neither. Using these models, we predicted ligand-binding sites, protein-protein interfaces, conserved residues, destabilising effects, and pathogenicity caused by missense mutations. We could explain 80% of these disease-associated mutations based on proximity to functional sites, structural destabilization, or pathogenicity. These mutations were more buried, pathogenic, closer to predicted functional sites and had higher predicted ddG of mutation compared to polymorphisms. Usage of models from the two state-of-the-art techniques and multiple predictors predicting the same mutation to have an effect provides higher confidence in our predictions. We explain 93 additional mutations based on RoseTTAFold models which could not be explained based solely on AlphaFold models.

Proteins

COCOMAPS-2: an improved web tool for characterizing the interface of protein-protein and protein-nucleic acids complexes

Tiziana Ricciardelli (King Abdullah University of Science and Technology), Mohit Chawla (King Abdullah University of Science and Technology), Luigi Cavallo (King Abdullah University of Science and Technology) and Romina Oliva (University Parthenope of Naples).

Abstract:

Over ten years ago we proposed COCOMAPS (bioCOmplex COntacts MAPS), a web tool available at: https://www.molnac.unisa.it/BioTools/cocomaps (1), which provides a thorough characterization of the interface in macromolecular complexes. The COCOMAPS hallmark is visualizing interfaces as 2D intermolecular contact maps (where a dot at the cross-over of two residues represents a contact), as sort of fingerprints of the interactions. In addition, the COCOMAPS output includes lists of: interacting residues (defined on the basis of a adjustable cut-off distance), residues at the interface (defined on the basis of the buried surface upon complex formation) and intermolecular H-bonds, as well as ready-to-run PyMOL scripts to generate an efficient visualization of the interface (see Figure).

Characterization of the interface by COCOMAPS can be considered complete at level of amino acid or nucleotide residues, making it indeed one of the users' preferred tools for the purpose. However, it may lack of a deeper understanding of the physico-chemical nature of the contacts themselves. To fill this gap, we are now complementing COCOMAPS with novel features aimed at a detailed understanding of the interactions at an atomic level. Integrated with a complete characterization of the interface in terms of: π-π, cation-π, anion-π, lone pair-π and sulphur-π stacking contacts, along with halogen bonds and water- or ion-mediated contacts at the interface, also reported in easy-to-read contacts maps, COCOMAPS-2 will be soon available to the community as a free server, proposing itself as one-stop tool.

1. Vangone A et al (2011) Bioinformatics 27:2915.

Proteins

Comparative clustering of eukaryote complexomes identifies novel taxon-specific protein complexes and interactors

Joeri van Strien (Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands), Felix Evers (Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, Netherlands), Madhurya Lutikurti (Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands), Alfredo Cabrera Orefice (Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands), Ulrich Brandt (Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands), Taco W.A. Kooij (Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, Netherlands) and Martijn Huynen (Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands).

Abstract:

Complexome profiling is a powerful omics approach to systematically interrogate the presence and composition of protein complexes in a cell or organelle. In recent years a number of complexome profiling datasets, from a variety of species, have become available. Large-scale comparative analysis of complexome profiles holds the potential to identify novel conserved interactions as well as to characterize differences between complexomes. However, as of yet no approaches exist that leverage complexome profiling data to allow systematic characterization of not just evolutionary conserved, but also taxon-specific elements of the complexome. Therefore, we present comparative clustering (CompaCt). Our method combines and compares interaction datasets from multiple species by using an integrative clustering approach. The resulting “super-clusters'' represent groups of interacting proteins, and consist of one or more subclusters each representing one species. Our novel method of integrating data from multiple species allows identification of both conserved as well as taxon-specific interactors. CompaCt was applied to a collection of 51 complexome profiling datasets from 9 eukaryote species. We are able to recover over 50 known protein complexes, the majority of which are represented in multiple species. We demonstrate that combined clustering of multiple complexomes outperforms separate clustering of individual complexomes. Furthermore, we pinpoint many novel candidate interactors and complexes in multiple species. Among these we identify a novel animal-specific candidate interacting with a mitochondrial oxidative phosphorylation complex, and a novel candidate human protein complex, related to the emp24 complex in Yarrowia lipolytica.

Proteins

Computational approach for the recognition of small molecule inhibitors for Toll-like receptor

Shailya Verma (National Centre for Biological Sciences, TIFR) and Ramanathan Sowdhamini (National Centre for Biological Sciences, TIFR).

Abstract:

Toll-like receptors (TLRs) are pattern recognition receptors present on the surface of cell playing crucial role in innate immunity. One of the TLRs, TLR4, recognizes LPS (Lipopolysaccharide) as ligand leading to the release of anti-inflammatory mediators as well as pro-inflammatory cytokine through signal transduction and domain recruitment. TLR4 forms homodimer at its intracellular TIR (Toll/interleukin-1 receptor) domain. This homodimer of TLR4 helps in the recruitment of TRAM/TICAM2 (TIR domain-containing adaptor molecule 2) molecule. TRAM contains its own TIR domain which in turn, dimerises and functions as an adapter protein to further recruit TRIF/TICAM1 (TIR domain-containing adaptor molecule 1) protein for mediating downstream signaling. Apart from LPS, TLR4 also recognizes endogenous ligands like fibrinogen, HMGB1 and hyaluronan in autoimmune conditions and sepsis. The aim of the study is to attenuate the signalling of TLR4-TRAM-TRIF cascade in these auto inflammatory situations. Our computational approach is to target TRAM, recognise small molecule inhibitors for TRAM.

We performed Homology modelling, Molecular docking, Ligand Clustering and Molecular Dynamics to identify 10 lead compounds from small molecules of natural origin, as contained in Super Natural II database. Apart from this, we performed cell-based reporter assay for studying the stimulation of TLR4 by monitoring the inhibition of NF-κB and AP-1 in presence of small molecule inhibitors. We also purified the TRAM protein and looked at the small molecule interaction using NMR. Through our study we identified lead compounds, that abrogate the downstream signalling of TLR4, which can be helpful in autoimmune conditions.

Proteins

Computational approach to identify multifunction by metamorphism in proteins

Israel Barrios Núñez (CABD-CSIC) and Ana Rojas M. (CABD-CSIC).

Abstract:

The hypothesis Sequence determines Structure determines Function (SSF), excludes the possibility of a protein to exhibit more than one physiological function. We know now that multifunction in proteins is not rare, a phenomena which has been amply documented (REFS) in different forms like IDP, Moonlighting, and Metamorphism. The SSF hypothesis is the basis of computational prediction of function, based mostly on transferring functions based on sequence similarity. When the same peptide acquires alternative and very different structural arrangements according to precise functions, multi-functionality is then

achieved by metamorphism. This event is relevant as inventing novel genes may not be essential to acquire novel functions.

Here we present the difficulties to predict the structure of certain metamorphs and provide some insights to computationally approach this task, which include the use of ML approaches.

Proteins

Computational function prediction in UniProt & AI/ML community engagement

Vishal Joshi (EMBL-EBI), Hermann Zellner (EMBL-EBI) and Maria Martin (EMBL-EBI).

Abstract:

Purpose of Automatic Annotation

Currently, manually reviewed records constitute only about 1% of UniProtKB; expert curation is time-intensive and most published experimental data focuses on a rather limited range of model organisms. Simultaneously, the number of unreviewed records is growing continuously, yet for a large proportion of these records there is no experimental data available. Currently UniProt uses two prediction systems UniRule and Association-Rule-Based Annotator (ARBA) to functionally annotate the unreviewed records automatically.

Metal Binding Site challenge

Artificial Intelligence and Machine Learning (AI/ML) are making giant strides in many fields of science and technology. We would like to work with computational biologists on methods for predicting functional and site annotations that are implementable by UniProt. For our first challenge to the AI/ML community, we are asking participants to predict protein metal binding sites. We aim to identify one or more software tools that are both accurate and scalable and can be applied within the UniProt production environment. It is due to commence in June 2022.

Protein Embeddings

Protein sequence embeddings are representations of protein sequences by a vector that can be used as input for machine learning models. Computation of protein embeddings can be computationally costly on a larger set of protein sequences. In order to reduce the redundant development efforts, we consider to provide sequence embeddings for proteins in UniProtKB to the AI/ML community.

We are interested in discussing these developments with interested groups in this community.

Proteins

Computational prediction of epitope-specific paratopes using convolutional neural networks

Dong Li (Université Libre de Bruxelles), Marianne Rooman (Université Libre de Bruxelles) and Fabrizio Pucci (Université Libre de Bruxelles).

Abstract:

Antibodies play a central role in the adaptive immune response of vertebrates through the specific recognition of exogenous or endogenous antigens. The rational design of antibodies has a wide range of biotechnological and medical applications such as disease diagnosis and treatment. However, there are currently no reliable methods to predict the paratopes (i.e. antibody's binding region) that recognize a specific epitope (i.e. antigen's binding region) and conversely, epitopes that recognize a given paratope. To fill this gap, we developed a machine learning-based tool for predicting paratopes able to bind epitopes of a given antigen. In a first step, the most important features of epitope-paratope binding were identified, such as the aromatic and charged character of the interacting residues. The epitope and the paratope patches were simplified into interacting two-dimensional patches, colored according to the values of the selected features, and pixelated. The specific recognition of an epitope image by a paratope image was achieved by using a convolutional neural network-based model which was trained on a set of paratope-epitope two-dimensional images derived from experimental structures of antibody-antigen complexes. Our method achieves very good performances in cross validation with an accuracy of 0.86, and is also able to recognize homologous epitope-paratope pairs. As a particular example, we successfully applied our method to the prediction of paratopes towards epitopes of the SARS-CoV spike protein.

Proteins

Conformine, a predictor of protein Conformational Variability from amino acid sequence

Jose Gavalda-Garcia (Vrije Universiteit Brussels) and Wim Vranken (Vrije Universiteit Brussel).

Abstract:

Proteins are dynamic and can change conformation over time. We introduce a knowledge-based metric to describe conformational regions for the protein backbone at the residue level, including a means to quantify how often an amino acid residue moves between these regions. This metric provides complementary characterisation of the protein in addition to other protein dynamics metrics, such as Random Coil Index. To calculate this metric (Conformational Variability), we performed Molecular Dynamics simulations of 100 proteins with diverse levels of disorder, for which each amino acid was assigned to one of five secondary structure categories for every sample time in the simulation. Then, the Conformational Variability was determined from how often changes in this secondary structure category occurred, calculated with the Frobenius matrix norm.

We then trained an estimator to predict Conformational Variability from amino acid sequence only. This estimator consists of a Long Short-Term Memory neural network which predicts Conformational Variability in multitask with synergetic problems: secondary structure propensity and ShiftCrypt index [1]. We trained on 118 sequences for the main task and secondary structure propensities and on 4500 sequences for the ShiftCrypt index. We performed a 5-fold cross-validation, which indicated a Pearson's correlation for the main task above 0.5 and its corresponding p-value near 0.

We present this estimator under the name ConforMine, a predictor of protein Conformational Variability from amino acid sequence. This tool will be incorporated in our tool suite "b2bTools", available in PyPI, Anaconda and Bioconda under the same name.

1 Orlando, G. et al. https://doi.org/10.1093/nar/gkaa391

Proteins

Conservation and evolution of Roquin-1 binding sites of known and novel targets in the T cell transcriptome

Giulia Cantini (Helmholtz Center Munich), Taku Ito-Kureha (Institute for Immunology, Biomedical Center (BMC), Faculty of Medicine, Ludwig-Maximilians-Universität Munich), Elaine H. Wong (Institute for Immunology, Biomedical Center (BMC), Faculty of Medicine, Ludwig-Maximilians-Universität Munich), Gesine Behrens (Institute for Immunology, Biomedical Center (BMC), Faculty of Medicine, Ludwig-Maximilians-Universität Munich), Lambert Moyon (Helmholtz Center Munich), Annalisa Marsico (Helmholtz Center Munich) and Vigo Heissmeyer (Institute for Immunology, Biomedical Center (BMC), Faculty of Medicine, Ludwig-Maximilians-Universität Munich).

Abstract:

Roquin-1 is an RNA-binding protein (RBP) which promotes mRNA deadenylation and degradation by binding mainly to 3' UTRs of transcripts. In T cells it represses mRNAs commonly associated with inflammatory pathways. Dysregulation of Roquin in mice or humans causes inappropriate immune activation and the development of autoimmune or autoinflammatory diseases. Mapping Roquin-1 interactions on RNA is therefore of high relevance to uncover the biological mechanisms in which Roquin is involved and identify novel therapeutic targets in immune-related diseases.

In this study we established a pipeline based on mouse iCLIP-seq data in primary T cells, from raw sequencing reads to peak calling, for the identification of a high-confidence set of Roquin-1 binding sites. Among the 6000 genes that encode Roquin-1 bound target mRNAs we confirm interactions with known targets like Nfkbid, Tnfrsf4 and Tnf and we find numerous new targets, which are also upregulated upon genetic inactivation of Roquin. Remarkably, many of these are key players of cell fate decisions in hematopoietic cells as well as T cell differentiation.

Currently, we expand our analyses to the human system and implement a framework to assess cross-species conservation of binding sites in orthologous targets on a sequence or structural basis. Interestingly, we find conserved binding and regulation of many mRNA targets involving either the conservation of the binding site(s) or fast evolution of new ones.

Our future work aims at understanding the molecular determinants of Roquin interaction with mRNA and searching for human disease-associated genetic variants overlapping with the binding sites.

Proteins

Darwin: a side-chain positioning program with electron density map constraints based on an exact optimization framework

Nadege Polette (Université Fédérale de Toulouse, ANITI, INRAE, MIAT UR 875, 31326 Toulouse, F), Mikael Grialou (Université Fédérale de Toulouse, ANITI, INRAE, MIAT UR 875, 31326 Toulouse, F) and David Allouche (Université Fédérale de Toulouse, ANITI, INRAE, MIAT UR 875, 31326 Toulouse, F).

Abstract:

Darwin is an MSL library-based program developed for the modeling of side-chain positioning (SCP) and computational protein design (CPD).

The problem formulation uses a fixed backbone; the side-chain conformations are picked from the "BEBL" conformer library. The objective function consists of a bi-criteria pairwise decomposition, composed of:

A physical energy component derived from the CHARMM36 force field.

Constraints based on the Cross-Correlation coefficient (C.C.) between model-derived electron density maps and a target density map (putatively : Synthetic, Rx, or Cryo-em)

The optimization problem thus formulated as "Cost Function Networks"(CFN) is then resolved with the toulbar2 solver. The method used is “exact”, in other words, it provides optimality proof of the output solution.

We first iteratively solved a CFN model in a learning stage (on 50 PDB) to calculate a global weighting factor on the cross-correlations constraints; thus balancing the effect of the electron density map on the physical energy and maximizing the quality of the output.

An evaluation stage, performed with synthetic maps on a benchmark set of 206 PDB -using the learned weighting factor- outperformed state-of-the-art methods such as FASPR and SCWRL4. Dihedral overlap, inter-atomic clash scores, and RMS deviation between model side chains versus the original problem were used as qualitative metrics.

These successful preliminary evaluations are encouraging for the use of Darwin for SCP application or towards computational design applications.

Proteins

DDGun: an untrained predictor of protein stability changes upon amino acid variants

Ludovica Montanucci (Cleveland Clinc), Emidio Capriotti (University of Bologna), Giovanni Birolo (University of Torino), Silvia Benevenuta (University of Torino), Corrado Pnacotti (University of Bologna), Dennis Lal (Cleveland Clinic) and Piero Fariselli (University of Torino).

Abstract:

To estimate the functional effect of single amino acid variants in proteins, it is fundamental to predict the change in the thermodynamic stability, measured as the difference in the Gibbs free energy of unfolding, between the wild-type and the variant protein (ΔΔG). Here we present the web-server of the DDGun method, which was previously developed for the ΔΔG prediction upon amino acid variants. DDGun is an untrained method based on basic features derived from evolutionary information. Despite being untrained, DDGun reaches prediction performances comparable to those of trained methods. Here we make DDGun available as a web server. For the web server version, we updated the protein sequence database used for the computation of the evolutionary features and we compiled two new data sets of protein variants to do a blind test of its performances. On these blind data sets of single and multiple site variants DDGun confirms its prediction performance, reaching an average correlation coefficient between experimental and predicted ΔΔG of 0.45 and 0.49 for the sequence-based and structure-based version, respectively. Besides being used for the prediction of ΔΔG, we suggest that DDGun should be adopted as a benchmark method to assess the predictive capabilities of newly developed methods.

Proteins

De novo proteins from rice have a potential for forming structured entities

Francisco J. Guzmán-Vega (King Abdullah University of Science and Technology), Yuanmin Zheng (King Abdullah University of Science and Technology), Afaque A. Momin (King Abdullah University of Science and Technology) and Stefan T. Arold (King Abdullah University of Science and Technology).

Abstract:

The evolution of new proteins is commonly thought to rely almost exclusively on the duplication and subsequent specialization of existing genes, benefiting from a vast library of complex protein folds and their functions. However, it has emerged that a significant number of new genes originate de novo from parts of the genome that were previously non-coding. De novo genes have been detected in several plant and animal species and were shown to support tumorigenesis, suggesting that they play a role in allowing rapid adaptation of cells and organisms. It is however thought that these de novo proteins do not have a stable fold, severely limiting their functional range. Here we use computational and experimental methods to analyze the structural landscape of a comprehensive set of 175 de novo proteins obtained in a previous comparative genomics study of 13 related rice accessions (Zhang et al., 2019). The predicted properties and structures suggest a progression from randomly generated proteins towards canonical rice proteins, with de novos as intermediates. Unexpectedly, several protein sequences are predicted to already adopt three dimensional folds, whereas other sequences are predicted to form complex folds through self-association. Hence, our analysis suggests that complex protein folds may arise relatively easily through duplication, linkage and diversification of self-associating de novo fragments. These results are now driving experimental testing of selected sequences, which, in turn, will also provide important feedback on the performance of current state-of-the-art structure prediction tools for cases where no homologous sequences are available.

Proteins

Deciphering protein secretion from brain to CSF for biomarker discovery

Katharina Waury (Vrije Universiteit Amsterdam), Renske De Wit (Vrije Universiteit Amsterdam) and Sanne Abeln (Vrije Universiteit Amsterdam).

Abstract:

The detection of low concentration molecules in a complex matrix like cerebrospinal fluid (CSF) is still challenging and compounds the discovery of novel brain-derived fluid biomarkers. It is thus beneficial to gain a deeper understanding of the processes within the brain that lead to the secretion of proteins to the CSF. We aimed to explore if the transport of proteins from the brain to the CSF can be predicted and which factors determine this process by utilizing feature analysis and machine learning.

We curated a human CSF proteome from seven exploratory mass spectrometry studies and overlapped it with the elevated brain proteome to obtained CSF and non-CSF brain protein classes. A logistic classifier was trained on a set of 75 features to correctly distinguish between these classes. Feature analysis was performed to identify the properties most important for brain protein secretion to CSF.

The logistic classifier achieved a balanced accuracy of 73,94%. Interestingly, prediction accuracy increased up to 85.63% when including only high confidence CSF proteins that have been detected across multiple mass spectrometry studies. Feature analysis revealed the properties most important to differentiate CSF and non-CSF brain proteins: subcellular localization, presence of a signal peptide, and glycosylation sites. The trained machine learning model can be utilized to identify novel CSF biomarker candidates within the brain proteome by identifying the proteins likely to be detectable in CSF.

Proteins

Deciphering the RRM-RNA recognition code: A computational analysis

Joel Roca-Martinez (VUB), Hrishikesh Dhondge (CNRS-Loria) and Wim Vranken (VUB).

Abstract:

RNA recognition motifs (RRM) are the most prevalent class of RNA binding domains in eukaryotes. Their RNA binding preferences have been investigated for almost two decades, and even though some RRM families are now very well described, their RNA recognition code has remained elusive. An increasing number of available RRM-RNA complexes can now be analysed, and are here investigated in an in-depth computational analysis that for the first time enables the definition of the RRM recognition code for canonical RRMs. We present and validate a computational scoring method to estimate the binding between an RRM and a single stranded RNA, based on structural data from a carefully curated alignment, which can predict likely RNA binding motifs based on the RRM protein sequence. Due to the importance and prevalence of RRMs in humans and other species, this tool could help design or improve RNA binding motifs with uses in medical or synthetic biology applications, leading towards the de novo design of RRMs.

Proteins

Deep-learning protein structure predictions suggest likely molecular functions for two uncharacterised polytopic membrane proteins from the P. falciparum apicoplast

David Murphy (University of Liverpool), Daniel Rigden (University of Liverpool), Shahram Mesdaghi (University of Liverpool), Filomeno Sanchez Rodriguez (University of Liverpool), Adam Simpkin (University of Liverpool) and J Javier Burgos-Mármol (University of Liverpool).

Abstract:

Malaria is a particularly burdensome disease to humanity caused chiefly by the still poorly understood parasite Plasmodium falciparum. Much of the pathogenic success of this and related parasites is due to the presence of the apicoplast, a comparatively poorly characterised biosynthetic organelle containing many proteins of unknown function. Here we present state of the art AlphaFold2 protein structure predictions together with further in silico analyses to infer molecular functions for three transmembrane apicoplast proteins. PF3D7_0622700 and PF3D7_0908100 are shown to belong to the polytopic Major Facilitator and Cation-Proton Antiporter superfamilies respectively, confirming previous suspicions of a transporter function for both. Importantly, our computational small-molecule docking screens further suggest the parasite-essential metabolite pyridoxal 5-phosphate (vitamin B6) is transported by PF3D7_0622700 making it a potential drug target, while PF3D7_0908100 likely transports a larger negatively charged metabolite. Other analyses were unable to suggest a likely molecular function for the essential protein PF3D7_1021300, but nonetheless still present interesting protein characteristics. These findings will aid the design of experimental assays to determine what apicoplast metabolites these proteins transport. This work highlights the power of high accuracy protein structure predictions to illuminate proteins of unknown structure and function abundant in the parasite and other disease-causing microorganisms.

Proteins

Designing next-generation kinase inhibitors using machine learning of structural and chemical features

Nicholas Clark (Harvard Medical School), Ratul Chowdhury (Harvard Medical School), Clemens Hug (Harvard Medical School), Caitlin Mills (Harvard Medical School), Peter Sorger (Harvard Medical School) and Mohammed Alquraishi (Columbia University).

Abstract:

Kinase inhibitors are one of the largest and most intensively studied class of anticancer drugs and they are increasingly be pursued for other diseases as well. However, their development is often hampered by a lack of clinical efficacy and toxicity. Recent data suggest that both phenomena are related to polypharmacology – the inhibition of multiple protein targets by a single drug. Although current practice emphasizes maximizing selectivity, several recently approved selective kinase inhibitors are less efficacious than their less selective analogues. Remarkably, more selective inhibitors can also be more toxic in some cases. Uncertainty about the precise spectrum of kinases inhibited by any drug also complicates pre-clinical use of drugs as research tools.

Polypharmacology is currently an accidental byproduct of medicinal chemistry campaigns rather than a design criterion. Our goal is to fundamentally change this by using state-of-the-art machine learning (ML) methods to precisely model first, the kinase protein structures and then interactions between the whole kinome and panels of kinase inhibitors. In this project, we propose using a combination of neural networks, “transformers,” and geometric constraints to train machine learning models that predict kinase binding probability based on chemical structure, and explicit kinase sequence on our lab’s experimentally labeled drug activity and selectivity data. The approach will integrate predictive and generative modeling to design a pipeline for AI-assisted development of novel kinase inhibitors with desired binding profiles. The most promising generated molecules will be synthesized via collaboration with a lab specializing in medicine chemistry to experimentally assess our predictions.

Proteins

Development of a free interactive web portal for cytometry data gating

Robin Cohen (Department of Sciences and Technology, Haute École en Hainaut, Mons), Gabor Beke (Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava), Lubos Klucar (Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava), Dana Cholujova (Cancer Research Institute, Biomedical Research Center, Bratislava) and Jana Jakubikova (Cancer Research Institute, Biomedical Research Center, Bratislava).

Abstract:

Cytometry is a well-established method, which can be used to identify, measure and characterise cells. This method generates a large amount of data, which are then typically analysed by a process called manual gating. Gating is the subsequent selection of areas of interest on scatterplots created by using a panel of markers. The purpose of gating is to identify different cell populations. In an experiment with many parameters measured, manual gating can be a very time-consuming process. Up to now, for gating we could choose between commercially available software and a few free tools (e.g. CytoExploreR R package), but their use is not intuitive and requires some level of programming skills. We have developed a free dynamic web application using d3js JavaScript library, FlowCore R library and PHP. As a main feature, user can draw multiple types of gates (e.g. rectangle, polygon or ellipsoid), which can be modified by moving, scaling and rotating. In the future, this application could serve as an alternative to paid software or other freely available tools, which require local installation. This work was supported by grants APVV-19-0212, APVV-20-0183 and MZSR 2019/14-BMCSAV-9.

Proteins

Development of continuous, protein-specific predictors of the impact of protein sequence variants

Selen Ozkan (Vall d'Hebron Research Institute), Natàlia Padilla Sirera (Vall d'Hebron Research Institute) and Xavier de la Cruz (Vall d'Hebron Research Institute).

Abstract:

Up to date, computational models developed for predicting the effect of protein sequence variants have mostly focused on predicting the binary version effect (benign/pathogenic). However, the need to advance in line with the requirements of Personalized Medicine has created an increasing interest in producing less simplified estimates of variant impact. New research efforts are made in producing continuous estimates comparable to functional assays. Here, we present our approach to this problem following the protein-specific method. We collected deep mutational scanning experiments data available in the literature for individual proteins. We trained specific predictors for each protein in our dataset with a set of sequence- and structure (AlphaFold)-based input features to build a family of predictors of functional impact of variants. Our model performances from a stringent leave-one-position-out cross-validation method display a statistically significant predictive ability. The success rates of the protein-specific tools are comparable or better than those obtained with other tools in the literature. We also investigated whether a given protein-specific predictor can serve for several proteins. Our preliminary results show that cross-predictions may have an accuracy comparable to that of auto-predictions, opening the possibility to extend the use of protein-specific to other proteins in the clinical genome.

Proteins

Discovering novel genes in bacteria: gene function prediction using protein embeddings and synteny

Aysun Urhan (Delft University of Technology), Bianca-Maria Cosma (Delft University of Technology) and Thomas Abeel (Delft University of Technology).

Abstract:

Despite the increasing amount of genetic data available, our understanding of protein function, and gene function prediction has not been able to keep up with the advances in sequencing technology. Recently, as deep learning has become more widely used, a promising line of work emerged when researchers started to employ ideas from natural language processing such as word embeddings and language models, to protein function prediction. However, the field of protein function prediction is currently dominated by human-centric studies, and there is little research dedicated to applications in the bacterial kingdom. In this project, we address this gap by developing a method that combines protein embeddings and synteny to transfer GO annotations to novel bacterial protein sequences. At its core, our model is a nearest neighbor classifier based on similarity in the embedding space instead of sequence homology. Furthermore, we build a database of known operon structures in bacteria by mining more than 300 thousand representative bacterial genomes. We show that the nearest neighbor model on its own can outperform conventional baseline models on a set of experimentally annotated E. coli and B. subtilis proteins. Next, we demonstrate that we can further improve the prediction performance when we incorporate synteny information based on our operon database. Our work is novel in the way we exploit known biological concepts in bacteria in conjunction with new deep learning techniques. Our findings show promising results, especially for novel protein sequences with no homologs in existing databases.

Proteins

DistilProtBert: A distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts

Yaron Geffen (Bar Ilan University), Yanay Ofran (Bar Ilan University) and Ron Unger (Bar-Ilan University).

Abstract:

Recently, Deep Learning models, initially developed in the field of Natural Language Processing (NLP), were applied successfully to analyze protein sequences. A major drawback of these models is their size in terms of the number of parameters needed to be fitted and the amount of computational resources they require. Recently, "distilled" models using the concept of student and teacher networks have been widely used in NLP. Here, we adapted this concept to the problem of protein sequence analysis, by developing DistilProtBert, a distilled version of the successful ProtBert model. Implementing this approach, we reduced the size of the network and the running time by 50%, and the computational resources needed for pretraining by 98% relative to ProtBert model. Using two published tasks, we showed that the performance of the distilled model approaches that of the full model. We next tested the ability of DistilProtBert to distinguish between real and random protein sequences. The task is highly challenging if the composition is maintained on the level of singlet, doublet and triplet amino acids. Indeed, traditional machine learning algorithms have difficulties with this task. Here, we show that DistilProtBert preforms very well on singlet, doublet, and even triplet-shuffled versions of the human proteome, with AUC of 0.92, 0.91, and 0.87 respectively. Finally, we suggest that by examining the small number of false-positive classifications (i.e., shuffled sequences classified as proteins by DistilProtBert) we may be able to identify de-novo potential natural-like proteins based on random shuffling of amino acid sequences.

Proteins

Drug repurposing for identification of potential spike inhibitors for SARS-CoV-2 using molecular docking and molecular dynamics simulations

Michal Lazniewski (Warsaw University of Technology), Doni Dermawan (Warsaw University of Technology) and Dariusz Plewczynski (Warsaw University of Technology).

Abstract:

For the last two years, the COVID-19 pandemic has continued to bring consternation on most of the world. The virus continues to evolve, thus requiring both vigilance and the necessity to find and develop a variety of therapeutic treatments, including the identification of specific antiviral drugs. Multiple studies have confirmed that SARS-CoV-2 utilizes its membrane-bound spike protein to recognize human angiotensin-converting enzyme 2 (ACE2). Thus, preventing spike-ACE2 interactions is a potentially viable strategy for COVID-19 treatment.

This work aims to identify potential drugs using an in silico approach. Molecular docking was carried out on both approved drugs and substances previously tested in vivo. This step was followed by a more detailed analysis of selected ligands by molecular dynamics simulations to identify the best molecules that thwart the ability of the virus to interact with the ACE2 receptor. Because the SARS-CoV-2 virus evolves rapidly due to a plethora of immunocompromised hosts, the compounds were tested against five different known lineages. As a result, we could identify substances that work well on individual lineages and these showing broader efficacy. The most promising candidates among the currently used drugs were zafirlukast and simeprevir with an average binding affinity of -22 kcal/mol for spike proteins originating from various lineages. From among the in vivo tested substances that concurrently exhibit promising free energy of binding and ADME parameters (indicating a possible oral administration) we selected the compound BDBM50136234. In conclusion, these molecules are worth exploring further by in vitro and in vivo studies against SARS-CoV-2

Proteins

Elucidating important structural features for the binding affinity of spike - SARS-CoV-2 neutralizing antibody complexes

Divya Sharma (Indian Institute of Technology Madras) and M. Michael Gromiha (Indian Institute of Technology Madras).

Abstract:

The coronavirus disease 2019 (COVID-19) has affected the lives of millions of people around the world. In an effort to develop therapeutic interventions and control the pandemic, scientists have isolated several neutralizing antibodies against SARS-CoV-2 from the vaccinated and convalescent individuals. These antibodies can be explored further to understand SARS-CoV-2 specific antigen-antibody interactions and biophysical parameters related to binding affinity, which can be utilized to engineer more potent antibodies for current and emerging SARS-CoV-2 variants. In the present study, we used structural bioinformatics and computational biophysics to analyze the interface between spike protein of SARS-CoV-2 and neutralizing antibodies in terms of amino acid residue propensity, pair preference, and atomic interaction energy. We observed that Tyr residues containing contacts are highly preferred and energetically favorable at the interface of spike protein-antibody complexes. We have also developed a regression model to relate the experimental binding affinity for antibodies using structural features, which showed a correlation of 0.93. Moreover, several mutations at the spike protein-antibody interface were identified, which may lead to immune escape (epitope residues) and improved affinity (paratope residues) in current/emerging variants. Overall, the work provides insights into spike protein-antibody interactions, structural parameters related to binding affinity, and mutational effects on binding affinity change, which can be helpful to develop better therapeutics against COVID-19.

Proteins

Enzyme and transporter annotation in UniProtKB using Rhea and ChEBI

Lionel Breuza (SIB Swiss Institute of Bioinformatics), Lucila Aimo (SIB Swiss Institute of Bioinformatics), Ghislaine Argoud-Puy (SIB Swiss Institute of Bioinformatics), Kristian Axelsen (SIB Swiss Institute of Bioinformatics), Emmanuel Boutet (SIB Swiss Institute of Bioinformatics), Cristina Casals-Casas (SIB Swiss Institute of Bioinformatics), Elisabeth Coudert (SIB Swiss Institute of Bioinformatics), Marc Feuermann (SIB Swiss Institute of Bioinformatics), Nadine Gruaz-Gumowski (SIB Swiss Institute of Bioinformatics), Damien Lieberherr (SIB Swiss Institute of Bioinformatics), Michele Magrane (EBI), Anne Morgat (SIB, Swiss Institute of Bioinformatics), Nevila Hyka-Nouspikel (SIB Swiss Institute of Bioinformatics), Lucille Pourcel (SIB Swiss Institute of Bioinformatics), Sylvain Poux (SIB Swiss Institute of Bioinformatics), Catherine Rivoire (SIB Swiss Institute of Bioinformatics), Shyamala Sundaram (SIB Swiss Institute of Bioinformatics), Rossana Zaru (EBI) and Alan Bridge (SIB Swiss Institute of Bioinformatics).

Abstract:

The UniProt Knowledgebase (UniProtKB, at www.uniprot.org) is a reference resource of protein sequences and functional annotation. Here we describe a broad ranging biocuration effort, supported by state-of-the-art machine learning methods for literature triage, to describe enzyme and transporter chemistry in UniProtKB using Rhea, an expert curated knowledgebase of biochemical reactions (www.rhea-db.org) based on the ChEBI ontology of small molecules (www.ebi.ac.uk/chebi/). This work covers proteins from a broad range of taxonomic groups, including proteins from human, plants, fungi, and microbes, and both primary and secondary metabolites. It provides enhanced links and interoperability with other biological knowledge resources that use the ChEBI ontology and standard chemical structure descriptors, and improved support for applications such as metabolic modeling, metabolomics data analysis and integration, and the use of advanced machine learning approaches to predict enzyme function and biosynthetic and bioremediation pathways.

Proteins

E-SNPs&GO: Embedding of protein sequence and function improves the prediction of human pathogenic variants

Matteo Manfredi (Biocomputing Group, University of Bologna), Castrense Savojardo (Biocomputing Group, University of Bologna), Pier Luigi Martelli (Biocomputing Group, University of Bologna) and Rita Casadio (Biocomputing Group, University of Bologna).

Abstract:

Massive DNA sequencing technologies produce an ever-increasing number of human single-nucleotide polymorphisms occurring in protein-coding regions and possibly changing protein sequences. Discriminating harmful protein variations from neutral ones is one of the crucial challenges in precision medicine. Computational tools based on artificial intelligence provide models for protein sequence encoding, bypassing database searches for evolutionary information. We leverage the new encoding schemes for efficient annotation of protein variants. We develop E-SNPs&GO, a novel method that, given an input protein sequence and a single residue variation, can predict whether the variation is related to diseases or not. The proposed method adopts an input encoding completely based on protein language models and embedding techniques, specifically devised to encode protein sequences and GO functional annotations. The model is trained on a dataset of 101,146 human protein single residue variants in 13,661 proteins, derived from public resources. When tested on a blind set comprising 10,266 variants, our method well compares with recent approaches released in literature for the same task, reaching a MCC score of 0.72. We propose E-SNPs&GO as a suitable, efficient and accurate large-scale annotator of protein variant datasets.

Availability: The method is available as a webserver at https://esnpsandgo.biocomp.unibo.it. Datasets and predictions are available at https://esnpsandgo.biocomp.unibo.it/datasets.

Proteins

Expression regulation of protein complex partners as a compensatory mechanism in aneuploid tumors

Gokce Senger (European Institute Of Oncology), Stefano Santaguida (European Institute Of Oncology) and Martin Schaefer (European Institute Of Oncology).

Abstract:

Aneuploidy, gain or loss of chromosome or chromosomal regions, is a common feature of cancer; however, how it contributes to tumor evolution is poorly understood. In this work, we aimed to understand the global consequences of whole-chromosome-level aneuploidies on the proteome of tumor cells by integrating aneuploidy, transcriptomic and proteomic data from hundreds tumor samples from TCGA/CPTAC. We observed that a surprisingly large number of abundance changes happens on other, non-aneuploid chromosomes. Moreover, we observed a general tendency for those changes to be complex partners of proteins from aneuploid chromosomes. We showed that this co-abundance association is a compensation mechanism to prevent proteotoxicity as a consequence of aggregation of orphan proteins and imbalance in protein complex stoichiometry. On the other hand, we observed that complexes of the cellular core machinery are under functional selection to maintain their stoichiometric balance in aneuploid tumors. Ultimately, we provided evidence that those compensatory and functional maintenance mechanisms are established through post-translational control and that the degree of success of a tumor to deal with aneuploidy-induced stoichiometric imbalance impacts the activation of cellular protein degradation programs and patient survival. Taken together, our findings describe the need for compensation mechanisms to deal with the stoichiometric imbalances in protein complexes induced by aneuploidy and highlight the importance of protein complex components as potential vulnerabilities for the identification of drug targets for clinical use.

Proteins

Features derived from protein 3D structure improve prediction of variant effects in novel proteins

Alexander Gress (Helmholtz Institute for Pharmacy Saarland, University of Saarland) and Olga V. Kalinina (10.1016/j.cels.2017.11.003).

Abstract:

The rise of deep mutational scanning (DMS) experiments had a huge impact on the field of variant effect prediction with machine learning models. The results coming from DMS experiments are extremely useful, as they provide direct experimental evidence on the effect of every possible mutation in a protein on its function. Yet they have to be handled with care, since the type of effect is different for different DMS studies, and, in particular, it differs fundamentally from the other type of effect that is relevant for genetic variants, their clinical effect. Modern prediction models heavily incorporate DMS data, whether as main training data set or as an additional source of information to boost the models performance. It has been shown (e.g. with a recent method Envision, doi:10.1016/j.cels.2017.11.003) that using features that describe protein structure information improves prediction of variant effects in particular for novel proteins, not seen by the model in training. So far only comparably simple structure-based features were used, and specifically features related to protein interactions have been ignored, due to the high complexity of their generation.

In this study, we will demonstrate the usage of complex structural features significantly enhances prediction performance. We present a new method to generate very comprehensive structure-based features. Leveraging high quality protein structure models by AlphaFold makes them very broadly applicable. We used this new method to train a prediction model for variant effect prediction in novel proteins that outperforms state-of-art approaches.

Proteins

Foldseek: fast and accurate protein structure search

Michel van Kempen (Max Planck Institute), Stephanie Kim (Seoul National University), Charlotte Tumescheit (Seoul National University), Milot Mirdita (Max Planck Institute), Cameron Gilchrist (Seoul National University), Johannes Soding (Max Planck Institute) and Martin Steinegger (Seoul National University).

Abstract:

Highly accurate structure prediction methods, such as AlphaFold2 and RoseTTAFold, are generating an avalanche of publicly available protein structures. Searching through these structures with current structural alignment tools is becoming the main bottleneck in their analysis. Here we propose Foldseek a fast and sensitive protein structures alignment method to compare large structure sets. Foldseek encodes structures as sequences over a 20-state 3Di alphabet. 3Di describes discretized tertiary residue-residue interactions, which is critical for reaching high sensitivities. Foldseek's novel local alignment stage combines structural and amino acid substitution scores to improve sensitivity without sacrificing speed. It reaches sensitivities similar to state-of-the-art structural aligners while being at least 20,000 times faster. The open-source Foldseek software is available at foldseek.com and a webserver at search.foldseek.com

Proteins

FrustraEvo: Assessing Protein Families Divergence In The Light Of Sequence and Energetic Constraints

Victoria Ruiz-Serra (Barcelona Supercomputing Center), Maria Freiberger (Buenos Aires University), Camila Pontes (Barcelona Supercomputing Center), Miguel Romero (Barcelona Supercomputing Center), Pablo Galaz-Davison (Institute for Biological and Medical Engineering, Pontificia Universidad Catolica de Chile), Cesar Ramirez-Sarmiento (Institute for Biological and Medical Engineering, Pontificia Universidad Catolica de Chile), Rodrigo Gonzalo Parra (European Molecular Biology Laboratory) and Alfonso Valencia (Barcelona Supercomputing Centre BSC).

Abstract:

Protein families evolve by accumulation of sequence variations that translate into changes in the folding pathways, structure and dynamics of the native state of their members. These changes are constrained by the energy landscape features that follow the principle of minimum frustration, i.e.: energetic minimisation of those interactions that are present in their native states [1]. Although the free energy is globally minimized, native states can be in conflict with their local environment [2]. These conflicting, frustrated, signals have been linked with functional aspects [3].

We present FrustraEvo, a tool that measures local frustration conservation patterns within and between protein families as a proxy to define functionally important residues either for stability or function and relate them to their sequence variability signatures. Thanks to recent advances in structure predictions, FrustraEvo can shed light into the functional understanding of structurally characterized protein families as well as of poorly characterized ones.

[1] Bryngelson, J.D. and Wolynes, P.G. (1987) 'Spin glasses and the statistical mechanics of protein folding', Proceedings of the National Academy of Sciences of the United States of America, 84(21), pp. 7524–7528.

[2] Ferreiro, D.U. et al. (2007) 'Localizing frustration in native proteins and protein assemblies', Proceedings of the National Academy of Sciences of the United States of America, 104(50), pp. 19819–19824.

[3] Freiberger, M.I. et al. (2019) 'Local frustration around enzyme active sites', Proceedings of the National Academy of Sciences of the United States of America, 116(10), pp. 4037–4043.

Proteins

Functional Impacts of Copy Number Imbalance in Aneuploid Cancer

Elle Loughran (Institut Curie / Trinity College Dublin), Emmanuel Barillot (Institut Curie), Andrei Zinovyev (Institut Curie), Nicolas Servant (Institut Curie) and Aoife McLysaght (Trinity College Dublin).

Abstract:

Cancer karyotypes are characterised by widespread copy number alterations and aneuploidy, which can cause stoichiometric imbalances between members of protein complexes. In the germline, stoichiometric imbalances affecting dosage-sensitive genes produce deleterious effects that restrict copy number variation and contribute to the lethality of nearly all germline aneuploidies.

It is unclear how tumours tolerate such high levels of aneuploidy. Using copy number profiles of aneuploid tumours (TCGA), we have found that human germline dosage-sensitive genes are not under significant gene dosage constraint in tumours, and CNAs often leave the gene copy numbers of protein complex members out of balance. However, recent work suggests that tumours do face pressure to maintain stoichiometric balance at the protein level and that the ability to buffer CNAs via dosage compensation is associated with higher tumour fitness (Senger et al, 2022).

We thus sought to assess the impact of stoichiometric imbalance on the functioning of protein complexes in aneuploid cancer. We first evaluate stoichiometric imbalance between protein complex members in tumours and cancer cell lines using proteomics datasets from CPTAC and the CCLE. We relate the degree of stoichiometric imbalance in a protein complex to the complex's activity using a combination of activity readouts including regulon enrichment analysis (VIPER), drug sensitivity and metabolomics. Finally, we will investigate factors that affect a complex's sensitivity to stoichiometric imbalance and develop a network-based method to quantify the cellular disruption caused by aneuploidy.

Proteins

Gaussian accelerated molecular dynamics for the convergence of allosterism

Oriol Gracia i Carmona (University of natural resources and life sciences (BOKU), Wien), Franca Fraternalli (King's College London) and Chris Oostenbrink (University of natural resources and life sciences (BOKU), Wien).

Abstract:

Allosterism is the process by which a modification, namely an organic molecule binding or a mutation, that happens outside the binding site of the protein generates a change in the global behaviour of the protein. These modifications in the behaviour can be generated by big conformational shifts, or by changes in the internal motions of the proteins without any drastic change in the overall protein conformation, sometimes referred to as "dynamic allosterism". Several improvements regarding the study of allosteric pathways have been published [1]. However, despite all the improvements, these methods require long time scales to converge, hampering its applicability to drug design pipelines.

To tackle this challenge one can use enhanced sampling techniques such as Gaussian Accelerated Molecular Dynamics (GAMD). GAMD is an enhanced sampling technique that works by adding an harmonical boosting potential to lift up the energy wells, allowing to produce microseconds worth of sampling in nanoseconds time scales [2]. However, there is no knowledge on how the addition of this bias could affect these subtle internal motions. In this work we have performed an in-depth analysis of the effects of GAMD on an allosteric model system, the Pyruvate kinase M2 (PKM2), and provided a comprehensive list of advantages and caveats of different possible settings.

References

[1] MACPHERSON, Jamie A., et al. (2019), Elife, 8. Jg., S. e45068.

[2] MIAO, Yinglong; MCCAMMON, J. Andrew. (2017), Elsevier, S. 231-278.

Proteins

Highly significant improvement of protein sequence alignments with AlphaFold2

Leila Mansouri (Centre for Genomic Regulation - CRG), Athanasios Baltzis (Centre for Genomic Regulation - CRG), Suzanne Jin (Centre for Genomic Regulation - CRG), Björn E. Langer (Centre for Genomic Regulation - CRG), Ionas Erb (Centre for Genomic Regulation - CRG) and Cedric Notredame (Centre for Genomic Regulation - CRG).

Abstract:

Protein sequence alignments are essential to structural, evolutionary, and functional analysis but their accuracy is often limited by sequence similarity unless molecular structures are available [1]. Such reliance on structural information poses an issue due to the lack of experimentally validated structures. However, the solution may lay on structure prediction methods able to reach experimental grade accuracy, as achieved by AlphaFold2 [2].

In this study, we analyze the performance of protein sequence alignments using AlphaFold2 predicted structures. We find that multiple sequence alignments estimated on AlphaFold2 predictions are almost as accurate as alignments computed on experimental structures and significantly superior to sequence-based alignments even when the predicted models are of relatively low quality. Specifically, we are able to show that using AlphaFold2 models, sequence alignment accuracy increases by 24 percentage points and results in alignments in which 94% of the residues are correctly aligned as judged using experimental structural information [5].

These results suggest that, besides structure modeling, AlphaFold2 encodes higher-order dependencies that can be exploited for sequence analysis.

[1] Rost, B. Twilight zone of protein sequence alignments. Protein Eng. Des. Sel. 12, 85–94 (1999).

[2] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

[3] Thompson, J. D., Plewniak, F., & Poch, O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15(1), 87–88 (1999).

Proteins

High-throughput analyses of internal voids in biomolecules and ligand transport through them with TransportTools library

Jan Brezovsky (International Institute of Molecular and Cell Biology in Warsaw & Faculty of Biology, Adam Mickiewicz University), Carlos Eduardo Sequeiros-Borja (International Institute of Molecular and Cell Biology in Warsaw & Faculty of Biology, Adam Mickiewicz University), Aravind Selvaram Thirunavukarasu (International Institute of Molecular and Cell Biology in Warsaw & Faculty of Biology, Adam Mickiewicz University), Bartlomiej Surpeta (International Institute of Molecular and Cell Biology in Warsaw & Faculty of Biology, Adam Mickiewicz University), Nishita Mandal (International Institute of Molecular and Cell Biology in Warsaw & Faculty of Biology, Adam Mickiewicz University) and Dheeraj Kumar Sarkar (International Institute of Molecular and Cell Biology in Warsaw & Faculty of Biology, Adam Mickiewicz University).

Abstract:

Living systems contain thousands of small organic molecules that need to arrive at their sites of action to exert their function. The transport of these molecules around the cell and beyond is governed primarily by tunnels formed within the internal voids of biomolecules. Hence, the investigation of these pathways is critical to drug discovery and protein engineering efforts. Unfortuantely, these pathways are often equipped with dynamic gates, making them transient and challenging to study. The advent of high-throughput molecular dynamics simulations has enabled the study of such pathways, and their rare utilization for ligand transport events. It is not uncommon to collect datasets of thousands simulations, imposing a substantial burden on researchers in establishing the identity of tunnels observed across all simulations, determining utilization of tunnels by ligands, and developing means of specific quantitative analyses.

To alleviate these difficulties, we developed a TransportTools Python library that provides access to efficient analyses of tunnels across extensive simulations; integrated data regarding tunnels and their utilization by ligands; and means to compare transport processes under different settings, contrasting transport in original and perturbed systems. Finally, we will present insights into biological problems connected with ligand transport obtained with these tools on model proteins: (i) discovering rare yet functional water tunnels, (ii) understanding the effect of mutations on transport pathways, and (iii) studying substrate selectivity for pathways leading to an enzyme active site.

This work was supported by National Science Centre, Poland (2017/25/B/NZ1/01307). The computations were performed at Poznan Supercomputing and Networking Center.

Proteins

Identification of protein-protein interaction hubs in multiple sclerosis

Gözde Yazıcı (Bilkent University), Burcu Kurt Vatandaşlar (Istanbul Medipol University), Bilal Kerman (University of Southern California), Emre Karakoç (Sanger Institute) and Can Alkan (Bilkent University).

Abstract:

Identifying and prioritizing disease-related genes is an important scientific problem to be able to develop proper treatments. Network science has become an important discipline to prioritize such proteins/genes. In this study, we analyze multiple sclerosis (MS), which is an autoimmune disease characterized by demyelination. Demyelination is the destruction of myelin, a sheath that facilitates fast transmission of electrical impulses in nerve cells, and the cells producing it (i.e., oligodendrocytes) by immune cells. As in all autoimmune diseases, this is because the immune cells attack the organism's own cells, likely because there is a perturbation in the protein-protein interaction network of oligodendrocytes and immune cells breaking the cell-to-cell recognition. Here we analyze the regular protein-protein interaction networks of the oligodendrocyte cell and two types of immune cells, macrophage and T-cell. We investigate the most important interacting proteins, the proteins that provide the interactions between the two cells, since they are powerful candidates to cause demyelination. Importance is based on a score related to the quantity and the quality of the connections of the proteins within the intercellular and intracellular networks. We prioritize the interacting protein pairs using network analysis techniques and integer programming using a model based on the set cover with pairs problem. We show that a significant percentage of the detected proteins by our model have already been associated with MS and/or some other autoimmune and neurodegenerative diseases. Our model can be used for other autoimmune diseases or processes where interactions of two cells play an important role.

Proteins

Impact of rare genetic variants on ADME protein dynamics identified in sub-Saharan Africa: CYP3A5 as a study model

Houcemeddine Othman (Sydney Brenner Institute for Molecular Biosciencem University of the Witwatersrand) and Jorge E.B da Rocha (Sydney Brenner Institute for Molecular Biosciencem University of the Witwatersrand).

Abstract:

Pharmacogenomics studies highlighted the importance of rare variants in the genetic makeup of African populations. These variants were predicted to show a significant impact on genes involved in Absorption Distribution Metabilisation and Excretion mechanisms (ADME) of which the Cytochrome P450 superfamily are the most important ones. In terms of population prevalence, rare variants of ADME genes show high diversity among population groups of the African continent. While genome-based variant prediction tools outlined the putative functional importance of ADME gene rare variants, little is however known about their impact at the protein structural level. In this regard, we performed molecular dynamics simulations of all CYP3A5 missense variants identified in 458 individuals from different sub-Saharan African countries. CYP3A5 protein was selected as a study model to assess the functional extent of rare variants' impact. In addition, CYP3A5 is involved in the drug metabolization of ritonavir and artemether, used respectively for the treatment of HIV and malaria, two prevalent diseases in Africa. In total, we identified 5 missense variants that were simulated in addition to the reference allele and Y53C variant, which has a known deleterious impact on enzyme activity. Microscale time simulation of the CYP3A5 unbound form showed minor structural drift compared to the reference allele. Local re-arrangement however was noticed in the B/C loop and F/G loop of the drug to the interaction site leading to a significant change in the pocket volume and the tunnels controlling the accessibility of the drugs to the catalytic site.

Proteins

Improving Tandem Repeats Proteins annotation and classification in RepeatsDB

Martina Bevilacqua (University of Padova), Damiano Clementel (University of Padova), Alexander Monzon (University of Padova), Jiachen Lu (University of Padova), Paula Arrias (University of Padova) and Silvio Tosatto (University of Padova).

Abstract:

RepeatsDB is a database of structured Tandem Repeats Proteins (TRPs). Repeated units can be classified according to their shape, but their functional characterization and proper identification are still open questions. The goal of RepeatsDB is to identify the repeated units and their type on all available protein structures in the Protein Data Bank (PDB). An annotation is the association between the type of the repeat unit and its begin and end position on the protein structure. Annotations can be manually generated by a biocurator looking at the protein structure of interest (reviewed entries), or can be generated automatically through the RepeatsDB-lite predictor (unreviewed entries). RepeatsDB contains about 100,000 annotations over 7,000 PDB entries (~28% of total entries in PDB). In turn, such PDB entries are bound to just 1'500 UniProt entries. However, most of these are automatically generated. We implemented a distributed annotation framework, where anyone in the world should be able to annotate TRPs. However, we will still rely on a small group of trusted biocurators and reviewers which will grant that quality standards are met. This is helped by automatic statistical checks which would preemptively signal them any outlier or strange cases. Instead, non-trusted biocurators are rewarded through a badges-and-medals system based on gamification. The framework is scalable and can be easily applied to other classification tasks, outside the scope of RepeatsDB and TRPs classification.

Proteins

Investigation of Bacterial Fibrillar Adhesins and their Binding Characteristics

Vivian Monzon (European Bioinformatics Institute - EMBL-EBI, Cambridge) and Alex Bateman (European Bioinformatics Institute - EMBL-EBI, Cambridge).

Abstract:

Fibrillar adhesins are a class of long filamentous surface proteins, which are found in a wide range of bacterial species. These proteins can mediate the interaction of the bacterium with its environment. This includes binding other bacterial cells during biofilm formation and also host cells during host-bacteria interactions. Consequently, fibrillar adhesins play a crucial role in the initial stage of infection processes. Their characteristic filamentous structure is based on repeating protein domains, which fold into a rod-like stalk structure. An adhesive domain is posed at the tip of this stalk. Different adhesive domains are known, which are mostly protein or carbohydrate binding.

We characterised fibrillar adhesins in-depth and developed a machine learning classification approach using a range of sequence-based identification features. We applied this approach on the Firmicute and Actinobacteria UniProt Reference Proteomes and could predict a high number of potential fibrillar adhesins. But more importantly, we also detected proteins without a known adhesive domain, giving us the possibility to find novel adhesive domain families. We further investigated this putative adhesin set by predicting their structure using AlphaFold and by searching for similar structures in the PDB database. We discovered protein domains whose sequence and/or structure is similar to known adhesive domains, such as the thioester adhesive domain. We also discovered domains with novel structure folds. Studying their function and potential binding partner could enhance the understanding of bacteria host interactions as well as the development of infection prevention mechanisms.

Proteins

Lower-order Statistics Facilitate Decoy-free FDR Control in Shotgun Proteomics

Dominik Madej (The Hong Kong University of Science and Technology) and Henry Lam (The Hong Kong University of Science and Technology).

Abstract:

False discovery rate (FDR) control is a major challenge in the statistical analysis of peptide-spectrum matches (PSMs) in mass spectrometry-based shotgun proteomics. The calculation of FDR requires a null model capturing the behavior of the incorrect target PSMs. Nowadays, most analytical tools use either decoy-based or decoy-free approaches to generate the appropriate null models. However, both paradigms may lead to suboptimal FDR control due to the insufficient number of decoys available or the violation of unjustified parametric assumptions. On a deeper level, the utility of lower-scoring, incorrect target PSMs in the characterization of the top-scoring analogs remains largely unexplored.

We present a novel decoy-free framework for constructing null models for the top-scoring target PSMs using the lower-scoring counterparts and the universal transformed e-value (TEV) score. The proposed approach leverages the theoretical link between the distributions of lower-order statistics to derive an accurate parametric null model for the largest-order statistic. We present the related parameter estimation procedure supplemented with the empirical adjustment required for obtaining the optimal null model of interest. We test the proposed method on a wide range of diverse datasets. We demonstrate that, in general, the novel order-based framework controls FDR at least as accurately as the popular decoy-based and decoy-free alternatives. Evaluation of the features of the null models derived using lower-order TEV distributions sheds more light on the limitations of the parametric methods in the statistical analysis of the proteomics data.

Proteins

Machine learning applications for the classification of serial ED data

Senik Matinyan (University of Basel), Burak Demir (University of Basel) and Jan Pieter Abrahams (University of Basel).

Abstract:

An important aspect in structural biology is the improvement of the 3-dimensional models of the proteins beyond the current limits. Single molecule electron diffraction, as an alternative approach to X-ray crystallography and single particle cryo-EM, has better signal to noise ratio and potential to increase the resolution of available protein models.

This technology requires many measurements at various positions, which can lead to the congestion of data collection pipelines during long runs of experiments. In line with this, only the minority of the recorded diffraction patterns will be used for structure determination, thus making it vital to introduce new concepts of data selection criteria.

To address this question, we implemented a convolutional neural network for the classification of diffraction data.

The proposed pre-processing and training workflow could efficiently distinguish between amorphous ice and the carbon support, providing a “proof of principle” of the method to identify positions of interest. While limited in its context this approach exploits inherent characteristics of diffraction patterns of narrow electron beam and sample interaction and may be extended to allow protein data classification and feature extraction.

Proteins

MobiDB: intrinsically disordered proteins in 2022

Alexander Monzon (BioComputing UP - University of Padova), Damiano Piovesan (BioComputing UP - University of Padova) and Silvio Tosatto (University of Padova).

Abstract:

The MobiDB database (URL: https://mobidb.org/) [1] provides predictions and annotations for intrinsically disordered proteins. Latest update of MobiDB (version 4) includes novel types of annotations and an improved update process. The new website has been re-designed, with a new user interface, a more effective search engine and advanced API for programmatic access. The new database schema gives more flexibility for the users, as well as simplifying the maintenance and updates. In addition, the new entry page provides more visualisation tools including customizable feature viewer and graphs of the residue contact maps. MobiDB v4 annotates the binding modes of disordered proteins, whether they undergo disorder-to-order transitions or remain disordered in the bound state. In addition, disordered regions undergoing liquid-liquid phase separation or post-translational modifications are defined. The integrated information is presented in a simplified interface, which enables faster searches and allows large customized datasets to be downloaded in TSV, Fasta or JSON formats. An alternative advanced interface allows users to drill deeper into features of interest. A new statistics page provides information at database and proteome levels. The new MobiDB version presents state-of-the-art knowledge on disordered proteins, disordered/ordered regions are assigned based on AlphaFold pLDDT score and improves data accessibility for both computational and experimental users.

Proteins

ModCRE: a structure homology-modeling approach to predict TF binding in cis-regulatory elements

Baldo Oliva (Universitat Pompeu Fabra), Oriol Fornes (University of British Columbia), Alberto Meseguer (Universitat Pompeu Fabra), Joaquim Aguirre-Plans (Universitat Pompeu Fabra), Patrick Gohl (UNIVERSITAT POMPEU FABRA), Patricia-Mirela Bota (UNIVERSITAT POMPEU FABRA), Ruben Molina-Fernandez (Universitat Pompeu Fabra), Altair Chinchilla(Universitat Pompeu Fabra), Ferran Pegenaute (Universitat Pompeu Fabra), Oriol Gallego (Universitat Pompeu Fabra), Narcis Fernandez-Fuentes (Aberystwith University) and Jaume Bonet (Ecole Polytechnique Federale de Lausanne).

Abstract:

Transcription factor (TF) binding is a key component of genomic regulation. There are numerous high-throughput experimental methods to characterize TF-DNA binding specificities. Their application, however, is both laborious and expensive, which makes profiling all TFs challenging. For instance, the binding preferences of ~25% human TFs remain unknown; they neither have been determined experimentally nor inferred computationally. Here, we introduce ModCRE, a web server implementing a structure homology-modelling approach to predict TF motifs and automatically model higher-order TF regulatory complexes. Starting from a TF sequence or structure, ModCRE predicts a set of motifs for that TF. The predicted motifs are then used to scan the DNA for occurrences of each of them, and the best matches are either profiled with a binding score or collected for their subsequent modeling into a higher-order regulatory complex with DNA, as well as other TFs and co-factors. Moreover, we demonstrate that incorporating high-throughput TF binding data, such as from protein binding microarrays, addresses the protein-DNA structure scarcity problem for deriving statistical potentials. In turn, these statistical potentials are proven to be capable predictors of TF motifs. We also show the conditional advantage of using ModCRE over a nearest-neighbor approach for predicting TF binding sites as well as an improvement in prediction accuracy when using a rank-enrichment selection system. Finally, as case examples, we apply ModCRE to model the interferon beta enhanceosome and the complex of SOX2 and 11 with a nucleosome.

Proteins

nf-core/proteinfold: a bioinformatics best-practice pipeline for protein 3D structure prediction

Athanasios Baltzis (Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology), Jose Espinosa-Carrasco (Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology), Luisa Santus (Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology), Martin Steinegger (School of Biological Sciences, Seoul National University), Harshil Patel (Seqera Labs) and Cedric Notredame (Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology).

Abstract:

The advances in deep learning frameworks have revolutionised protein studies and contributed to unprecedented accurate predictions of protein structures. The release of AlphaFold2 has materialised this major breakthrough[1]. AlphaFold2 has also paved the way towards the development of a new generation of publicly available prediction tools whose combination constitutes a powerful toolkit for the systematic study of unknown proteomes[2]. Despite their obvious usefulness, the large-scale deployment of these new methods across the research community remains hampered by technical hurdles. The main challenge is the dependency of each prediction method onto a wealth of external software and sequence databases. This very problem of dependencies can be specifically addressed by Nextflow. Here we present nf-core/proteinfold (https://nf-co.re/proteinfold), a Nextflow pipeline developed according to nf-core guidelines[3] that enables the use of state of the art techniques in protein structure modelling. These best-practice guidelines ensure that the pipeline is scalable, reproducible and portable for execution on the major cloud providers as well as HPC infrastructures. We foresee that this development endeavour will have a significant impact in a variety of biological analyses based on protein structures by granting access to an open-source, community developed resource to obtain protein folds.

[1]Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

[2]Mirdita, M. et al. ColabFold: making protein folding accessible to all. Nat. methods 19(6), 679-682 (2022).

[3]Ewels, P. A. et al. The nf-core framework for community-curated bioinformatics pipelines. Nat. Biotechnol. 38(3), 276-278 (2020).

Proteins

Novel ADP-ribosyltransferase families in the Legionella genus

Marianna Krysińska (Department of Biochemistry and Microbiology, Warsaw University of Life Sciences SGGW, Warszawa, Poland), Krzysztof Pawłowski (Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, TX, USA) and Marcin Gradowski (Department of Biochemistry and Microbiology, Warsaw University of Life Sciences SGGW, Warszawa, Poland).

Abstract:

The Legionella bacterium, the malicious intracellular pathogen of eukaryotes gets inside the host/victim cell by using an impressive set of 330+ effector proteins. Thus, it builds a suitable environment for itself to live and replicate. These effectors, delivered into the host cell, affect its signalling and metabolic pathways. Most of the effectors are still poorly underwood and not characterised.

Motivated by the richness of Legionella effector repertoires and their oftentimes atypical biochemistry, also by several atypical Legionella effector ADP-ribosyltransferases and pseudo-ADP-ribosyltransferases discovered by us and by others (lpg0080, lpg0181/Lart1, SidE), we undertook an in silico survey and exploration of the “pan-ARTome” of the Legionella genus.

In this study, we present a bioinformatics survey of the several novel ART effectors families discovered with use of non-standard bioinformatic approaches.

Notably, some of the novel ART families are also present in other bacterial taxa, including other pathogens, often phylogenetically very distant from Legionella.

This taxonomic spread involves several more intracellular pathogens, e.g. Salmonella enterica enterica sv. Typhi, Chlamydia trachomatis, reflecting the intense evolutionary arms race between the pathogen and the host.

While presenting functional predictions for novel ART families, we discuss implications for novel therapeutic scenarios against Legionella infections.

Proteins

Novel binding site descriptors built upon inverse virtual screening

Arnau Comajuncosa-Creus (IRB Barcelona), Miquel Duran-Frigola (IRB Barcelona), Xavier Barril (Universitat de Barcelona) and Patrick Aloy (IRB Barcelona).

Abstract:

Pocket descriptors embed relevant features of protein binding sites in the shape of numerical vectors [1]. Unlike small molecule fingerprints, strategies to derive binding site descriptors are scarce and usually exhibit limited applicability. We herein present PocketVec, a novel strategy to derive comprehensible binding site descriptors based on the assumption that similar pockets bind similar ligands and should thus result in similar rankings for a fixed set of docked chemical compounds. We are able to provide a unique, meaningful and handy descriptor for each binding site of interest, in a similar way molecular fingerprints do for small molecules. We benchmarked PocketVec descriptors in several pocket similarity exercises [2] and showed remarkable great performances when evaluating the sensitivity to the binding site definition and flexibility and also when detecting similar pockets in unrelated proteins. Finally, we combined PocketVec descriptors with molecular signatures [3,4] in order to predict protein-ligand interactions from a proteome-wide perspective and we proved that the use of PocketVec descriptors led to a significant enrichment in such predictions.

[1] Fernández-Torras, A., et al. Curr Opin Chem Biol 66 (2022): 102090.

[2] Ehrt, C. et al. PLoS Comput. Biol. 14.11 (2018): e1006483.

[3] Duran-Frigola, M., et al. Nat. Biotechnol. 38.9 (2020): 1087-1096.

[4] Bertoni, Martino, et al. Nat. Commun. 12.1 (2021): 1-13.

Proteins

Origin and evolution of Cas9 and Cas12 proteins

Darius Kazlauskas (Institute of Biotechnology, Life Sciences Center, Vilnius University), Lukas Valančauskas (Institute of Biotechnology, Life Sciences Center, Vilnius University) and Česlovas Venclovas (Institute of Biotechnology, Life Sciences Center, Vilnius University).

Abstract:

Viruses and their hosts are involved in a constant battle for survival. Bacteria and Archaea fight against invaders using various molecular weapons including adaptive defense systems known as CRISPR-Cas. These systems are divided into two classes based on the CRISPR-Cas effector complex. Class II systems have large multidomain proteins (Cas9/Cas12/Cas13) as effectors. It was suggested that Cas9 and Cas12 evolved from IscB and TnpB protein families, respectively. The relationship between Cas9/IscB and Cas12/TnpB is limited to RuvC domain. There are some functionally equivalent regions such as PAM-interacting domain (or TAM-interacting domain in the case of IscB/TnpB) and 'Wedge' domain, but their mutual relationship has not been explored. Even the detailed evolutionary relationship between RuvC domains of Cas9 and Cas12 has not been analyzed. Here, we attempted to trace back evolutionary relationships of both Cas9 and Cas12 proteins. We identified the conserved evolutionary core of Cas9 and Cas12 proteins and structural-functional domains that constitute this core. We have further explored putative relationships of these domains and the structural-functional context of related domains in other proteins. Thus, this study provides insights into the origins of both Cas9 and Cas12 families.

Proteins

PACT – Prediction of Amyloid Cross-interactions by Threading

Jakub Wojciechowski (Wrocław University of Science and Technology) and Małgorzta Kotulska (Wrocław University of Science and Technology).

Abstract:

Amyloids are protein aggregates most commonly known for their role in the development of severe neurodegenerative diseases such as Alzheimer's or Parkinson's disease. However, the unique features of such structures were utilized by many organisms for a wide range of physiological roles including biofilm formation and hormone storage. More recent studies have shown that in some cases the presence of amyloid aggregates can affect the aggregation kinetics of other proteins. This so called cross-seeding or, more generally, cross-interactions turned out to be crucial for understanding comorbidity of amyloid related diseases, including Alzheimer's disease and type II diabetes. Despite the importance of the process, our understanding of it is still very limited due to costly and time consuming experiments required to study such interactions. To overcome this problem, we have developed PACT method. The

method is based on modeling of the heterogenous fibril, formed by two sequences of interest. Such a model is then assessed using DOPE statistical potential implemented in Modeller software. The main assumption of the method is that pairs of interacting amyloids will be more energetically favorable than negative cases. Based on that, it is possible to find an energy threshold for cross-interacting pairs. Importantly, the method can work with long protein fragments and, as a purely physicochemical model, it relies very little on the training data. The method, for the first time, opens the possibility of high throughput study of amyloid interactions.

Proteins

PanPA: Construction and Alignments of Panproteome Graphs

Fawaz Dabbaghie (Institute for Medical Biometry and Bioinformatics, Faculty of Medicine, Heinrich Heine University Düsseldorf), Sanjay Srikakulam (Drug Bioinformatics, Helmholty Institute for Pharamceutical Research Saarland, Saarbrücken), Olga Kalinina (Drug Bioinformatics, Helmholty Institute for Pharamceutical Research Saarland, Saarbrücken) and Tobias Marschall (Institute for Medical Biometry and Bioinformatics, Faculty of Medicine, Heinrich Heine University Düsseldorf).

Abstract:

Compared to eukaryotes, prokaryote genomes are more diverse through different mechanisms, including a higher mutation rate and horizontal gene transfer. Therefore, using one linear reference genome cannot capture the full diversity spectrum. As a result, the reference used introduces bias. Graph-based pangenomes are one proposed solution in recent years. However, working with DNA sequences is challenging in interspecies comparisons even within clades of related organisms. In contrast, amino acid sequences have higher similarity as they are more susceptible to selection and more conserved. Moreover, coding regions in prokaryotes cover the majority of the genome. Thus, building panproteomes instead of pangeomes can be a good alternative, enabling the analysis of more distant prokaryotes and keeping higher sequence similarity while at the same time not losing much of the genome as non-coding regions.

We present PanPA, a command-line software that takes set of multiple sequence alignments of individual proteins or protein clusters, indexes them, builds a directed graph for each of them, and aligns DNA or amino acid query sequences back to these graphs. We first showcase that PanPA generates correct alignments by building a panproteome from 1,350 E. coli assemblies and aligning back a sample of the original sequences, demonstrating that they align back correctly. We also demonstrate the ability to align sequences from more distant organisms by collecting coding regions of 1,073 Salmonella enterica assemblies and aligning them against E. coli reference genome, pangenome, and panproteome using BWA, GraphAligner, and PanPA respectively, where PanPA aligned 22% more sequences.

Proteins

PrankWeb3 - ligand binding site prediction for PDB and AlphaFold structures

David Jakubec (Faculty of Mathematics and Physics, Charles University), Petr Škoda (Faculty of Mathematics and Physics, Charles University), Radoslav Krivak (Faculty of Mathematics and Physics, Charles University), Marian Novotny (Faculty of Science, Charles University) and David Hoksza (Faculty of Mathematics and Physics, Charles University).

Abstract:

PrankWeb is a web-based state-of-the-art ligand-binding site prediction tool. We introduce a new version with two major and an array of minor improvements. The major improvements involve a new, faster and more accurate evolutionary conservation estimation pipeline and the ability to carry out LBS predictions in situations where no experimental structure is available.

The original version of PrankWeb utilized an evolutionary conservation calculation pipeline which could take up to several hours to finish for one protein. In PrankWeb 3.0 we replaced the evolutionary rate-based conservation scores with an entropy-based metric. The new conservation calculation pipeline utilizes the HMMER3 package for fast and sensitive sequence similarity searches against the UniRef50 sequence database. This change led to more accurate predictions and enabled us to reduce the average time required for the conservation score calculations down to a few minutes.

To extend the functionality of PrankWeb to proteins with no experimental structures available, we have pre-computed the LBS predictions for the structural models from the newly developed AlphaFold database (ADB) and integrated the results into our web server. The user has now the option to enter a UniProt accession number; then, if available, the corresponding model is fetched from the ADB, LBSs are predicted and presented to the user.

The minor improvements include the ability to deploy PrankWeb as a Docker container, support for the mmCIF file format, improved public REST API access, or the ability to batch download the LBS predictions for the whole PDB archive and parts of the ADB.

Proteins

Predicting the binding affinity of antibodies with antigens

Nahin Khan (Qatar Computing Research Institute, Hamad Bin Khalifa University), Hossam Almeer (Qatar Computing Research Institute, Hamad Bin Khalifa University), Joao Palotti (Qatar Computing Research Institute, Hamad Bin Khalifa University), Sanjay Chawla (Qatar Computing Research Institute, Hamad Bin Khalifa University) and Ehsan Ullah (Qatar Computing Research Institute, Hamad Bin Khalifa University).

Abstract:

Antibodies are a critical component of the adaptive immune system in the human body's defense against infection. The effectiveness of this immune response relies significantly on the strength of the binding of antibodies with specific antigens that can act as identifiers of invasive pathogens. It has been shown that antibody-antigen binding has some predictability and this has been used in research for the study and design of novel and known antibodies. Machine learning approaches have shown the potential to predict antibody-antigen binding affinity. Given the fast and constantly evolving nature of many viruses, such computational tools could present feasible methods to predict the efficacy of both naturally and artificially acquired immunity against emerging viral variants. We propose a computational framework to estimate antibody-antigen binding affinity based on protein sequence data. The proposed data-driven deep learning framework uses antibody and viral protein amino acid sequences as features instead of relying on predicted or crystallized structures of viral proteins and antibodies to predict binding affinity of viral antigens with antibodies.

Proteins

Predictive and comparative analysis of interaction hotspots of ACE2 orthologs in potential intermediate hosts for SARS-CoV-2

Myeongji Cho (Honam National Institute of Biological Resources), Nara Been (Seoul National University) and Hyeon S. Son (Seoul National University).

Abstract:

In this study, the potential host range of SARS-CoV-2 was estimated through evolutionary distance analysis of ACE2 protein sequences derived from 46 species, along with prediction and evaluation of hotspot conservation in the secondary structures of ACE2 orthologs. Prediction of interaction hotspots was achieved by classifying sequence segments with structural differences using Uversky's algorithm, to transform the boundary line equation that separates folded from disordered proteins. Species specificity of ACE2 orthologs that may affect the virus-host interaction was analyzed. Amino acid residues predicted to be interaction hotspots were mapped to multiple alignment results of ACE2 sequences in 46 species, including Homo sapiens. For hotspot residues predicted to have different distribution patterns among species, interspecies conservation rates were calculated and compared. Estimation of the potential host range of SARS-CoV-2 was performed based on the evolutionary distance and interspecies hotspot conservation rate of the ACE2 sequences of 45 animal species. The risk of interspecies infection in humans was estimated to be highest in primates, followed by mammals, rodents, and reptiles, in terms of evolutionary distance alone. Interestingly, among reptiles, snakes were estimated to have a high risk of interspecies infection with humans, suggesting that they have potential as an intermediate host for SARS-CoV-2.

Proteins

ProtVar: Protein coding variant annotation

Rizwan Ishtiaq (EMBL-EBI), James Stephenson (EMBL-EBI), Alok Mishra (EMBL-EBI) and Maria Jesus Martin (EMBL-EBI).

Abstract:

Whilst often benign, variation to the wildtype sequence of proteins can alter their function, sometimes to the detriment of the organism. Understanding the impact of variants can therefore be critically important to several fields including drug resistance, diagnosis of genetic diseases and synthetic biology. To consider the impact of variants two broad lines of evidence can be used. Firstly, understanding the role of the wild type amino acid via structural and functional annotations. Secondly, considering previously reported perturbations in that position or region. Whilst this data can be accessed from a variety of sources, retrieving and collating it could be prohibitively time consuming.

UniProt collates protein coding genetic variants for several species from external databases and literature. By mapping these to protein space they can be annotated with the wealth of functional information contained within UniProt. Variants can be viewed per protein in parallel with all functional annotations in the interactive UniProt Feature Viewer. Annotations and co-located variants for individual human variants can also be viewed via the ProtVar web-tool which shows the variant position in protein structures and also allows access via user entered genomic coordinates. These resources can be accessed programmatically via APIs, and categories of variants, for example enzymes, metal binding or transmembrane proteins can be filtered using the advanced search.

The suite of variant mapping and visualisation tools available in UniProt provides researchers with a single location, with regularly updated data, to gather variant and functional information to help them consider potential variant impact.

Proteins

Reduced structural flexibility of eplet amino acids in HLA proteins

Diego Amaya (INRIA), Romain Lhotte (INSERM), Magali Devriese (INSERM), Constantin Hays (INSERM), Jean-Luc Taupin (INSERM) and Marie-Dominique Devignes (Inria, LORIA).

Abstract:

Abstract :

A recent study [1] has demonstrated the importance of reduced structural flexibility to identify favorable regions in antigens for antibody binding. Amino acid flexibility is expressed through the Normalized Root Mean Square Fluctuation (N-RMSF) during a molecular dynamics (MD) simulation. However this study is restricted to 61 proteins where sequence dissimilarity was favored. In the present work we tested the underlying hypothesis in [1] applied to the case of HLA proteins, which has the peculiarity of high sequence similarity. Moreover, we were interested in studying the difference in flexibility at the eplet rather than epitope level.

For this purpose, we performed MD simulations of 203 HLA proteins. The N-RMSF of residues having at least a Relative Solvent Accessible Surface Area (RSASA) of 20% was then calculated. The list of confirmed eplets in the HLA system was obtained from EpRegistry. The Kolmogorov-Smirnov test was used to verify whether there is a statistically significant difference between the two distributions of N-RMSF (eplet residues vs non-eplet residues).

As a result, we found that there is a significantly reduced flexibility for eplet residues compared with non-eplet residues. This result opens the door to the use of structural flexibility to identify antibody binding sites on HLA proteins.

References

D.G. Kim et al, « Epitopes of Protein Binders Are Related to the Structural Flexibility of a Target Protein Surface », J. Chem. Inf. Model., 2021

Proteins

Schistosoma mansoni MEG family proteins in the environment of host-parasite interactions

Stepanka Nedvedova (Department of Chemistry, FAFNR, Czech University of Life Sciences, Prague and ISA, Universite Claude Bernard Lyon 1), Kristyna Peterkova (Department of Parasitology, Faculty of Science, Charles University, Prague), Vojtech Vacek (Department of Zoology and Fisheries, FAFNR, Czech University of Life Sciences, Prague), Petr Mateju (Department of Animal Science and Food Processing, FTA, Czech University of Life Sciences, Prague), Lukas Konecny (Department of Parasitology, Faculty of Science, Charles University, Prague), Jan Dvorak (Department of Zoology and Fisheries, FAFNR, Czech University of Life Sciences, Prague), Adriana Erica Miele (Institut des Sciences Analytiques, Universite Claude Bernard Lyon 1), Francesca Fiorini Tregouët (Institut de Biologie et Chimie des Protéines, Universite Claude Bernard Lyon 1) and Maggy Hologne (Institut des Sciences Analytiques, Universite Claude Bernard Lyon 1).

Abstract:

Schistosomiasis is a vector-borne parasitic disease affecting over 250 million people in Africa, the Caribbean, South America, South-East Asia and Mediterranean Europe; it is one of the most devastating parasitic diseases worldwide. In the egg secretome of Schistosoma mansoni, about 188 proteins have been identified, of which only a few have been well characterized. One group of highly expressed proteins of the S. mansoni egg secretome is represented by an enigmatic group of genes referred to as MEGs (Micro-Exon Genes). S. mansoni MEGs contain short symmetric exons comprising about 80% of the entire gene sequence. Proteins encoded by this group of genes represent a unique system of protein variants generated by alternative splicing. Our lab identified three MEG members in the egg transcriptome, yet they belong to the most highly expressed proteins in mature eggs. Based on reported interactions between S. mansoni eggs and their human host, transcriptomic analyses and already performed experiments, we suggest that these highly expressed MEG proteins and their splice variants could play an essential role in host-parasite interactions. This project aims to use a combination of computational and experimental work to determine their structure and interaction partners. In silico methods used to achieve this objective are ab inito structure prediction/homology modelling; interaction studies will be performed using extracellular matrix interaction partner prediction (MatrixDB) or molecular docking. These methods are accompanied by the expression of S. mansoni MEG proteins and their biophysical analysis (CD, DLS, SAXS) and subsequent structure determination by NMR.

Proteins

Single-cell map of Acute Myeloid Leukaemia

Alice Driessen (IBM Research), Susanne Unger (University of Zurich), An-Phi Nguyen (IBM Research), Burkhard Becher (University of Zurich) and Maria Rodriguez Martinez (IBM, Zurich Research Laboratory).

Abstract:

Acute myeloid leukaemia (AML) is a haematological cancer in the bone marrow, with accumulation and expansion of immature cells of the myeloid lineage. Standard treatment of AML is chemotherapy, which does not achieve durable remission in most patients. Personalised medicine including immunotherapies have the potential to target chemotherapy resistant cells and achieve long-term remission. Identifying suitable targets for AML therapy is hampered by the heterogeneity and complex clonal composition of the cancer, as well as its complex evolution as the disease progresses. We aim to build a single-cell cytometry AML map to identify malignant cells and place them along the developmental trajectory using data from 20 patients and three time points over the course of the disease. We train a variational auto-encoder structure on healthy cells, which learns cellular reconstruction as well as the cell type classification. The latent space of the auto-encoder provides a meaningful representation of the healthy bone marrow cells to which we can map new cells. We use the trajectory assignment to segment patients into groups as well as to study the time evolution of the disease in terms of the distribution of malignant cells across the myeloid lineage.

Proteins

SparseChem: Fast and accurate machine learning model for small molecules

Adam Arany (ESAT/STADIUS, University of Leuven), Jaak Simm (Katholieke Universiteit Leuven), Martijn Oldenhof (Katholieke Universiteit Leuven) and Yves Moreau (Katholieke Universiteit Leuven).

Abstract:

It is a challenging task to build and train a machine learning model using very high dimensional sparse data which is the case for quantitative structure–activity relationship (QSAR) modelling. Therefore, SparseChem offers an easy and efficient way to train an industry-scale (millions of input compounds) multi-task QSAR deep learning model with high-dimensional sparse input features. It is possible to train classification, regression and censored regression models, or combination of them from command line or directly from Python. SparseChem allows for inference and training to run on both CPU and GPU hardware. Out of the box many metrics are supported which are computed per task such as AUC-PR, AUC-ROC, F1, Kappa for classification and correlation, R2, RMSE for regression. In the context of a European Innovative Medicines Initiative (IMI) project for federated learning called MELLODDY which gathers 10 pharmaceutical companies, academic research labs, large industrial companies and startups, the SparseChem library was used to build single partner and federated models.

Proteins

Spatial relationships in the urothelial cancer microenvironment, the potential for immune-based subtyping and immunotherapy response prediction

Alberto Gil-Jimenez (Netherlands Cancer Institute (NKI)), Nick van Dijk (Netherlands Cancer Institute (NKI)), Yoni Lubeck (Netherlands Cancer Institute (NKI)), Maurits L. van Montfoort (Netherlands Cancer Institute (NKI)), Erik Hooijberg (Netherlands Cancer Institute (NKI)), Annegien Broeks (Netherlands Cancer Institute (NKI)), Bas van Rhijn (Netherlands Cancer Institute (NKI)), Daniel J. Vis (Netherlands Cancer Institute (NKI)), Michiel S. van der Heijden (Netherlands Cancer Institute (NKI)) and Lodewyk F. A. Wessels (Netherlands Cancer Institute (NKI)).

Abstract:

Background: Immune checkpoint inhibitors (ICI) can achieve remarkable clinical responses in urothelial cancer (UC). However, it remains unclear which aspects of the tumor microenvironment (TME), usually characterized by immune cell density, determine a patient's response. Importantly, density metrics ignore cells' spatial arrangement (SA) relative to each other.

Methods: We quantified the TME SA in multiplex immunofluorescence data of 24 pretreatment UC biopsies with the first-nearest neighbor (1-NN) distance statistic using a conventional approach (G-function), and a novel approach that fitted a Weibull distribution. Furthermore, we classified the tissue compartments into tumor or stroma based on cancer and non-cancer cell local density. We performed a simulation study to identify SA parameters' sources of variation. Lastly, we associated the TME parameters with ICI (ipilimumab+nivolumab) response.

Results: We found that density perturbations affected the SA metrics of rare cell types (i.e., B-cells) but not of abundant cell types (i.e., cancer cells). The G-function quantification correlated with the Weibull parameters. Nevertheless, the G-function’s distance threshold created a variable effect size and statistical power in association studies. The SA metrics outperformed immune cell density in ICI response prediction. Specifically, no immune cell density metric discriminated between ICI response groups. In contrast, the CD8 T-cell or macrophage SAs to their closest cancer cell did. Furthermore, low 1-NN distances from CD8 T-cells to B-cells were associated with non-response.

Conclusion: We created a framework to quantify, interpret and analyze SAs, and illustrated their superior clinical relevance compared to density metrics for ICI treatment patient stratification.

Proteins

Structural and functional insights into P protein

Shahram Mesdaghi (University of Liverpool).

Abstract:

Recent innovations in computational structural biology have opened up an opportunity to revise our current understanding of the structure and function of clinically important proteins. This study centres on human P protein (Oca2) which is located on mature melanosomal membranes. Mutations of P protein can result in a form of oculocutanous albinism (OCA) which is the most prevalent and visually identifiable form of albinism. Sequence analysis predicts P protein to be a member of the SLC13 transporter family, however, it has not been classified into any existing SLC families. The modelling of P protein with AlphaFold 2 and other advanced methods shows that, like SLC13 members, it consists of a scaffold and transport domain and displays a pseudo inverted repeat topology that includes re-entrant loops. This finding contradicts the prevailing consensus view of its topology. In addition to the scaffold and a transport domains the presence of a cryptic GOLD domain is revealed that is likely responsible for its trafficking from the endoplasmic reticulum to the Golgi prior to localisation at the melanosomes and possesses known glycosylation sites. Exploiting the AlphaFold2 multimeric modelling protocol allowed the modelling of a plausible homodimer. Analysis of the putative ligand binding site of the model shows the presence of highly conserved key asparagine residues that suggest P protein may be a Na+/dicarboxylate symporter. Known critical pathogenic mutations map to structural features present in the repeat regions that form the transport domain.

Proteins

Structural modeling and computational analysis of uncharacterized anti-CRISPR Cas proteins

Lukas Valančauskas (Institute of Biotechnology, Life Sciences Center, Vilnius University), Darius Kazlauskas (Institute of Biotechnology, Life Sciences Center, Vilnius University) and Česlovas Venclovas (Institute of Biotechnology, Life Sciences Center, Vilnius University).

Abstract:

In prokaryotic organisms CRISPR-Cas systems can provide immunity against invading bacteriophages. However, some phages have the capability to inhibit CRISPR-Cas systems through the usage of anti-CRISPR-Cas (Acr) proteins. Acrs have been shown to interact with Cas proteins in multitude of ways thereby reducing the ability of Cas proteins to bind nucleic acids or hydrolyze them.

There is a large and diverse set of proteins, experimentally identified as Arcs, but with unknown mechanism of action. To explore structures of these Arcs, we applied a large-scale structure modeling using AlphaFold. Structural models revealed a lack of prevailing or unifying structural motifs common to analyzed Acrs, supporting their highly divergent nature. Nonetheless, a sizable portion of models showed structural similarity to known protein structures while displaying little similarity at the sequence level. Detection of similar structural domains was also facilitated by structural modeling of Arc oligomers as some Acrs make up compact structures using more than one chain. As a result, this study not only provided structural characterization of many Arcs, but also explored various approaches that could be used for analysis of these highly divergent proteins.

Proteins

Structural modelling of odorant receptors from Aedes aegypti and search for natural repellents

Vikas Tiwari (National Centre for Biological Sciences) and Ramanathan Sowdhamini (National Centre for Biological Sciences).

Abstract:

Insects have well-developed olfactory system to assist in multiple behaviours including foraging, mating and host detection. Olfaction mediated host preference can be of negative value to humans if the insect is an agricultural pest or a disease vector. Odorant receptors (ORs) play an important role in mediating the olfactory behaviour. ORs require a co-receptor (Orco) for their proper localization and function. Inhibitor design against ORs from disease vectors like Aedes aegypti can be facilitated by availability of structural information. OR4 of A. aegypti imparts human host preference and therefore OR4 is a potential target for repellent design. In this study, the structures of OR4 and Orco of A. aegypti were modelled using homology modelling, followed by MD simulation in membrane-aqueous environment, in order to assess the model stability. The database of natural products (Super Natural II) and FDA approved drugs (DrugBank) along with known repellent molecules were screened for potential inhibitor using molecular docking approach. The docked complexes were assessed for binding energy calculation using MMGBSA and interaction stability through MD simulation. We observed that common repellents like DEET bind strongly to conserved Orco compared to finely tuned OR4. Many natural compounds were ranked better than known repellents which can be further tested experimentally. An example compound “Mulberroside A” from “White mulberry” plant is among top 300 hits against OR4 and the extracts of white mulberry plant have been shown to have repellent property against certain pests.

Proteins

Structure-activity relationships generated for peptaibols produced by Trichoderma: via accelerated MD simulations

Dóra Balázs (University of Szeged, Department of Microbiology), Chetna Tyagi (University of Szeged, Department of Microbiology), Tamás Marik (University of Szeged, Department of Microbiology), András Szekeres (University of Szeged, Department of Microbiology), Csaba Vágvölgyi (University of Szeged, Department of Microbiology) and László Kredics (University of Szeged, Department of Microbiology).

Abstract:

The filamentous fungal species from the genus Trichoderma play an important role in agriculture and biotechnology due to their potential application in biocontrol of phytopathogenic microorganisms and their plant growth-promoting properties. Numerous Trichoderma species produce secondary metabolites with favourable properties, out of which peptaibiotics make the largest group. Due to the way of synthesis by non-ribosomal peptide synthetases (NRPSs) and the incorporation of non-proteinogenic amino acids to the sequences, the peptaibols are characterized by a high degree of amino acid variability in their sequences. Using modern molecular modelling techniques, we can gain a deeper insight into the structural characteristics of peptaibols.

In our study, purified peptaibol extracts from six Trichoderma species were investigated through in vitro and in silico analysis. The effects of peptaibol extracts were studied against nine commonly known Gram-positive and Gram-negative bacteria and the minimal inhibitory concentrations (MIC, mg ml-1) were determined. In parallel with laboratory tests, accelerated molecular dynamics (aMD) simulations were performed on the selected peptaibol sequences to explore the folding mechanisms and representative structures. The structure-activity relationships (SARs) of the peptaibols characterized in two distinct groups were investigated. The most characteristic difference between the peptaibols are the 'Gly-Leu-Aib-Pro' and 'Gly-Aib-Aib-Pro' amino acid motifs, which have a significant effect on the structure dynamics and stability, and appears to affect the expressed bioactivity. Correlations established with SARs can lead to an efficient selection of peptaibiotic compounds for the practical application in agriculture and plant treatment.

Proteins

Studying the effect of phosphorylations on protein backbone dynamics

David Bickel (Vrije Universiteit Brussel) and Wim Vranken (Vrije Universiteit Brussel).

Abstract:

A protein's function is intrinsically linked to its fold and biophysical characteristics. As part of the complex system that regulates proteins in the cell, phosphorylations commonly modulate protein activities, protein-protein interactions, or their subcellular localization. An accurate description of the link between these phosphorylations and phenotypic effects at the protein level, such as changes in backbone dynamics or conformation, is still missing. With few experimental data available, molecular dynamics provides a suitable framework to study the influence of phosphorylations on protein dynamics on the molecular level.

In our study we set up a simulation framework to study the influence of phosphorylations on the local dynamics and conformational preferences. Using glycine-based backbones, we analyzed how isolated sidechains influence the backbone dynamics. Thereafter, we applied an enumerative approach running simulations in the millisecond timescale to explore every possible local sequence context.

This systematic study will allow us to project the knowledge gained from the simulations onto any phosphorylation-site in the whole proteome and predict changes in the local protein dynamics. These in turn may allow for a biophysics-based classification of phosphorylation-sites.

Proteins

Superfamily analysis of rice protein structures reveals a variety of stress tolerance mechanisms

Fatima Shahid (Universiti Kebangsaan Malaysia), Nicola Bordin (University College London), Christine Anne Orengo (University College London) and Su Datt Lam (Universiti Kebangsaan Malaysia).

Abstract:

Rice serves as a staple food for half of the world's population. However, a large amount of rice yield is lost due to biotic/abiotic stresses affecting rice protein structures. Earlier, understanding rice protein structures was difficult due to the low number of experimentally solved structures. Recently, the AlphaFold Protein Structure database provided models for 48 organisms, including rice. However, the quality of these models needs to be validated. This study focused on extracting rice AlphaFold models and chopping them into CATH domains using our in-house pipeline. The domains were evaluated using a series of model quality assessment (MQA) filters. Filtered protein domains are classified into CATH superfamilies using established tools. More than 50k rice domains were identified and ~50% of the domains passed our MQA filter. To give more context, we compared the rice superfamilies with the model plant Arabidopsis thaliana. Both rice and Arabidopsis share a large number of common superfamilies (e.g. leucine-rich repeat, ribonuclease inhibitor, zinc finger), implying both plants have similar survival strategies. Surprisingly, we found 700 more Ribonuclease inhibitor domains in rice compared to Arabidopsis. Ribonuclease inhibitor is involved with cell defence against pathogens. We found another rice-unique Ricin domain superfamily. Ricins behave like ribosome-inactivating proteins and have been implicated to be stress-related. Overall, this is the first large-scale study of plant protein structures and will be helpful in future to obtain more insights into plant survival under stress conditions.

Proteins

Targeting SARS-CoV-2 Endoribonuclease: A structure-based virtual screening supported by in vitro analysis

Ibrahim Mohamed (Biophysics Department, Faculty of scinece, Cairo University), Abdo Elfiky (Biophysics Department, Faculty of scinece, Cairo University), Mohamed Fathy (Biophysics Department, Faculty of scinece, Cairo University), Sara Mahmoud (Centre of Scientific Excellence for Influenza Viruses (CSEIV), National Research Centre, Cairo, Egypt) and Mahmoud Elhefnawi (Informatics and Systems Department, Division of Engineering Research, National Research Centre, Cairo).

Abstract:

Researchers worldwide are focused on discovering compounds that can interfere with COVID-19 life cycle. One of the important non-structural proteins is endoribonuclease since it is responsible for processing viral RNA to evade detection of the host defense system. This work investigates a hierarchical structure-based virtual screening approach targeting NSP15. Different filtering approaches to predict the interactions of the compounds have been included in this study. Using a deep learning technique, we screened 823,821 compounds from five different databases (ZINC15, NCI, Drug Bank, Maybridge, and NCI Diversity set III). Subsequently, two docking protocols (extra precision and induced fit) were used to assess the binding affinity of the compounds, followed by molecular dynamic simulation supported by the MM-GBSA free binding energy. Interestingly, one compound (ZINC000104379474) from the ZINC15 database has been found to have a good binding affinity of -26.1399 Kcal/Mol. The VERO-E6 cell line was used to investigate its therapeutic effect in vitro. Half-maximal cytotoxic concentration and Inhibitory concentration 50 were determined to be 0.9 mg/ml and 0.01 mg/ml, respectively, therefore the selectivity index is 90. In conclusion, ZINC000104379474 was shown to be a good hit for targeting the virus that needs further investigations in vivo to be a drug candidate.

Proteins

The alteration of structural network by transient association between proteins

Vasam Manjveekar Prabantu (Indian Institute of Science), Himani Tandon (MRC Laboratory of Molecular Biology), Sankaran Sandhya (Ramaiah University of Applied Sciences) and Narayanaswamy Srinivasan (Indian Institute of Science).

Abstract:

Proteins often interact (predominantly, non-covalent bonding) with other interacting units inside a cell. Generally, there is an impact on the overall structural topology of a protein due to such transient interactions that in turn enable proteins to mediate their function. Such alteration of the structural topology upon binding of a partner can be studied by employing protein structural networks (PSNs), which are the node-edge representative models of protein structures, reported as a robust tool for capturing interactions between residues. Several methods have been optimised to collect meaningful, functionally relevant information by studying alteration of structural networks. In this work, PSNs are used along with spectral decomposition of graphs to study the impact of protein-protein interactions on the intra-residue communication within the bound and unbound forms. A detailed analysis of the structural network of interacting partners is performed across a dataset of bound and corresponding unbound protein structures. The variation in network parameters at, around and far away from the interface shows that residue-residue communication within a protein is not only impacted at the interfaces, but also at sites away from the interfaces. In interesting case studies, an allosteric mechanism of structural impact is presented from community and communication-path detection methods.

Proteins

The clinical importance of tandem exon duplication-derived substitutions

Laura Martinez-Gomez (Spanish National Cancer Research Centre (CNIO)), Fernando Pozo Ocampo (Spanish National Cancer Research Centre (CNIO)), Thomas A. Walsh (EMBL-EBI), Federico Abascal (Wellcome Trust Sanger Institute) and Michael Tress (Spanish National Cancer Research Centre).

Abstract:

Tandem exon duplication-derived substitutions are splice events in which duplicated adjacent exons are spliced in a mutually exclusive manner. Although tandem exon duplications are a relatively common occurrence, tandem exon duplication-derived substitutions are not.

In depth manual curation of the 236 events annotated in the human gene set showed that tandem exon duplication-derived substitutions are substantially older than any other type of splice event. In fact, 91% are present in at least one fish species, compared to just 4% of all other alternative splice events. We traced 21 events back to a last common ancestor present prior to the split between vertebrates and invertebrates, more than 670 million years ago. This makes them among the oldest known splice events.

Two further characteristics support the functional relevance of tandem exon duplication-derived splice events: they are translated in much higher numbers than other splice events, and they are substantially more clinically important. Tandem exon duplication-derived events have proportionally 27 times more clinically important mutations as other splice events.

The functional importance of exons in tandem exon duplication-derived substitutions is related to their homology, not their splicing mechanism. Non-homologous mutually exclusively spliced exons are neither more conserved, more expressed, nor more clinically important than any other type of splice event.

Curiously, despite their clear importance, homologous exons are a relatively little studied class of splice variants. We hope that the annotation of a complete set of homologous substitutions for the human genome will inspire research into these important, highly conserved splice events.

Proteins

The importance of protein-protein interactions in Toll-like receptor 8 functioning

Maria Bzówka (Tunneling Group, Biotechnology Centre, Silesian University of Technology, Gliwice, Poland), Weronika Bagrowska (Tunneling Group, Biotechnology Centre, Silesian University of Technology, Gliwice, Poland) and Artur Góra (Tunneling Group, Biotechnology Centre, Silesian University of Technology, Gliwice, Poland).

Abstract:

Toll-like receptors (TLRs) are transmembrane proteins which trigger a proinflammatory response. Each TLR contains three domains: leucine-rich repeats (LRRs), transmembrane helix, and cytoplasmic Toll/IL1 receptor (TIR). The common mechanism of TLR signalling is that the interaction of a ligand with LRRs induces the formation of a receptor dimer or changes the conformation of a preexisting dimer. The TIR domain can then host adaptor protein, such as MyD88, and ensure further signal transduction. For some TLRs, e.g. TLR8, the proteolytic cleavage of the long Z-loop located in the LRRs is necessary to enable the interaction with a ligand. It is assumed that this process is performed by furin.

The study of the signal transduction by TLR is hampered by the lack of available structures and models which could be used to describe the phenomena. Therefore, we implemented state-of-the-art in silico methods including AI-supported protein structure prediction and a small-molecule tracking approach in order to get insight into the molecular basis for protein-binding regions within TLR8. We performed the analysis of LRRs-furin to describe the proteolytic cleavage and of TIR-MyD88 to characterise the adaptor protein binding. We believe that obtained models can provide a starting point for the design of experimental validation of the proposed findings.

The work was supported by the Ministry of Science and Higher Education, Poland from the budget for science for the years 2019-2023, as a research project under the "Diamond Grant" programme [DI2018 014148]. This research was supported in part by PL-Grid Infrastructure.

Proteins

Understanding the significance of inter-protein bifurcated interactions in protein-protein complexes

Sneha Bheemireddy (Indian Institute of Science, Bangalore, India), Revathy Menon (National Centre for Biological Sciences (NCBS)) and Narayanaswamy Srinivasan (Indian Institute of Science).

Abstract:

Multi-protein assemblies play a crucial role in several cellular processes. Studying the functional basis of such complexes begins with the analysis of protein-protein interactions. Several studies have highlighted the significance of interfacial residues in protein-protein complexes and their role in conferring stability and specificity to the complex. While they have been studied in detail, inter-protein bifurcated interactions are still an unexplored corner. In this work, the features of inter-protein bifurcated interactions in multi-protein assemblies have been investigated. We begin our study by generating a dataset of heteromeric complexes of known 3-D structures. Upon screening of these complexes, we found that inter-protein bifurcated interactions are present in over 600 multi-protein assemblies. Arg, Tyr, and Leu are the highly occurring amino acids in bifurcated inter-protein interactions. Van der Waals interactions, hydrophobic interactions, and salt bridges are the most frequent interaction types. Further, we found that most of these residues are hotspots, and they are moderate to highly conserved, with a few exceptions. We could explain the biological significance of bifurcated interactions through a few case studies. Overall, this study expands the knowledge on protein-protein interactions paving the way for the learning of multi-protein assemblies.

Proteins

What is hidden in the darkness? A large-scale approach to make sense of all natural unknown proteins

Joana Pereira (Biozentrum and SIB Swiss Institute of Bioinformatics, University of Basel) and Torsten Schwede (Biozentrum and SIB Swiss Institute of Bioinformatics, University of Basel).

Abstract:

The collection of all theoretically possible protein sequences is known as the “protein universe”, a multidimensional space where individual proteins correspond to points whose distances are defined by their homologous relationships. In this space, protein families form galaxies and superfamilies clusters of galaxies, surrounded by dark areas seemingly unexplored by nature. However, the number of protein sequences deposited in protein repositories is increasing exponentially and the number of “hypothetical proteins” and “proteins of unknown function” is increasing proportionally. This can be due to the low sensitivity of the methods used, but also to the presence of sequences belonging to novel biological systems populating dark areas of the protein universe.

We have set up a computational approach to shed light on these sequences. We analysed all sequences in UniRef and classified them based on the annotation level of their close homologs. We found that 40% of all UniRef50 clusters are composed of proteins with a maximum of 5% coverage of functional annotations, i.e. they are seemingly dark. These are widespread in the tree of life, but the most common are from marine sediment metagenomes. We then modelled the sequence space covered by these sequences with protBERT sequence embeddings. Preliminary results suggest that bacterial and eukaryotic sequences occupy distinct areas of the landscape, and that dark proteins complement the areas occupied by those well-annotated (i.e., bright proteins). While most form bridges between well-studied areas, some sequences form well-delimited clusters that may correspond to protein families unrelated to those previously described.

Proteins

Where do they come from, where do they go? Prediction of protein subplastid localization and origin with PlastoGram.

Katarzyna Sidorczuk (University of Wrocław), Przemyslaw Gagat (University of Wroclaw), Jakub Kała (Warsaw University of Technology), Henrik Nielsen (Technical University of Denmark), Filip Pietluch (University of Wrocław), Paweł Mackiewicz (University of Wroclaw) and Michal Burdukiewicz (Medical University of Białystok).

Abstract:

Due to the complex history, plastids possess proteins encoded both in the nuclear and plastid genome. Moreover, these proteins localize to various subplastid compartments. Since protein localization is tightly associated with its function, prediction of subplastid localization is one of the most important steps in plastid protein annotation providing insight into their potential function. Therefore, we created a robust ensemble model for prediction of protein subplastid localization and origin. PlastoGram classifies a protein as nuclear- or plastid-encoded and predicts its localization considering envelope, stroma, thylakoid membrane or thylakoid lumen. For the latter, the import pathway (Sec or Tat) is also predicted. We made PlastoGram easy to use in different settings, both as a web server and an R package. Considering that no new tool has been created since 2017, we believe that PlastoGram offers a lot of new possibilities in the field of plastid protein annotation.

Systems

A functional analysis of omic network embedding spaces reveals key altered functions in cancer.

Sergio Doria-Belenguer (Barcelona Super-Computing Center), Alexandros Xenos (Barcelona Super-Computing Center), Gaia Ceddia (Barcelona Super-Computing Center), Noel Malod (Barcelona Super-Computing Center) and Natasa Przulj (Barcelona Super-Computing Center).

Abstract:

Advances in omics technologies have revolutionized cancer research by producing massive datasets. Common approaches to deciphering these complex data are by embedding algorithms of molecular interaction networks. These algorithms find a low-dimensional space in which similarities between the network nodes are best preserved. Currently available embedding approaches mine the gene embeddings directly to uncover new cancer-related knowledge. However, these gene-centric approaches produce incomplete knowledge, since they do not account for the functional implications of genomic alterations. We propose a new, function-centric perspective and approach, to complement the knowledge obtained from omic data. We introduce the Functional Mapping Matrix (FMM), which captures the distances between all available functional annotations in the embedding space of genes, that we obtain by embedding molecular interaction networks. We use our FMM to explore the functional changes in the most prevalent cancers (breast, prostate, lung, and colorectal) compared to their corresponding control tissues.

Systems

A mathematical model for strigolactone biosynthesis in plants

Abel Lucido (Universitat de Lleida), Ester Vilaprinyo (Universitat de Lleida) and Rui Alves (Universitat de Lleida).

Abstract:

Strigolactones mediate plant development, trigger symbiosis with arbuscular mycorrhizal fungi, are abundant in 80% of the plant kingdom and help plants gain resistance to environmental stressors. They also induce germination of parasitic plant seeds that are endemic to various continents, such as Orobanche in Europe or Asia and Striga in Africa. While strigolactones are crucial hormones, their biosynthesis is less than well understood. The genes involved in the early stages of strigolactones biosynthesis are well known in several plants. The regulatory structure and the latter parts of the pathway, where flux branching occurs to produce alternative strigolactones, are less understood.

Here we present a computational study that collects the available experimental evidence and proposes alternative biosynthetic pathways that are consistent with that evidence. Then, we test the alternative pathways through in silico simulation experiments and compare those experiments to experimental information in order to identify the most likely pathway design.

Our results predict the differences in dynamic behavior between alternative pathway designs. Independent of design, the analysis suggests that feedback regulation is unlikely to exist in the strigolactone biosynthesis. In addition, it indicates that engineering the pathway to produce higher amounts of strigolactones could be most easily achieved by increasing the production of beta-carotenes. Finally, we find that changing the ratio of alternative strigolactones produced by the pathway can be done by changing the activity of the enzymes after the flux branching points.

Systems

A Mechanistic Cellular Atlas of the Rheumatic Joint

Naouel Zerrouk (Sanofi R&D, GenHotel (Université Paris-Saclay)), Sahar Aghakhani (GenHotel (Université Paris-Saclay), Lifeware (Inria Saclay Île-de-France)), Vidisha Singh (GenHotel (Université Paris-Saclay)), Franck Augé (Sanofi R&D) and Anna Niarakis (GenHotel (Université Paris-Saclay), Lifeware (Inria Saclay Île-de-France)).

Abstract:

Rheumatoid Arthritis (RA) is an autoimmune disease of unknown etiology involving complex interactions between environmental and genetic factors. The disease affects multiple cellular functions leading to synovial inflammation, bone erosion and cartilage destruction in the patients' joints. The resident synoviocytes of macrophage and fibroblast types can interact with innate and adaptive immune cells and contribute to the disease's debilitating symptoms. A detailed, mechanistic mapping of the molecular pathways and cellular cross-talks is essential to understand the complex biological processes and different disease manifestations. In this work, we present the RA-Atlas, a manually curated atlas of the rheumatic joint to recapitulate existing knowledge related to the disease's onset and progression. This state-of-the-art atlas is compliant with the Systems Biology Graphical Notation (SBGN) standard representation. It includes a global RA-map (covering signaling, gene regulation and metabolic pathways) and cell-specific molecular interaction maps for CD4+ TH1 cells, fibroblasts, M1 and M2 macrophages. Overall, the different molecular interaction maps were obtained by updating pre-existing maps (if available) with information extracted from literature or pathway databases and enriched using omic data, in bulk and single cell resolution. The RA-Atlas is available as online interactive maps on the standalone web server MINERVA (https://ramap.uni.lu/minerva/), allowing visual exploration of experimental data, gene set enrichment analysis, pathway export or drug query. The maps can be used as a source of high-quality curated information for disease-related pathways, as a template for omic data analysis and as a starting point for dynamic computational models.

Systems

A user-friendly strain design pipeline enabling the reproducible identification of Knock-Out strategies for overproducing metabolites of interest.

Álvaro Gargantilla Becerra (SPANISH NATIONAL CENTER FOR BIOTECHNOLOGY (CNB-CSIC)), David San León Granado (SPANISH NATIONAL CENTER FOR BIOTECHNOLOGY (CNB-CSIC)) and Juan Nogales Enrique (SPANISH NATIONAL CENTER FOR BIOTECHNOLOGY (CNB-CSIC)).

Abstract:

Bio-based industry is starting to be a crucial player in our economy and thus in our daily lives. It is plausible to imagine a bio-based economy in the upcoming years, provided that enough efforts will be made in the optimisation and standardisation until the overall process becomes cost-effective. To address this challenge, computational strain optimisation algorithms based on genome-scale metabolic models (GEMs) have proven to be useful, as they can identify deletion strategies yielding overproducing strains for different products of interest. Concerning these approaches, a widely established strategy is to use the genetic deletions to redirect cell metabolism in a way that chemical synthesis is made obligatory when the cell grows at its maximum rate, thus is Growth-Coupled (GC). Although enabling the design of high producer phenotypes that are evolutionarily robust, these strategies are hard to find in the vast design space of GEMs and for some metabolites, there is none. Besides, the design algorithms required to deploy the design are intimidating for those inexperienced in computational methods, hampering its usage across the scientific community. Accounting for all this, we developed a user-friendly strain design pipeline which merges several well-known algorithms exploring GC and non-GC approaches. Our pipeline is implemented via a web-based interactive notebook, requiring minimal inputs to construct reproducible strategies yielding to overproducing strains. Through the pipeline, some graphical outputs are produced to inform the user about the strategies. The reliability of the pipeline was thoroughly validated for strategies that have been experimentally tested, proving its practical application.

Systems

An agent-based model of tumor-associated macrophage differentiation in chronic lymphocytic leukemia

Nina Verstraete (Cancer Research Center of Toulouse INSERM UMR 1037), Malvina Marku (Cancer Research Center of Toulouse INSERM UMR 1037), Marcin Domagala (Cancer Research Center of Toulouse INSERM UMR 1037), Helene Arduin (Cancer Research Center of Toulouse INSERM UMR 1037), Julie Bordenave (Cancer Research Center of Toulouse INSERM UMR 1037), Jean-Jacques Fournié (Cancer Research Center of Toulouse INSERM UMR 1037), Loic Ysebaert (Cancer Research Center of Toulouse INSERM UMR 1037), Mary Poupot (Cancer Research Center of Toulouse INSERM UMR 1037) and Vera Pancaldi (Cancer Research Center of Toulouse INSERM UMR 1037).

Abstract:

In chronic lymphocytic leukemia (CLL), cancerous B cells can drive monocytes to differentiate into Nurse-Like Cells (NLC) which are able to protect the cancer cells from spontaneous apoptosis, hindering the efficacy of immunotherapy. Physical contact between B-CLL cells and monocytes is required for the formation of NLC but the precise mechanisms by which leukemic cells influence this differentiation are still unknown. Building on an in vitro model of leukemia, we propose here a two-dimensional agent-based model simulating intercellular interactions and the monocyte to Nurse-Like Cell differentiation.

Using time-course measurements of B-CLL cell viability and concentration to optimize the model parameters, we were able to reproduce the experimentally observed dynamics. We further tested the model's predictive power by simulating specific NLC production features in relation to varying measured proportions of monocytes in each patient in the co-cultures. Our results suggest that this model could be made patient-specific using their blood monocytes counts, which is a routinely measured variable. Finally, the parameter sensitivity analysis suggested a strong role for phagocytosis from the myeloid cells to ensure the cancer cells survival, especially in the initial phases of the co-culture. This finding suggests that monitoring and potentially modulating phagocytosis could play a role in the control of NLC polarization in CLL and also help understanding tumor-associated macrophages formation even in solid tumors [1].

1. Nina Verstraete et al. An Agent-Based Model of Monocyte Differentiation into Tumor-Associated Macrophages in Chronic Lymphocytic Leukemia. https://www.biorxiv.org/content/10.1101/2021.12.17.473137v2

Systems

An interpretable Graph Convolutional Network for predicting disease-causing genes involved in SARS-CoV-2 infection.

Samuele Firmani (Helmholtz Center Munich), Valter Bergant (Technical University of Munich), Christoph Ogris (Helmholtz Center Munich), Annalisa Marsico (Helmholtz Center Munich) and Andreas Pichlmair (Technical University of Munich).

Abstract:

Viral infections are multisystemic diseases involving genetic and molecular alterations in the host. This makes the discovery of disease related genes and their associated molecular mechanisms challenging.

Using host proteins as input nodes, protein protein interactions (PPIs) to build the input graph and integrating data from multiple omics, Graph Convolutional Networks (GCNs) can address this problem classifying unlabelled host proteins according to their relation with the neighborhood and their associated feature vectors.

GCNs have already been proved successful in predicting cancer related genes with high accuracy and we applied the same principle to SARS-CoV-2 infection, integrating transcriptomics and proteomics data from A549 cells with the aim of finding novel host factors.

Currently our method already predicts as potential host factors several genes that are well known to be related to SARS-CoV-2 infection and are reported in several independent publications.

We apply explanation techniques, such as Layer-wise relevance propagation, to prioritize relevant predictions to be tested in CRISPR-Cas9 knockout screens. We believe that in the future the integration of more omics and the use of different PPIs will increase the overall graph homophily and consequently improve the prediction of SARS-CoV-2 host factors.

In addition, other model agnostic interpretation methods will be implemented for a comparative analysis, in order to improve model interpretation. We believe that novel predicted and experimentally validated host factors will aid the development of new antiviral drug therapies specific for SARS-CoV-2 infection.

Systems

Between viral targets and differentially expressed genes in viral infections: the sweet spot of disease mechanisms for therapeutic intervention

Carme Zambrana (Barcelona Supercomputing Center), Sam Windels (Barcelona Supercomputing Center), Noel Malod-Dognin (Barcelona Supercomputing Center) and Nataša Pržulj (Barcelona Supercomputing Center).

Abstract:

Viral infections continue to cause pandemics, highlighting the importance of uncovering their disease mechanisms to enable drug repurposing to treat them. In our previous work, we discovered that “common neighbours” (CNs), genes that directly connect viral interactors (VIs) and differentially expressed genes (DEGs) in the human interactome, are key to COVID-19 mechanisms and promising targets for drug repurposing. Here, we use our CN concept across different viral diseases to uncover disease mechanisms and repurpose drugs. For five well-studied viruses (Influenza, SARS-COV-2, SARS-COV-1, HIV and HCVM), we demonstrate that the most central CNs in the interactome (with high betweenness centrality) are involved in disease mechanisms. However, DEG data is unavailable for many diseases, making identifying CNs impossible. Therefore, we show that the most central genes connected to VIs significantly overlap with CNs, enabling us to perform a pan-viral analysis covering 13 viruses (8 without DEG data). On average, we predict 500 disease-related genes for each virus (validating 30% through gene-disease databases), from which 86 are shared across all viruses, showing our methodology’s capability to uncover viral “pan-disease” mechanisms. Our top-ranking prediction, PLEKHA4, regulates the Wnt signalling pathway, known to be involved in viral infections, including Influenza and Hepatitis C. On average, 23% of all predicted CNs and 55% of the top 10 are drug targets, providing potential drug repurposing candidates. Currently, we are working on predicting candidates for drug repurposing for the remaining CNs. Our CN concept is universal and can enable insight into viral and other infectious diseases.

Systems

Bioinformatic analysis of prescription patterns and drug combinations of patients under peritoneal dialysis treatment

Michail Evgeniou (Medical University Vienna), Fabian Eibensteiner (Medical University Vienna), Klaus Kratochwill (Medical University Vienna) and Paul Perco (Medical University Innsbruck).

Abstract:

Patients with end stage kidney disease receiving peritoneal dialysis (PD) suffer from a number of comorbidities requiring them to take an extensive amount of drugs regularly. Despite the fact that data on unwanted side effects or drug interactions are available, no systematic analysis on the synergistic effect of drug combinations in this patient population has been published. We therefore aimed systematically to investigate the impact of drug combinations on PD-associated mechanisms on a molecular level.

We generated a set of 630 drugs of PD-patients at the Medical University of Vienna. These drugs were assigned to their respective classes based on the anatomical therapeutic chemical (ATC) classification system. We created network-based mechanism of action (MoA) molecular models for these compounds through literature-derived drug-associated gene sets and protein-protein interaction networks. We then clustered individual drugs based on the similarity of their constructed MoA molecular models.

We shortlisted seven drug categories with a prescription frequency ten times above average, in particular proton pump Inhibitors, and drugs for hyperkalemia management. Based on the constructed MoA molecular models we identified six clusters using the Jaccard coefficient as correlation measure. We identified drug combinations being used in clinical practice in this patient cohort addressing different mechanisms, thus mechanistically being complementary, as well as combinations sharing target pathways.

This systematic analysis of drug combinations in the context of PD-related mechanisms and pathways has the potential to guide future therapeutic intervention in this patient population.

Systems

Biologically informed neural network identifies Unfolded Protein Response as key pathway in critical COVID-19

George Gavriilidis (INAB, CERTH), Stella Dimitsaki (INAB, CERTH), Fotis Psomopoulos (INAB, CERTH) and Vasileios Vasileiou (INAB, CERTH).

Abstract:

COVID-19 multi-omics have been thoroughly analyzed with machine learning; however, the inference of therapeutically actionable insights has been hampered by highly dimensional biological data and poorly interpretable Artificial Intelligence components (10.1136/bmjinnov-2020-000648). Prior biomedical knowledge can significantly enhance neural networks applied on multi-omics since it can constrain the model from exploring unnecessary hypothesis spaces (https://doi.org/10.1093/bib/bbab454). Considering the above, we designed COV-PASnet which combines pathway-associated sparse deep neural network (PASnet) with explainable Artificial Intelligence Shapley values (https://doi.org/10.1371/journal.pone.0231166). COV-PASnet was able to robustly demarcate critical from non-critical COVID-19 cases (AUC:92%, F1-score: 69%) when applied on 2 large plasma proteomic datasets (training MGH: https://doi.org/10.1016/j.xcrm.2021.100287, testing DC: https://doi.org/10.1016/j.cell.2020.10.037). Strikingly, one of the most predictive pathways via pathway layers' node ranking was the Unfolded Protein Response (UPR), mainly due to highly abundant Death Receptor 5 (DR5/TNFRSF10) (https://doi.org/10.1016/j.molcel.2017.06.017). Since the literature is scant for UPR signaling in coronaviruses, we next analyzed scRNA-seq data from COVID-19-patient-derived peripheral-blood mononuclear cells and discovered UPR-prone plasmablasts through Enrichr (GO:0036500, padj.<0.001) (doi: 10.1002/cpz1.90) and KEA3 (IRE1 kinase: 1st in critical COVID-19, 108th in non-critical cases based on MeanRank metric)(https://doi.org/10.1093/nar/gkab359). Overall, our herein work shows that our biologically informed neural network called COV-PASnet led to the identification of a previously unexplored tenet of critical COVID-19 in circulating plasmablasts called UPR signaling. Considering how important antibody secretion by plasmablasts is for COVID-19 clearance (https://doi.org/10.1016/j.cell.2020.08.025) and that persistent UPR via DR5 promotes apoptosis (10.1126/science.1254312), we believe that this pathway merits further pharmacological investigation, in the search of novel COVID-19 therapeutics.

Systems

Changing terpenoid biosynthesis in rice through synthetic biology

Oriol Basallo (Universitat de Lleida - IRBLleida), Rui Alves (Universitat de Lleida - IRBLleida) and Ester Vilaprinyo (Universitat de Lleida - IRBLleida).

Abstract:

Squalene and many high valued chemicals in the pharmaceutical, biotechnological, cosmetic, and biomedical industries belong to the terpenoid family. Biosynthesis of these chemicals relies on polymerization of the IPP and/or DMAPP monomers that are synthesized by plants in the cytosolic MVA and plastidic MEP pathways.

IPP/DMAPP are building blocks for developmental hormones, making it difficult to redirect IPP/DMAPP towards production of non-cognate plant chemicals without affecting plant viability. Developing plants for use as a platform to produce high value terpenoids is an important biotechnological goal that requires increasing their IPP/DMAPP production flux. Rice is a potential platform for achieving this.

As such, we created mutant rice lines with additional biosynthetic pathways for IPP/DMAPP production. These lines express three different versions of an exogenous MVA pathway in the plastid, in addition to the normal endogenous pathways. We collected data for changes in macroscopic and molecular phenotypes, gene expression, isoprenoids and hormone abundance in those lines.

We developed and analyzed data-centric, line specific, multilevel mathematical models that integrate all the data. These models connect the effects of variations in hormones and gene expression to changes in macroscopic plant phenotype and metabolite concentrations within the MVA and MEP pathways of WT and mutant rice lines.

Our models reveal how an exogenous IPP/DMAPP biosynthesis pathway affects the flux of biosynthesis of terpenoid precursors. They also allow us to quantify the effect of hormonal regulation on the alternative IPP/DMAPP biosynthetic pathways, enabling the prediction of macroscopic plant characteristics from molecular data.

Systems

Characterising Alternative Splicing Effects to Protein Interaction Networks with LINDA

Enio Gjerga (Section of Bioinformatics and Systems Cardiology, University Hospital Heidelberg, 69120 Heidelberg, Germany) and Christoph Dieterich (Section of Bioinformatics and Systems Cardiology, University Hospital Heidelberg, 69120 Heidelberg, Germany).

Abstract:

Alternative RNA Splicing plays a crucial role in defining protein function. However, despite its relevance, there is a lack of tools that characterize effects of splicing on protein interaction networks in a mechanistic manner (i.e. presence or absence of protein-protein interactions due to RNA splicing). To fill this gap, we present LINDA (Linear Integer programming for Network reconstruction using transcriptomics and Differential splicing data Analysis) as a method that integrates resources of protein-protein and domain-domain interaction, transcription factor targets, and differential splicing/transcript analysis to infer splicing-dependent effects on cellular pathways and regulatory networks.

We have applied LINDA to a panel of 94 shRNA knock-down experiments in HepG2 and K562 cells from the ENCORE initiative. Briefly, RNA-seq data after knock-down of RNA binding proteins were used to uncover splicing-dependent changes in the interactome on these two types of cells. For example, for HepG2 case studies, pathway sets related to Cell Cycle, FOXO Transcription and Cyclin D - CDK were the ones mostly affected by splicing.

LINDA has been implemented as an R-package and is available at https://dieterich-lab.github.io/LINDA along with examples and tutorials.

Systems

Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data

Daniel Dimitrov (UKHD) and Julio Saez-Rodriguez (UKHD).

Abstract:

The growing availability of single-cell data, especially transcriptomics, has sparked an increased interest in the inference of cell-cell communication. Many computational tools were developed for this purpose. Each of them consists of a resource of intercellular interactions prior knowledge and a method to predict potential cell-cell communication events. Yet the impact of the choice of resource and method on the resulting predictions is largely unknown. To shed light on this, we systematically compare 16 cell-cell communication inference resources and 7 methods, plus the consensus between the methods' predictions. Among the resources, we find few unique interactions, a varying degree of overlap, and an uneven coverage of specific pathways and tissue-enriched proteins. We then examine all possible combinations of methods and resources and show that both strongly influence the predicted intercellular interactions. Finally, we assess the agreement of cell-cell communication methods with spatial colocalisation, cytokine activities, and receptor protein abundance and find that predictions are generally coherent with those data modalities. To facilitate the use of the methods and resources described in this work, we provide LIANA, a LIgand-receptor ANalysis frAmework as an open-source interface to all the resources and methods.

Systems

Composing and organizing metabolic model perturbations and constraint systems with COBREXA.jl

Miroslav Kratochvíl (Luxembourg Centre for Systems Biomedicine, University of Luxembourg), St Elmo Wilken (Institute for Quantitative and Theoretical Biology, Heinrich Heine University Düsseldorf), Laurent Heirendt (Luxembourg Centre for Systems Biomedicine, University of Luxembourg), Reinhard Schneider (Luxembourg Centre for Systems Biomedicine, University of Luxembourg), Christophe Trefois (Luxembourg Centre for Systems Biomedicine, University of Luxembourg) and Wei Gu (Luxembourg Centre for Systems Biomedicine, University of Luxembourg).

Abstract:

COBREXA.jl is a constraint-based metabolic modeling and analysis (COBRA) toolbox for Julia language [1] that facilitates construction and running of large-scale analyses on HPC platforms. This poster elaborates the design features of COBREXA.jl: Allowing many different model representations, and providing a system of modifiers that are used to intuitively combine existing algorithms and model perturbations as "building blocks" over various model types. Apart from improving the efficiency in parallel processing, this allows the users to easily create complex novel workflows for meaningful use-cases.

We demonstrate how the approach simplifies implementation of complex analyses: On a model enriched with various constraints derived from multiomic measurements, we run a large-scale parallel virtual screening that estimates model viability and variability over many hypothetical model states. The available analysis extensions improve the correspondence of the simulation results with reality, benefiting applications in bioengineering and personalized medicine. In the poster, we formalize the software design patterns that enable this extensibility.

Additionally, we summarize our observations from automated construction of large compound models using COBREXA.jl. We identify specific interoperability deficiencies in the published model metadata that impair the feasibility of performing construction tasks automatically. The poster suggests several ways to reduce the ambiguity in metadata annotation that would further aid both the model reproducibility and the ability to easily create biologically accurate multi-organ and multi-organism models.

[1] Kratochvíl, Heirendt, Wilken, ... & Gu. (2022). COBREXA.jl: constraint-based reconstruction and exascale analysis. Bioinformatics, 38(4).

Systems

Computational prediction of time-course drug transcriptomic responses

Michio Iwata (Kyushu Institute of Technology) and Yoshihiro Yamanishi (Kyushu Institute of Technology).

Abstract:

Identifying the dynamic human cell line response to drug therapies is necessary to determine the time-course mode of drug action in medical and pharmaceutical research. Most drugs are bioactive compounds that interact with target proteins implicated in a disease of interest. As chemical perturbagens, they modulate the activity of biological systems for treating the disease. However, the effect of drug perturbations on the cellular system is not well time-dependently investigated; thus, there is an incomplete picture of their mode of action. In this study, we developed a novel computational method to predict time-course drug-induced transcriptomic responses of the cellular system. We performed dynamic sensitivity analyses with a mathematical model of the biological pathway, constructed drug-induced enzymatic sensitivity signatures, and predicted drug-induced gene expression values at new time points for which no data were observed. By conducting enrichment analyses related to biological pathways and gene ontology terms, the use of time-course drug-induced gene expression data enabled the detection of different biological functions in a time-dependent manner. Overall, the proposed method can be used to increase our understanding of drug-induced transcriptomic responses over time.

Systems

Construction of the cystic fibrosis biological network from the meta-analysis of transcriptomic studies

Matthieu Najm (Institut Curie, Mines ParisTech), Loredana Martignetti (Institut Curie, Mines ParisTech), Matthieu Cornet (Mines ParisTech, Institut Imagine, Hôpital Necker - Enfants Malades), Mairead Aubert (Institut Imagine, Hôpital Necker - Enfants Malades), Isabelle Sermet-Gaudelus (Institut Imagine, Hôpital Necker - Enfants Malades), Laurence Calzone (Institut Curie, Mines ParisTech) and Véronique Stoven (Institut Curie, Mines ParisTech).

Abstract:

Cystic Fibrosis (CF) is a disease caused by mutations in the gene encoding CFTR. However, the overall pathophysiology cannot be easily linked to the loss of the CFTR function alone. Our hypothesis is that CFTR belongs to a yet not fully deciphered protein network, whose functions are disrupted by the absence of CFTR, thus participating in some deleterious CF phenotypes.

Using a systems biology approach and a meta-analysis of transcriptomic studies, we built a consistent biological network summarizing the molecular disruptions of CF human airway epithelial cells (HAEC). The aim is to identify new potential therapeutic strategies, complementary to those rescuing CFTR. We applied and compared pathway-based algorithms on publicly available transcriptomic datasets. From this meta-analysis, we established the list of the most common differentially expressed genes (DEG) and biological pathways between CF and non-CF samples. Finally, from these lists and from Protein-Protein Interaction databases, we built the biological network summarizing CF HAEC.

While the DEGs can be very different between studies, the differentially expressed pathways are very consistent between studies of different tissue samples, cell cultures or techniques. We noticed that studies in which these pathways were not found have fewer secretory cells in CF samples than in non-CF samples. Finally, topological analyses of the network connecting these pathways revealed how CFTR is linked to the different CF phenotypes.

This study identified the possible contribution of secretory cells in disease phenotypes, but can also help determine which proteins to include in a Boolean model to simulate in silico experiments.

Systems

Context-specific investigations of combined miRNA effects with DIANA-miRPath v4.0

Spyros Tastsoglou (Hellenic Pasteur Institute/DIANA-Lab, Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly), Giorgos Skoufos (DIANA-Lab, Hellenic Pasteur Institute/Dept. of Electrical & Computer Engineering, Univ. of Thessaly), Marios Miliotis (Hellenic Pasteur Institute/DIANA-Lab, Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly), Dimitra Karagkouni (DIANA-Lab/Hellenic Pasteur Institute/Harvard Medical School, Dept. of Pathology, Beth Israel D

Abstract:

microRNAs (miRNAs) are negative post-transcriptional modulators shaping expression landscapes in (patho-)physiological states. The many-to-many relationship miRNAs and their targets share implicates miRNAs in most biological processes. Investigating combined miRNA functions against known pathways thus constitutes a challenge. Here, we present the latest version of DIANA-miRPath, an online platform that utilizes robust miRNA targets, pathways, terms and ontologies to conduct miRNA functional analysis.

We employed biased urn statistics to account for known biases in targeting resources and non-parametric testing options to enable, for the first time, detection of cell-types and tissues in which miRNA-targeted pathway members are expressed at lower levels. An internal library of resources to test against includes cell types from human single-cell atlases, GTEx tissues and TCGA cancer types. Additionally, a novel application is dedicated to identification of enriched pathways containing positively/negatively selected miRNAs and targets from CRISPR knockout screens, enabling selection-based analyses in phenotypic contexts.

miRPath v4.0 was designed into a modular form and upgraded, enabling analyses using millions of high-quality experimentally supported (TarBase, LncBase, microCLIP and miRTarBase) and predicted (microT-CDS and TargetScan) interactions against gene sets and ontologies from KEGG and Reactome pathways, GO-terms, MSigDB and PFam. Custom resources can also be provided in all modules in order to cater results on non-reference miRNAs (e.g., RNA-edited forms) or even non-model species.

Since 2009, miRPath has become a reference tool for interrogating miRNA regulatory activity. Its last version pushes the envelope, achieving cell-type-level resolution in context specificity, which is unprecedented in computational investigations of miRNA function.

Systems

Coupling of poro-aniso-hyperelastic and solute transport finite element models in a High-Performance Computing framework, for the study of Intervertebral Disc Degeneration

Dimitrios Lialios (Barcelona Supercomputing Center), Mariano Vazquez (Barcelona Supercomputing Center), Beatriz Eguzkitza (Barcelona Supercomputing Center), Eva Casoni (Barcelona Supercomputing Center) and Maria Paola Ferri (Barcelona Supercomputing Center).

Abstract:

Intervertebral Disc degeneration (IVDD) is the leading cause of lower back pain, affecting more than 10% of the global population. In silico modeling could contribute to the understanding of the dynamics of IVDD. Patient specific medical solutions can be explored by exploiting the high scalability of supercomputers, regarding both the number of cases simulated and the refinement level of the finite element (FE) mesh.

In this work, the highly parallelizable FE solver, Alya, developed by BSC is used to perform multiscale simulations to understand the mechanisms dictating IVDD. IVDs are known to exhibit poro-aniso-hyper-elastic behavior. The anisotropic and hyperelastic behavior are accounted for by combining the Holzapfel-Gasser-Ogden and modified neo-Hookean material models in a solid mechanics solver. Coupled with the later, a porous mechanics solver whose permeability adapts to the deformation of the solid matrix, is developed. Finally, oxygen, glucose and lactate represent the key nutrients involved in IVDD. Simulations aiming to study the changes in their concentration levels are performed by coupling a solute transport solver to the poro-mechanical one.

The present work is part of the HORIZON MSCA Disc4All project, aiming to produce multi-disciplinary tools applied to the study of IVDD. The requirement for integration is satisfied by adopting the BioBB framework. The framework allows for multiple instantiations of Alya, enabling the efficient simulation of cohorts, as well as the coupling with Physicell, an HPC ready Agent Based solver to create a holistic tool for the study of IVDD.

Systems

DEPI v3: A systems biology and artificial intelligence based patient stratification and drug positioning platform for neurodevelopmental disorders

Laura Pérez-Cano (Stalicla), Francesco Sirci (Stalicla), Igor Ariz-Extreme (Stalicla), Daniel Boloc (Stalicla), Sara Azidane (Stalicla), José Hidalgo (Stalicla), Rubén Sabido (Stalicla), Lynn Durham (Stalicla) and Emre Guney (Stalicla).

Abstract:

Neurodevelopmental disorders (NDDs) are a group of highly heterogenous and prevalent disorders characterized by abnormal brain development. These disorders remain an area of high unmet medical need due to the lack of specific pharmacological treatments addressing the core symptoms, especially social communication deficits. Most clinical trials in NDDs over the past decades have failed to show efficacy across patients recruited using behavior-based diagnosis despite bearing unrelated molecular pathophysiology. Here, we present the DEPI platform, STALICLA's data-driven solution to meet the biological stratification needs in complex NDD populations. DEPI is the first systems biology and multi-omics based, AI module driven platform in the NDD space for the development of precision medicine-based treatments. It uses curated NDD-risk catalogs to identify pathway-level perturbations associated to clinical observations. Furthermore, it combines supervised and unsupervised machine learning methods to characterize molecular signatures (genetic, transcriptomic and metabolomic) across individuals through mechanistic endophenotyping. Importantly, the platform facilitates matching of the right treatments to the right patients with NDDs. The platform has been proven clinically successful by: i) identifying a first clinically actionable subgroup of patients with Autism Spectrum Disorder (ASD Phenotype 1), and the corresponding tailored treatment, STP1, with outstanding and unprecedented target engagement results in a Phase 1b clinical trial; and ii) blindly recalling known drug responder patients with sensitivity and specificity values ~80% for patients with fragile X syndrome.

Systems

DrDimont: Explainable drug response prediction from differential analysis of multi-omics networks

Pauline Hiort (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Julian Hugo (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Justus Zeinert (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Nataniel Müller (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Spoorthi Kashyap (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Jagath C. Rajapakse (School of Computer Science and Engineering, Nanyang Technological University), Francisco Azuaje (Genomics England, London), Bernhard Y. Renard (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam) and Katharina Baum (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam).

Abstract:

Motivation: While it has been well established that drugs affect and help patients differently, personalized drug response predictions remain challenging. Solutions based on single omics measurements have been proposed, and networks provide means to incorporate molecular interactions into reasoning. However, how to integrate the wealth of information contained in multiple omics layers still poses a complex problem.

Results: We present DrDimont, Drug response prediction from Differential analysis of multi-omics networks. It allows for comparative conclusions between two conditions and translates them into differential drug response predictions. DrDimont focuses on molecular interactions. It establishes condition-specific networks from correlation within an omics layer that are then reduced and combined into heterogeneous, multi-omics molecular networks. A novel semi-local, path-based integration step ensures integrative conclusions. Differential predictions are derived from comparing the condition-specific integrated networks. DrDimont's predictions are explainable, i.e., molecular differences that are the source of high differential drug scores can be retrieved. We predict differential drug response in breast cancer using transcriptomics, proteomics, phosphosite, and metabolomics measurements and contrast estrogen receptor positive and receptor negative patients. DrDimont performs better than drug prediction based on differential protein expression or PageRank when evaluating it on ground truth data from cancer cell lines. We find proteomic and phosphosite layers to carry most information for distinguishing drug response.

Availability: DrDimont is available on CRAN: https://cran.r-project.org/package=DrDimont.

Systems

Estimating Strengths of Causal Interactions for Gene Regulatory Networks in Yeast

Adriaan Ludl (Universitet i Bergen), Tom Michoel (Computational Biology Unit, Department of Informatics, University of Bergen) and Mariyam Khan (University of Bergen).

Abstract:

Causal inference from genomics and transcriptomics data is a powerful approach for reconstructing causal gene regulatory networks. Instrumental variable methods (IV) use a local expression quantitative trait locus (eQTL) as a randomized instrument for a gene's expression level and assign target genes based on distal eQTL associations.

We are developing methods to estimate the strength of causal interactions between regulatory and target genes in cases where multiple genes share multiple eQTLs. Our method is based on multi-variate IV methods to solve the genomic linkage problem and disentangle the relative trans-effects of multiple genes with linked cis-regulatory sites.

We give an overview of the method and results of our first evaluation of this method on a dataset of 1012 segregants from a yeast (Saccharomyces cerevisiae) cross, and the YEASTRACT database of experimentally validated regulatory interactions between genes.

Systems

Ethanol-induced sex-based differences in the extracellular vesicles lipidome

Carla Perpiñá-Clérigues (UBB, Principe Felipe Research Center (CIPF) - School of Medicine and Dentistry, University of Valencia), José F. Català-Senent (Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF)), Susana Mellado (School of Medicine and Dentistry, University of Valencia - Department of Mol. and Cellular Pathology of Alcohol (CIPF)), Consuelo Guerri (Department of Molecular and Cellular Pathology of Alcohol, Príncipe Felipe Research Center (CIPF)), María Pascual (School of Medicine and Dentistry, University of Valencia - Department of Mol. and Cellular Pathology of Alcohol (CIPF)) and Francisco García-García (Bioinformatics and Biostatistics Unit, Príncipe Felipe Research Center (CIPF)).

Abstract:

Lipids represent essential components of extracellular vesicles (EVs), playing structural and regulatory functions. Importantly, lipidic dysregulation has been linked to several inflammatory and neurological disorders. Thus, exosome lipidomics is emerging as an innovative field for discovering novel lipid species with biomedical applications. Likewise, EVs isolated from adolescents exposed to alcohol intoxication demonstrated a sex-based difference in their microRNA profiles.

Accordingly, we applied a lipidomics computational strategy using R language programming in order to examine how acute ethanol intoxication affects the lipid composition of plasma EVs, differently by sex in adolescents and the involvement of the immune response.

After an exploratory analysis experimental groups (ethanol and control groups of females and males) were compared with a differential abundance analysis. Annotation of the lipids in their corresponding classes and class enrichment were carried out to evaluate the biological function. This strategy was performed in human subjects and WT and TLR4-KO mice. The latter to explore the role of the toll-like receptor 4 (TLR4) in the response.

We identified a higher enrichment of specific EV lipid species in human female adolescents (e.g., PA, LPC, unsaturated FA and FAHFA) than in males (e.g., PI). These lipid species participate in the formation, release, and uptake of EVs and the activation of the immune response; therefore, results suggest that female adolescents who binge drink alcohol also display increased levels of EV biogenesis and neuroinflammatory spread than males. Our findings also support the potential use of EV-enriched lipids as biomarkers of ethanol-induced neuroinflammation during adolescence.

Systems

Evaluation of machine-learning methods using diverse representations and feature selection techniques for cell type annotation in single-cell transcriptomics data

Hyojin Kim (Uniklinik RWTH Aachen), Rafael Kramann (Uniklinik RWTH Aachen) and Sikander Hayat (Uniklinik RWTH Aachen).

Abstract:

Multiple computational approaches have been developed for single-cell transcriptomics analysis. However, these methods use different input features and representations, which leads to heterogeneous interpretation. In order to reduce method-based discrepancies and find a uniform way for cell type annotation, we evaluated 3 major processes involved in automated cell-type annotation: 1) feature selection, 2) choice of underlying representation and 3) machine learning methods. For feature selection, genes were selected by 7 different ways based on highly variable score using Seurat (vst) and Symphony (pooled_vst), celltype-wise differentially expressed score by wilcoxon and MAST, cluster-wise probability, specificity and TF-IDF transformed score. The feature selected data were represented in 4 different ways using PCA, harmony, scVI and scANVI. Lastly, Random Forest (RF) and Multi-Layer Perceptron (MLP), and algorithms implemented in Seurat and Symphony, were applied for cell-type annotation. We tested ~70 combinations across the 3 processes using 244660 cells from PBMC, Kidney, Lung and Heart human/mouse datasets covering ~80 samples. For annotation evaluation in homogenous conditions, each dataset was evaluated separately by dividing into training (70%) and test data (30%) by stratified sampling while tuning parameters of models using 5 cross-fold validation. We found that vst (feature selection), scANVI (representation) and MLP (method), along with vst-PCA-MLP, pooled_vst-PCA-MLP and pooled_vst-scANVI-MLP showed top performances based on macro-F1 from 0.95 to 0.89, respectively. Additionally, we evaluated performance while training on reference datasets by predicting cell-types on completely unseen query datasets (F1 0.88). Our evaluation will be useful for standardising cell-type annotation across datasets.

Systems

Flexible nets to optimize antibody production in chinese hamster ovary cells

Teresa Joven (Department of Computer Science and Systems Engineering, University of Zaragoza), Jorge Lázaro (Department of Computer Science and Systems Engineering, University of Zaragoza), Jorge Júlvez (Department of Computer Science and Systems Engineering, University of Zaragoza), Nicole Borth (ACIB, Vienna. Department of Biotechnology, University of Natural Resources and Life Sciences, Vienna), Diana Széliová (Department of Analytical Chemistry, University of Vienna) and Jürgen Zanghellini (Austrian Centre of Industrial Biotechnology, Vienna. Department of Analytical Chemistry, University of Vienna).

Abstract:

Antibodies are therapeutic proteins with a multitude of applications in medicine such as the treatment of viral infections, different types of cancer and common diseases such as psoriasis and multiple sclerosis. Chinese Hamster Ovary (CHO) cells are the most broadly used cells for the manufacturing of antibodies because of their advantages compared with other mammalian cells. The current design of systems for the production of antibodies is mainly based on "trial and error" methods that manipulate CHO cells. This leads to high expenditure of time and money, in addition to obtaining suboptimal process performance. The use of mathematical models has the potential to greatly accelerate and facilitate the design and optimization of antibody production. Starting from a systematic and formal approach, the aim is to achieve an automatic design of the whole process that allows optimal productivity to be reached. To this end, we are developing both mathematical models and algorithms for the design and optimization of antibody manufacturing systems. The mathematical models are based on flexible nets, a novel modeling formalism able to accommodate uncertain parameters and nonlinear dynamics. Flexible nets are used to develop a comprehensive model that encompasses both the metabolic network of CHO cells and the dynamics of the bioreactor in which the cells are cultured. Thus, the model integrates macroscopic (dilution rate, substrate concentration, cell density, etc.) and microscopic variables (intracellular metabolic fluxes) simultaneously which enables a global optimization of the system.

Systems

GALLANT: A standardized workflow for multi strain Genome-scale metabolic modeling.

David San León Granado (Spanish National Center for Biotechnology) and Juan Nogales Enrique (Spanish National Center for Biotechnology).

Abstract:

Genome-scale metabolic models (GEMs) of bacteria are one of the most important drivers of recent advances in metabolic engineering and systems biology. However, single strain GEMs define the metabolic capabilities of the strain in question, thus limiting the ability to perform metabolic studies at the species level. To overcome this, here we present GALLANT, a standardized and reproducible workflow that returns multi-strain GEMs to analyze the unique capabilities of individual strains, but with the added value that it is also able to group GEMs according to their pan-genome to perform studies of metabolic potential at the species level. The main goal of GALLANT is to deliver output GEMs to the highest standards. To this end, GALLANT uses a high-quality, customized GEM database as a template and it combines it with other databases like BIGG and KEGG to return final GEM drafts. To reduce the need for subsequent manual curation, it features an optimized automatic gap-filling module for multi-strain modeling, while built-in GEM quality assurance tools such as MEMOTE guide the user in the process of refining final results. GALLANT's robust modular design supports parallel delivery of single strain GEMs and it is fully geared up to be run on high performance servers in order to scale up the process and reduce execution times considerably. We demonstrate that GALLANT provides a standardized, fast and accurate process to build multi-strain GEMs and, in doing so, it improves the current understanding of potential metabolisms.

Systems

Gene-essentiality based drug signature helps repurposing non-cancer drugs

Jing Tang (University of Helsinki) and Wenyu Wang (University of Helsinki).

Abstract:

Cancer drugs often kill cancer cells independent of their putative targets. The lack of understanding on drug-target interactions prevents biomarker identification and ultimately leads to high attrition in clinical trials. In this study, we explored whether an integration of loss-of-function genetic and drug sensitivity screening data could help identify the mechanisms of actions of drugs. We constructed a gene-essentiality drug signature by integrating loss-of-function genetic and drug sensitivity screening data. A machine learning model was developed, where the coefficients of all the genes were considered as the gene-essentiality signature of the drug. We compared the gene-essentiality signatures against structure-based fingerprints as well as the gene expression signatures in both supervised and supervised target predictions. We showed that the gene-essentiality signature can predict drug targets and their downstream signalling pathways. We then confirmed the validity of our framework in the PRISM dataset generated by the large-scale drug screening experiment. Finally, we predicted the targets for the non-cancer drugs in the PRISM screens that explain better their anticancer efficacy, which may pave ways for drug repositioning.

Systems

Guiding CAR T cell experimental design using probabilistic graphical models

Alice Driessen (IBM Research), Rocío Castellanos Rueda (ETH Zürich), Constance le Gac (IBM Research), Nicolas Deutschman (IBM Research), Sai Reddy (ETH) and Maria Rodriguez Martinez (IBM, Zurich Research Laboratory).

Abstract:

Chimeric antigen receptor (CAR) T cells are a promising new approach in cancer immunotherapy. Their safety, efficacy and phenotype depend heavily on the design of the CAR, which intracellular tail contains up to three domains derived from a range of cellular signalling receptors. Due to its modular design and the multitude of possible domains, there is a vast combinatorial space of CAR designs. There are substantial efforts to improve CAR T cells based on CAR designs. However, testing the effect of each CAR design experimentally is very resource and labour intensive, and not feasible beyond a few hundred different combinations. Therefore, we aim to predict T cell phenotypes upon expression of different CAR designs, informed by single-cell RNA sequencing of a small library of 30 CAR designs using combinations of five different domains. As a single CAR design is not associated with a single T cell phenotypes but rather a distribution over phenotypes, we use a probabilistic graphical model that links CAR domains to transcription factor activity and subsequently to phenotypes. The sequencing data is used as a prior for the transcription factor activity. Using this probabilistic model, we predict the transcription factor activity and the distribution over phenotypes of new combinations of CAR domains. This would give us insights into the transcriptional activity of new CAR domains and guide CAR T cell therapy.

Systems

Hybrid system based gene regulation models of circadian cycles

Lelde Lace (Institute of Mathematics and Computer Science, University of Latvia), Gatis Melkus (Institute of Mathematics and Computer Science, University of Latvia), Karlis Cerans (Institute of Mathematics and Computer Science, University of Latvia) and Juris Viksna (Institute of Mathematics and Computer Science, University of Latvia).

Abstract:

We present hybrid system based gene regulation models of mammalian circadian cycle and the results of model behaviour analysis. The models cover genes of two recently proposed biological models with correspondingly 5 gene and 3 gene core oscillators. The advantage of hybrid system based modelling framework is very limited model dependence on parameter values, which are described only at qualitative (comparative) level and only in cases when they explicitly affect models' observable behaviour. The models we propose here represent gene regulatory networks in terms of genes and gene products (proteins), protein binding sites, gene regulatory functions, and some comparative constraints on protein growth rates and binding site affinities. Although such models do not provide accuracy that can be achieved by differential equation based formalisms, they are less dependent from parameter fitting and can provide predictions on some biological aspects of gene regulation that are not dependent form the choice of particular parameter values.

With all the inherent limitations on quantitative accuracy, the developed hybrid system models can well replicate the gene expression periodicity and timing offsets consistently with the known data from experimental observations as well as with simulations that can be obtained from the currently proposed differential equation models. The work also includes developments of new analysis methods, in particular, for analysis of available trajectories in model state spaces and derivation of constraints that are needed for state transition trajectories to satisfy a number of specific properties that are required for biological feasibility of the models.

Systems

Identification of transcriptional network disruptions associated to drug resistance in cancer with TraRe

Charles Blatti (National Centre for Supercomputing Applications (NCSA) University of Illinois at Urbana-Champaign), Jesús de la Fuente (TECNUN School of Engineering of University of Navarra), Huanyao Gao (Mayo Clinic), Irene Marín Goñi (Centre for Applied Medical Research (CIMA) University of Navarra), Zikun Chen (University of Illinois at Urbana-Champaign), Sihai Zao (University of Illinois at Urbana-Champaign), Winston Tan (Mayo Clinic), Richard Weinshilboum (Mayo Clinic), Krishna Kalar

Abstract:

Background: Genes and their corresponding pathways form networks regulating various cellular functions that are critical in tumor development and response to therapy. Differentially expressed genes are the downstream effect of differences in global cell de-regulation in different phenotypic groups, and thus are unable to provide mechanistically grounded insights. Hence, the identification of significant changes in transcriptional network structures between different subgroups can help discover novel molecular diagnostics and prognostic signatures, and shed light on how the cell changes its behavior in response to drugs.

Results: To address this limitation, we developed a new computational method, termed TraRe (available in Bioconductor), that through sparse Bayesian models unravels transcriptional dynamics from RNA-Seq data. The unique value of TraRe is that it attempts to examine phenotypically-driven regulatory differences at three distinct levels: gene co-expression modules, specific regulons, and individual regulators. We applied TraRe to transcriptomic data from 46 metastatic castration-resistant prostate cancer (mCRPC) patients with Abiraterone (Abi) response clinical data and uncovered hampered immune response regulatory modules that showed strong differential regulation in Abi-resistant patients. These modules were replicated in an independent mCRPC study. Further, we experimentally validated key rewiring predictions and their associated transcription factors. Among them, ELK3, MXD1, and MYB were found to have a differential role in cell survival for Abi-response-specific settings, suggesting therapeutic targets for mCRPC.

Conclusions: Collectively, these findings shed light on the underlying regulatory mechanisms driving Abi response, demonstrating that our method is useful for the broad biomedical community to uncover complex regulatory dynamics from RNAseq data.

Systems

Identifying key multifunctional components shared by critical cancer and normal liver pathways via sparseGMM

Shaimaa Bakr (Department of Electrical Engineering, Stanford University), Kevin Brennan (Stanford Center for Biomedical Informatics Research, Stanford University), Pritam Mukherjee (Stanford Center for Biomedical Informatics Research, Stanford University), Josepmaria Argemi (Liver Unit, Clinica Universidad de Navarra, Hepatology Program, Center for Applied Medical Research), Mikel Hernaez (Center for Applied Medical Research, University of Navarra) and Olivier Gevaert (Stanford Center for Biomedical Informatics Research, Stanford University).

Abstract:

Despite the abundance of multi-modal data, suitable statistical models that can improve our understanding of diseases with genetic underpinnings are challenging to develop. Here we present SparseGMM, a novel statistical approach for gene regulatory network discovery. SparseGMM uniquely uses latent variable modeling with sparsity constraints regulators to learn gaussian mixtures from multi-omic data. By combining co-expression patterns with a Bayesian framework, sparseGMM quantitatively measures confidence in regulatory genes and uncertainty in target gene assignment by computing the entropy of a gene. We apply SparseGMM to liver cancer and normal liver tissue data and evaluate the discovered gene modules in an independent single-cell RNA-seq data set. We show that SparseGMM can recover diverse cancer and normal liver functions, as well as shared biology between cancer and normal. Further, sparseGMM identifies PROCR as a potential regulator and therapeutic target to inhibit angiogenesis, and PDCD1LG2 and HNF4A as important regulators of immune response and blood coagulation in cancer, respectively. Single cell analysis evaluation reveals that sparseGMM is also able to decouple myeloid and lymphoid biological processes in liver cancer. In agreement with molecular heterogeneity of cancer, we show that target genes have significantly higher entropy in cancer compared to normal liver; among these high entropy genes are key multifunctional components shared by critical cancer pathways, such as p53 and estrogen signaling. Finally, sparseGMM is a broad tool that will help the biomedical and computational biology community dissect the regulatory dynamics of transcriptomic data. It is available as a docker container to facilitate adoption.

Systems

Image-based simulation of morphogen gradient formation during zebrafish epiboly.

Justina Stark (Technische Universität Dresden; MPI-CBG Dresden; Center for Systems Biology Dresden), Rohit Krishnan Harish (Technische Universität Dresden; Center for Regenerative Therapies Dresden), Michael Brand (Technische Universität Dresden; Center for Regenerative Therapies Dresden) and Ivo F. Sbalzarini (Technische Universität Dresden; MPI-CBG, Dresden; Center for Systems Biology Dresden).

Abstract:

Graded concentration fields of morphogens are crucial during embryonic development, as they provide positional information for cell differentiation and tissue morphogenesis. Yet, how exactly these morphogen gradients are formed and maintained remains poorly understood. This is due to the abundance of factors involved, including localized sources and sinks, interaction with extracellular matrix molecules, and complex 3D tissue shapes.

Mathematical and computational models can help understand the role each factor plays. Many of the existing models, however, neglect the complex and dynamic tissue shape and reduce dimensionality to 2D or even 1D, although it is known that the 3D geometry has a significant impact on morphogen gradient formation.

We address this gap by deriving a realistic 3D computational model of a zebrafish embryo during epiboly from light-sheet microscopy volumes. We use this image-based model to simulate Fgf8a and Fgf3 gradient formation in the dynamically deforming extracellular space, as the embryo undergoes epiboly.

Our simulation numerically solves continuous reaction-diffusion equations in order to test whether the source-diffusion-sink model, which has been assumed so far, is sufficient to explain the shapes of the morphogen gradients observed. In particular, we show how the gradient is coupled to different factors, like tissue shape, production and degradation rates, location and surface area of source and sink, as well as the interaction with the extracellular matrix.

This presents a complete image-based simulation workflow that helps understand the mechanisms underlying morphogen gradient formation in the complex and deforming shapes of a growing embryo.

Systems

Implementation of cellular transport mechanisms within a multiscale simulation framework

Othmane Hayoun-Mya (Barcelona Supercomputing Center), Arnau Montagud (Barcelona Supercomputing Center), Miguel Ponce-de-Leon (Barcelona Supercomputing Center) and Alfonso Valencia (Barcelona Supercomputing Center).

Abstract:

Agent-based models (ABMs) constitute an advantageous approach for multicellular systems modeling. Nonetheless, cell and tissue ABMs usually do not include detailed transport mechanisms which interface between the agent or cell and its chemical microenvironment. Given that these mechanisms are a key biological process, we set to develop and implement a compendium of general transport mechanisms into a cellular ABM-based multiscale software: PhysiCell.

We consider six different mechanisms which encompass both passive (simple and facilitated diffusion through carriers and channels) and active transport (both primary and secondary). Transport mechanisms were described as systems of Ordinary Differential Equations (ODEs), conceptually grounded on Fick's Diffusion Laws and Michaelis-Menten kinetics.

We assessed density equilibrium dynamics in each mechanism through simple experiments, confirming saturable, nonlinear dynamics. We also performed unit testing on simple models to check for unexpected dynamics. Lastly, a PhysiCell-based template of drug resistance acquisition was then built by combination of three models. We argue that our models improve PhysiCell base code, which accounts only for linear transport dynamics. Furthermore, we made an effort to assemble modular models with interoperable units, which will facilitate connection with specific agent rules and will make them more user-friendly by simplifying model fitting with experimental data.

Systems

Improving xylose-fermenting yeast for 2G ethanol production via constraint-based modeling combined with omics data analysis

Lucas Carvalho (UNICAMP), Vítor Pereira (University of Minho), Guido Araujo (UNICAMP), Marcelo Carazzolle (UNICAMP), Gonçalo A. G. Pereira (LGE/Unicamp) and Miguel Rocha (UMinho).

Abstract:

The use of cellulosic ethanol (2G-ethanol), a technology that produced bioethanol from non-food lignocellulosic biomass as a renewable alternative, is a great ally for sustainability and the bioeconomy. In 2G ethanol, the consumption of xylose, one of the main products of the biomass hydrolysis process, can be done in two main pathways by the xylose-fermenting yeast: xylose isomerase (XI) and oxidative-reductive (OXR). Constraint-based modeling (CBM) is an approach widely used to study biochemical networks, in particular the reconstruction of the metabolic network through genome-scale metabolic models (GSMM) integrated with of omics data (transcriptomics, proteomics and metabolomics) through regulatory network algorithms and transcriptome-guided parsimonious flux analysis. In this work, we propose the development and application of methods based on omics data analysis, integrated with genome-scale metabolic models to understand the mechanisms behind the adaptative regulation of xylose-fermenting yeasts in 2G ethanol fermentation through XI and OXR pathways. For this, we apply parsimonious FBA algorithm with omics data and genome-scale regulatory model in different carbon sources (glucose and xylose) between these two pathways to understand their differences in the application of GSMM. We aim to build predictive models trained with omics data and metabolic simulation features that will enable us to learn how to improve the engineered yeast to improve ethanol production.

Systems

Inferring biologically relevant molecular tissue substructures by agglomerative clustering of digitized spatial transcriptomes with multilayer

Marco Antonio Mendoza Parra (French National sequencing Center / Genoscope) and Julien Moehlin (Inserm U1110 Strasbourg).

Abstract:

Developments in spatially-resolved transcriptomics (SrT) are providing means to interrogate organ/tissue architecture from the angle of the gene programs defining their molecular complexity. However, computational methods to analyze SrT data under-exploits the spatial signature retrieved within the maps. Inspired by contextual pixel classification strategies applied to image analysis, we have developed MULTILAYER, allowing to stratify SrT maps into functionally-relevant molecular substructures. For this, MULTILAYER applies agglomerative clustering strategies within contiguous locally-defined transcriptomes (herein defined as gene expression elements or gexels), combined with community detection methods over spatially-reconstructed gene networks for graph partitioning.

From the analysis of multiple published SrT datasets using MULTILAYER, we show it can be as performant as existing methods for detecting differentially expressed genes but in addition, it reveals highly-resolved tissue substructures reflecting anatomical classification but also tumorigenic progression signatures within cancer samples. Furthermore, we will illustrate the power of using MULTILAYER in a 3-dimensional context, i.e. via the analysis of multiple contiguous tissue sections, or the use of this strategy for revealing chromatin epigenetic status from the recent developments in spatial cut&tag technology.

Overall, MULTILAYER provides a digital perspective for the analysis of spatially-resolved transcriptomes and anticipates the application of contextual gexel classification strategies for developing self-supervised molecular diagnostics solutions.

Systems

Inferring pathway activities from gene expression data using perturbation transcription profiles of the LINCS-L1000 dataset

Bence Szalai (Turbine Simulated Cell Technologies Ltd.), Szabolcs Hetey (Turbine Simulated Cell Technologies Ltd.), Péter Szikora (Turbine Simulated Cell Technologies Ltd.), Kristóf Szalay (Turbine Simulated Cell Technologies Ltd.) and Dániel V Veres (Turbine Simulated Cell Technologies Ltd.).

Abstract:

Identifying altered pathway activities in diseases can provide mechanistic insight regarding disease process, thus can help identifying new therapeutic options. While pathway activities are defined on the level of protein interactions, they are frequently inferred from transcriptomics data. Classical methods infer pathway activity from the expression of pathway member genes. Recently several methods were developed to infer pathway activity a more data driven manner, from the expression of pathway regulated genes, however these methods are based on extensive manual curation of low-scale perturbation experiments. Large scale perturbation gene expression profiles, like LINCS-L1000 can help to build more robust and comprehensive pathway activity inference tools.

We developed TREX method (TuRbine EXpression based pathway activity) to infer pathway activities for 73 cancer related pathways. For each modelled pathway, we collected gene expression profiles from the LINCS-L1000 study, where pathway members were perturbed with CRISPR KO. Using these gene expression profiles, we built a multiple linear regression model to predict the activity changes of the investigated pathways. We benchmarked TREX on shRNA and drug perturbation profiles from LINCS-L1000, drug perturbation profiles from the PANACEA database and on a genome scale CRISPRi scRNAseq. Our benchmarking results showed that TREX can effectively recover the mechanism of action / pathway target of diverse perturbations.

TREX method allows pathway activity calculation from transcriptomics data for a large number of cancer related pathways, and contributes to the mechanistic understanding of perturbed cell states.

Systems

In-silico perturbation of transcription factors in a deep model of gene expression can predict tissue-specific regulation

Yuhu Liang (University of Copenhagen), Viktoria Schuster (University of Copenhagen), Thilde Terkelsen (University of Copenhagen) and Anders Krogh (University of Copenhagen).

Abstract:

Transcription factors (TF) are proteins participating in biological processes via binding to genomic DNA. TFs can regulate genes expression alone or with other proteins. To study the a TF, knock-out or knock-down experiments are usually used in the lab, and the gene expression is compared between the case and control. However, the experiments are done in cell cultures or laboratory animals. It is difficult to interpret the gene expression changes in these experiments, because they may be caused by secondary effects and further downstream processes. Here, we propose to do in-silico TF perturbations in a neural network trained to predict gene expression from TF expression with the aim to replace or supplement lab experiments with in-silico experiments. We trained a neural network using scaled expression values of TF genes as input, and predict the output expression of all protein coding genes. We selected a neural network after model searching, which has two hidden layers with 2,000 and 8,000 hidden units, respectively. Our data set consisted of 19,081 samples from GTEx. We perturbed a transcription factor in a given tissue by decreasing the expression value by 5% in samples from the selected tissue and measured the average relative change in expression of all the genes (q'). 2.5% of the genes at the two tails of the q'-distribution were selected as differentially expressed genes. The differential gene set could annotate hundred of GO terms, which should compare to few of GO terms can be annotated if we use random gene set. One transcription factor, HIFIA, for example, can regulate acetylation and phosphorylation. We found acetylcholine receptor binding and activation of phospholipase C activity as significant GO terms. Calcuium signaling and miRNA can increase/decrease HIF1A, we found around 20 significant calcium-related GO terms. We also found MIR21, a miRNA that can be regulated by HIF1A, and it was annotated in 29 GO terms. We show several other result indicating that changes in gene expression values after perturbing a TF are consistent with expectations and propose that this can function as a supplement to lab experiments.

Systems

Integrating and mining time-dependent single-cell RNA-seq data: Parkinson's disease application

Katarina Mihajlovic (Barcelona Supercomputing Center), Gaia Ceddia (Barcelona Supercomputing Center), Noël Malod-Dognin (Barcelona Supercomputing Center), Gabriela Novak (University of Luxembourg), Alexander Skupin (University of Luxembourg), Dimitrios Kyriakis (University of Luxembourg) and Nataša Pržulj (Barcelona Supercomputing Center).

Abstract:

Parkinson's disease (PD) is one of the most common neurodegenerative disorders that currently cannot be cured. Although its etiology is unknown, the development of PD is characterized by perturbations of many molecular pathways. The underlying mechanisms of PD could be unveiled by exploiting the information hidden in the molecular networks, combined with data obtained with the emerging single-cell RNA-sequencing (scRNA-seq) techniques. To fuse all these omics data, we develop a new non-negative matrix tri-factorization (NMTF)-based method, NetSC-NMTF, that integrates four molecular networks (protein-protein, genetic and metabolic interaction, and co-expression) with a scRNA-seq dataset of PD, or a control cell line, at a specific differentiation stage, to obtain eight condition-specific (control vs. PD samples at one stage) sets of gene embedding vectors (i.e., “gene embeddings”). To analyze these gene embeddings and uncover novel PD-associated genes, we propose a 2-step downstream methodology, which mines gene embeddings and identifies genes that group together with the known PD genes and prioritizes them according to the changes of their embedding positions between PD and control cell conditions (largest change first). This allows us to identify 226 PD-related gene predictions that are largely supported (48.7%) in the literature. Then, we validate our predictions using KEGG enrichment analysis, highlighting 17 PD-related pathways. Finally, we further investigate our top-20 prioritized genes, uncovering six already known PD genes and, more interestingly, 14 new and promising PD-associated genes, out of which eight are known drug targets, presenting potential candidates for developing novel therapeutic treatments for PD.

Systems

ISMARA: completely automated inference of gene regulatory networks from high-throughput data.

Piotr Balwierz (Biozentrum University of Basel), Mikhail Pachkov (Swiss Institute of Bioinformatics, Biozentrum University of Basel), Phil Arnold (Biozentrum University of Basel), Andreas Gruber (University of Konstanz), Mihaela Zavolan (Swiss Institute of Bioinformatics, Biozentrum University of Basel) and Erik van Nimwegen (Swiss Institute of Bioinformatics, Biozentrum University of Basel).

Abstract:

Understanding the key players and interactions in the regulatory networks that control gene expression and chromatin state across different cell types and tissues in metazoans remains one of the central challenges in systems biology. Our laboratory has pioneered a number of methods for automatically inferring core gene regulatory networks directly from high- throughput data by modeling gene expression (RNA-seq) and chromatin state (ChIP-seq) measurements in terms of genome wide computational predictions of regulatory sites for hundreds of transcription factors and micro-RNAs. These methods have now been completely automated in an integrated web server, called ISMARA (ismara.unibas.ch) that allows researchers to analyze their own data by simply uploading RNA-seq or ChIP-seq data-sets, and provides results in an integrated web interface as well as in downloadable flat form. For any data-set ISMARA infers the key regulators in the system, their activities across the input samples, the genes and pathways they target, and the core interactions between the regulators.

We believe that, by empowering experimental researchers to apply cutting-edge computational systems biology tools to their data in a completely automated manner, ISMARA can play an important role in developing our understanding of regulatory networks across metazoans.

Systems

Isoform-level quantification for single-cell RNA sequencing

Lu Pan (Karolinska Institutet), Huy Dinh (University of Wisconsin - Madison), Yudi Pawitan (Karolinska Institutet) and Trung Nghia Vu (Karolinska Institutet).

Abstract:

Estimation of isoform expression is still a challenging problem for bulk RNA-Seq, and the problem is even harder for scRNA-Seq. The 3' bias from high-throughput scRNA-Seq such as Chromium Single Cell 3' 10× Genomics generates substantial similarities between the read statistics from different isoforms, making isoform-level quantification highly challenging. Several groups have made a great effort to estimate isoform-level expression from full-length scRNA-Seq data. However, to date, no method has been developed for isoform expression quantification from 3' bias high-throughput scRNA-Seq data. To address these issues, we have developed a single-cell isoform quantification tool called Scasa, to estimate isoform expression from scRNA-Seq data by relying on the concepts of transcription clusters and isoform paralogs.

Systems

Learning and reasoning with Bayesian networks to reconstruct cell signalling networks and automate hypothesis generation from phosphoproteomics data

Magdalena Huebner (Queen Mary University of London), Conrad Bessant (Queen Mary University of London) and Pedro Cutillas (Queen Mary University of London).

Abstract:

The modulation of protein activities through the addition or the removal of phosphate groups, catalysed by kinases and phosphatases, plays an important role in cell signalling. Uncovering the circuitry of the resulting phosphorylation networks is crucial for understanding the underlying mechanisms of diseases such as cancer. One major challenge in reconstructing protein signalling networks from phosphoproteomics data is turning lists of differentially occupied phosphosites into causal regulatory relationships. Furthermore, considering all the relevant prior knowledge for hypothesis generation is time-consuming, leaving many of the available measurements unexplained and datasets under-analysed. Bayesian networks present a promising framework for modelling complex dependencies among signalling molecules. Their probabilistic nature allows accounting for the uncertainty inherent to most biological data, while their graphical properties facilitate their interpretation in the context of prior knowledge. Even so, attempts to apply Bayesian networks to protein signalling networks have been confined to individual pathways. In this study, we employ the Bayesian FGES search algorithm to learn the structure of proteome-wide signalling networks from LC-MS/MS-based phosphorylation data acquired for MCF-7 cells treated with 61 kinase inhibitors. We then combine Bayesian inference with logic modelling to generate explanations for observed data points and suggest hypotheses for further experimental testing.

Systems

Leveraging multi-omic network embeddings to attain mechanistic insights into acute-on-chronic liver failure

Matthias Fabian Meyer-Bender (Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health), Pia Erdoesi (Molecular Hepatology Section, Department of Medicine II, Medical Faculty Mannheim, Heidelberg University), Maren Büttner (Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health), Ioannis Deligiannis (Helmholtz Pioneer Campus, Helmholtz Zentrum München, German Research Center for Environmental Health), Rizqah Kamies (Helmholtz Pioneer Campus, Helmholtz Zentrum München, German Research Center for Environmental Health), Ersin Karatayli (Department of Medicine II, Saarland University Medical Center, Saarland University, Homburg, Germany), Matthias Ebert (Molecular Hepatology Section, Department of Medicine II, Medical Faculty Mannheim, Heidelberg University), Frank Lammert (Department of Medicine II, Saarland University Medical Center, Saarland University, Homburg, Germany), Christine von Törne (Research Unit Protein Science, Helmholtz Zentrum München, German Research Center for Environmental Health), Stefanie Hauck (Research Unit Protein Science, Helmholtz Zentrum München, German Research Center for Environmental Health), Celia P. Martinez-Jimenez (Helmholtz Pioneer Campus, Helmholtz Zentrum München, German Research Center for Environmental Health), Nikola Mueller (Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health), Steven Dooley (Molecular Hepatology Section, Department of Medicine II, Medical Faculty Mannheim, Heidelberg University), Michael Menden (Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health), Seddik Hammad (Molecular Hepatology Section, Department of Medicine II, Medical Faculty Mannheim, Heidelberg University) and Christoph Ogris (Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health).

Abstract:

Acute-on-chronic liver failure (ACLF) is a major complication in patients with chronic liver diseases, and its molecular dynamics remain elusive. Systems biology approaches aim to gain mechanistic insights by interpreting complex signaling networks, however this poses computational challenges. In order to address this, here we investigated transcriptomic and proteomic measurements from a previously established mouse model, and consecutively integrated the data into a multi-modal network via a method called KiMONo. Our methodology makes use of both experimental data and prior knowledge (such as protein-protein interactions), produces a customized network containing high quality interactions relevant to ACLF, and leverages node embeddings to facilitate downstream analyses. Concretely, we used DeepWalk to project the multi-omic network onto a latent space, thereby paving the way for vector-based analysis methods such as visualization, clustering, and subsequent pathway analysis. This enabled us to visualize and investigate network properties such as the heterogeneity of different omics layers. In addition, we used the epsilon-neighborhoods of selected nodes to obtain clusters, a pathway enrichment of which highlighted a variety of pathways that were previously linked to ACLF (e. g. complement and coagulation cascades (p=8.54e-50), fatty acid degradation (p=4.13e-21)), as well as several new ones.

Despite numerous challenges, our results showed that embedding-based methods are able to enhance multi-omic data integration pipelines. Representation learning is becoming increasingly important for the analysis of multi-modal networks, hence the exploration of its strengths and limitations is an important step to facilitate its adoption into mainstream analysis pipelines.

Systems

Leveraging systems biology and machine learning for automatic drug repurposing in the rare disease landscape

Marina Esteban (Clinical Bioinformatics Area, Fundación Progreso y Salud (FPS), Sevilla; IBiS), Carlos Loucera (Clinical Bioinformatics Area, Fundación Progreso y Salud (FPS), Hospital Virgen del Rocio, Sevilla, 41013, Spain), Maria Peña-Chilet (Clinical Bioinformatics Area, Fundación Progreso y Salud (FPS),Sevilla; BieR - (CIBERER); IBiS) and Joaquín Dopazo (Clinical Bioinformatics Area, Fundación Progreso y Salud (FPS), Sevilla; BieR - (CIBERER); IBiS; ELIXIR).

Abstract:

The understanding and treatment of rare diseases remain a challenge for the healthcare system. Merging the knowledge about disease mechanisms and drug action, together with arising machine-learning methodologies, we present a model to define the Disease-Map for automatic drug-targets prioritization.

First, for the Disease Map construction, we selected ORPHANET's well-known genes responsible for each rare disease. Then, using the KEGG pathway database and HiPathia R package, we constructed the Disease-Maps by extracting all the subpathways containing those genes. Next, we run HiPathia over each Disease-Map using the GTEx RNA-seq dataset. We selected approved known drug-target genes (KDTs) by the FDA or in late clinical trials from the Drugbank database, establishing the Disease-Map environment. Moreover, we implemented Multi-Output Random Forest regression, from GTEx expression data to the circuits activity values obtained with the HiPathia algorithm, to infer the effect of the selected KDTs over the activity regulation of the constructed Disease map. In order to degranulate a per-subpathway relevance, we used the SHapley Additive exPlanations.

From 3822 rare diseases with known ORPHANET genes, a total of 156 rare diseases with 3 or more genes in KEGG pathways database were selected. From a total of 1098 possible affected subpathways, the disease maps obtained vary from 1 to 409. Furthermore, the application of multi-output regression methodologies revealed relevant potential known drug targets for each disease that are, in most cases, functionally related to the rare disease phenotype. These findings would lead to a more efficient drug selection for rare disease patients' treatment.

Systems

Longitudinal analysis of biological age (Phenoage) in UKBiobank participants reveals key factors driving ageing trajectories

Laura Bravo (University of Birmingham), Victor Cardoso (University of Birmingham), John William (University of Birmingham), Dominic Russ (University of Birmingham), Samantha Pendleton (University of Birmingham), Furqan Aziz (University of Birmingham), Archana Sharma-Oates (University of Birmingham), Animesh Acharjee (University of Birmingham), Georgios Gkoutos (University of Birmingham) and Janet Lord (University of Birmingham).

Abstract:

Background: Age is the greatest single risk factor for many chronic diseases. The emerging premise is that the biological ageing process might be a common pathogenic factor, providing a novel therapeutic target. Interestingly, key mechanisms proposed to drive biological ageing converge on inflammation-related pathways. Previous studies have aimed to objectively define biological age using different biomarkers. In this study we used one of these biomarkers (PhenoAge) longitudinally, together with disease development and multimorbidity to identify key genetic and lifestyle factors underlying the transition from a healthy to unhealthy state.

Methods: PhenoAge measurements were calculated from blood marker information in UKBiobank (UKBB) participants using clinical read codes and repeat UKBB assessment visits. A PhenoAge trajectory was then fitted and compared to chronologic age creating four categories describing the participant's health trajectory: Healthy remaining Healthy; Healthy becoming Unhealthy; Unhealthy becoming Healthy; Unhealthy remaining Unhealthy. Lifestyle and health-related factors were analysed alongside information on 60 chronic-inflammatory diseases. Pairwise-disease association, Dynamic Time Warping (DTW) and rule-based associations were measured per category for all diseases. GWAS was performed for each category and results were linked using public databases (DisgeNET and TreeWas).

Results: Older males with higher BMI, comorbidities, harmful lifestyle (i.e smoking) and lower parental lifespan were seen in the unhealthy categories. GWAS results showed an enrichment of metabolic related processes with APOE, FEN1 and HSPA1L/B/A as significant genes. Disease clusters found through DTW revealed increased multimorbidity patterns in colitis related diseases or inflammatory arthropathies with the genes mentioned above, linking these pathologies together.

Systems

Machine learning for functional reconstruction and analysis of Biochemical interactions controlling Colorectal cancer phenotypes

Victor Olorunshola (Queens University Belfast), Ian Overton (Queens University Belfast), Sandra Van Schaeybroeck (Queens University Belfast), Alan Murphy (Imperial College London) and Erola Pairo-Castineira (The Roslin Institute, University of Edinburgh, United Kingdom).

Abstract:

The 60% expected increase in deaths from Colorectal Cancer (CRC) by 2030 (1), which is already the fourth leading cause of cancer-related deaths in the world raises a pressing need for better clinical tools for CRC patient stratification and treatment. The advances in functional genomics technologies have resulted in a 'data deluge' and enabled determination of a 'parts list', including comprehensive characterisation of the protein coding genome. A current challenge is understanding the roles of individual component parts and how they work together in the cell to determine biological function in health and disease. My research applies supervised machine learning for context-specific biological network reverse engineering, integrating polyomics data, focused on CRC to capture molecular features specific to tumour stage and subtype. A major research activity involved developing a pipeline that integrates functional genomics data (transcriptome and methylation) and extracts features using supervised machine learning to establish a context-specific genome scale biochemical interactions network. Networks for the following CRC contexts- stage II, stage III, CMS2 and CMS3 have been successfully produced. This predicted context specific network approach is been explored as a precision medicine discovery tool capable of providing mechanistic insights underlying CRC stages/subtypes and also inform risk stratification.

Reference

1. Arnold, M. et al. (2017) 'Global patterns and trends in colorectal cancer incidence and mortality.', Gut. BMJ Publishing Group, 66(4), pp. 683–691. doi:10.1136/gutjnl-2015-31091

Systems

Mathematically mapping the network of cells in the tumor microenvironment

M. van Santvoort (Eindhoven University of Technology), Ó. Lapuente-Santana (Eindhoven University of Technology), F. Finotello (University of Innsbruck), W.L.F. van der Hoorn (Eindhoven University of Technology) and F. Eduati (Eindhoven University of Technology).

Abstract:

The tumor microenvironment (TME) is composed of malignant and non-malignant cells which communicate with each other determining tumor development and response to treatment. Unfortunately, the TME consists of many intertwined elements that cannot be measured fully, thus locking away the information contained in the TME. To address this problem, we propose a mathematical model to reconstruct the TME on a patient level by integrating several sources of information, so that fine-grained information about their tumor can be extracted and analyzed.

Our model inputs consist of patient-specific information about: 1. Cellular composition derived from bulk RNA sequencing (RNAseq) data by combining different deconvolution methods (9 cell types including tumor cells, immune cells and cancer associated fibroblasts); 2. Scoring of 1784 literature-derived ligand-receptor interactions based on their expression. Additionally, the model uses information about which ligands and receptors can be expressed by each cell type based on single-cell RNAseq data. With these inputs we construct an ensemble of possible communication networks by matching ligands and receptors uniformly at random to cells known to secrete them. By analyzing the commonalities in this ensemble, we are able to identify “signature properties” of the TME.

To validate the model we focused on two pathway types: direct communication between cells and triangles of three communicating cells. Preliminary analysis shows that the model can extract communication pathways that distinguish known TME subtypes and responders from non-responders to immune checkpoint inhibitors. Additionally, triangle pathways seem to be the most relevant, whilst being consistent with important direct communication pathways.

Systems

Metabolic Atlas - exploration and visualization of metabolic networks for model organisms

Mihail Anton (National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Chalmers University of Technology), Nanjiang Shu (National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University), Ingrid Hyltander (National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University), Malin Klang (National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University), Per Johnsson (National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University), Shan Huang (Department of Biology and Biological Engineering, Chalmers University of Technology, Sweden), L. Thomas Svensson (Department of Biology and Biological Engineering, NBIS, SciLifeLab, Chalmers University of Technology, Sweden) and Jens Nielsen (Chalmers University of Technology, BioInnovation Institute).

Abstract:

Models in systems biology are used to understand biological processes by facilitating data interpretation, analysis, and prediction. By joining thousands of reactions, metabolites, and genes into large metabolic networks, genome-scale metabolic models (GEMs) have become valuable tools to study metabolism. The incorporation of enzymatic parameters is, for example, one of the ways to further improve the prediction of GEMs, in addition to the integration of omics data.

Metabolic Atlas, through the web platform freely available at https://metabolicatlas.org, presents the entire content of open GEMs for easy browsing and analysis. This is achieved through both tabular and map views (2D and 3D). In addition, Metabolic Atlas aims to meet the needs of the community through the development of specific tools and features through iterative releases.

The history of Metabolic Atlas began with a focus on the human metabolic model (Pornputtapong et al., 2015). The present website has been re-developed from the ground up, following open-source standards, by first integrating Human1, an integration and extensive curation of the most recent human metabolic models (Robinson et al., 2020), and Yeast8, a consensus metabolic model for S. cerevisiae (Lu et al., 2019). Following a database redesign and the addition of a performant 3D viewer, 5 more models have been integrated (Wang et al., 2021). A new major release is under active development, aiming to further facilitate computational applications and metabolic engineering.

Systems

Metabolic Reprogramming in Rheumatoid Arthritis Synovial Fibroblast: a Hybrid Modeling Approach

Sahar Aghakhani (GenHotel (Université Paris-Saclay), Lifeware (Inria Saclay Île-de-France)), Sylvain Soliman (Lifeware (Inria Saclay Île-de-France)) and Anna Niarakis (GenHotel (Université Paris-Saclay), Lifeware (Inria Saclay Île-de-France)).

Abstract:

Rheumatoid Arthritis (RA) is an autoimmune disease characterized by a highly invasive pannus formation consisting mainly of synovial fibroblasts (RASFs). This pannus leads to cartilage, bone and soft tissue destruction in the affected joint. RASFs' activation is associated with metabolic alterations resulting from dysregulation of extracellular signals transduction and gene regulation. Deciphering the intricate mechanisms at the origin of this metabolic reprogramming may provide significant insight into RASFs' involvement in RA's pathogenesis. In this work, we present the first hybrid RASF-model: a combination of a cell-specific qualitative regulatory network with a global metabolic network. We developed our regulatory network as a boolean model covering the main signaling and gene regulation pathways in RASFs along with their impact on metabolism. Regarding the metabolic network, we used the MitoCore model, a curated constraint-based model for simulating human central metabolism. Both models are encoded in the Systems Biology Markup Language (SBML) standard for biological computational models. Our automated hybrid framework exploits the regulatory network's asynchronous stable states as additional constraints on the metabolic network's components. Subsequent metabolic flux analysis allows to assess RASFs' regulatory outcomes' impact on its central metabolic flux distribution. Simulations of our hybrid RASF-model reproduce the experimentally observed metabolic reprogramming induced by signaling and gene regulation in RASFs. They also enable further hypotheses on the potential reverse Warburg effect in RA. RASFs may undergo metabolic reprogramming to turn into "metabolic factories", producing high levels of energy-rich fuels and nutrients for neighboring demanding cells.

Systems

miRarmature: a time series analysis pipeline for paired miRNA and RNA-seq data reveals new regulatory dynamics

Ranjan Kumar Maji (Goethe University and Uniklinikum Frankfurt), Ariane Fischer (Institute of Cardiovascular Regeneration, Goethe University, Frankfurt), Martin C. Simon (Wuppertal University), Stefanie Dimmeler (Institute of Cardiovascular Regeneration, Goethe University, Frankfurt) and Marcel H. Schulz (Goethe University and Uniklinikum Frankfurt).

Abstract:

Time series analysis of small and long RNA sequencing profiles poses challenges for data integration along with deriving novel biological hypothesis. We built a time series microRNA analysis pipeline, miRarmature, that explores the miRNA arm 'switching' and 'shifting' dynamics and its effects on modulating gene expression patterns in time series data. The pipeline enables us to (a) compute a cell/tissue type specificity index to recover and rank the miRNAs that have the most variance in expression specific to the cell/tissue type, (b) detect and evaluate significant miRNA 5'/3'-arm switching dynamics, (c) associate miRNA arm dynamics with significant changes in target gene expression at the specific time points, and finally (d) identify and visualize the time

points where miRNA arm shift/switch events result in a significant change in target gene expression. We implemented our pipeline for bulk, long RNA-seq (for gene and miRNA hairpin abundance estimates) and small RNA-seq (for miRNA estimates) pro-files and applied it to cell types measured in Acute Myocardial Infarction (AMI) model in mouse hearts at 6 different days. Using miRarmature, we explored cell type-specific miRNAs in the heart that show differential arm dynamics and significant changes in target expression along time. Our analysis reveals not only the diverse mature miRNA arm modulation in response to AMI stress, but also reflects that the processing and generation of mature miRNA 5'/3' arms upon injury, acts in a cell-type and time-dependent manner.

Systems

Modeling the metabolic switch in Ewing Sarcoma from single-cell transcriptomic profiles

Marianyela Petrizzelli (Institut Curie) and Andrei Zinovyev (Institut Curie).

Abstract:

Ewing sarcoma, a genetically stable pediatric soft tissue and bone tumor, represent a model system to study. The major genetic cancer driver consists of a chromosomal translocation that fuses the 5' end of EWSR1 gene to the 3' end of a member of the ETF family, most commonly the FLI1. EWSR1-FLI1 activity induces changes at the molecular and phenotypic levels and raises questions on how it alters cell metabolism. Here we explore the metabolic pathways activated upon EWSR1-FLI1 activity by combining constraint-based modeling and scRNA-seq from patient-derived xenografts (PDXs) (Aynaud et al. 2020). We develop a pipeline to construct context-specific genome scale models (csGEMs) from scRNA-seq data using the GIMME algorithm (Becker and Palsson, 2008). The pipeline performs scRNA-seq dimensionality reduction, clustering, and cluster annotation; it constructs csGEMs for each cluster and performs flux balance analysis to predict metabolic fluxes. Assuming that clusters with the same annotation across multiple PDXs have a similar distribution of metabolic fluxes, we infer the best parameters that discriminate clusters at the metabolic flux level by supervised kernel PCA. We show, using a thermodynamically curated reduction of Recon3 model (Masid et al. 2020), that the set of reactions that discriminate the group of clusters annotated as EWS-high to other group of clusters involve serine, glycine, tyrosine, tryptophan, and cysteine amino acids; those that discriminate EWS-high from EWS-low clusters, among others, involve glutathione and glutamine metabolites. Our results show consistent results with known metabolic pathways activated upon EWSR1-FLI1 action and suggest new ones associated to it.

Systems

Molecular cross-talk communication between Intermuscular Adipose Tissue and Skeletal Muscle under progressing Insulin Resistance

Amare Wolide (HMGU) and Dominik Lutter (HMGU).

Abstract:

Background: Intermuscular Adipose Tissue (IMAT) seems associated with the insulin sensitivity (IS) of Skeletal Muscle(SM) (Sachs S, et al, 2019). Our study was initiated to understand the molecular cross-talk communications in these tissues in progressive Insulin Resistance (IR).

Methods: For this study, a total of 42 from 3 different IS study groups were recruited. Overnight fasting blood and muscle biopsy from Vastus lateralis were taken for clinical and RNA-Seq studies. Additional variables such as age, gender, fat mass, fat-free mass, and body mass index were studied using One-way ANOVA. Correlation and differental expressions analysis was employed to dissect the molecular crosstalk between IMAT and SM. Signficant gene pairs from both analysis were mapped to the Sender-Receiver Database(SRD), and matching pairs were selected for further analysis. Our SRD was curated after a rigorous database search and text mining of protein-coding genes. In this context, the sender represents a protein that could transmit information from the signaling cell to the target cell where the receiver protein is located.

Results: We discovered increased communication from IMAT to SM in the augmented IR. Consequently, a higher number of edge, node, and degree centrality were observed in T2D and Overweight groups compared to Lean groups. Some hits predict IS and displayed IR signatures.

Conclusion: Observed gene rewiring in the communication could be a possible target for further study. Moreover, we will inspect genetic interactions, expression, and functional differences between groups and, eventually, demonstrate potential targets to modulate SM insulin resistance.

Systems

Multi-modal based predictions of vaccine-induced immune responses.

Fabio Affaticati (University of Antwerp), Abdulkader Azouz (Université libre de Bruxelles), Esther Bartholomeus (University of Antwerp), Benson Ogunjimi (University of Antwerp), Stanislas Goriely (Université libre de Bruxelles), Kris Laukens (University of Antwerp) and Pieter Meysman (University of Antwerp).

Abstract:

The determinants of seroconversion outcome for vaccine administration remain poorly understood, despite several prior studies which have sought to couple baseline immunological characteristics to vaccine response. It is likely that these determinants are dependent on both the vaccine composition as well as the target population. In this study, we built classification methods to analyse potential vaccine-response biomarkers within transcriptomics data.

Transcriptomics data of several cohorts prior to and following injection with the Pfizer/BioNTech SARSCoV2 mRNA vaccine were generated. On these features, we built a random forest classification model that shows encouraging results on predicting vaccination response, within and across different population types. In line with the literature, innate immune system and inflammation related genes emerged among the most relevant features for classification at time points prior to vaccination.

In accordance with prior studies, we show that it is possible to use baseline transcriptomics data to predict part of the vaccine response. The remainder of the vaccine response is likely determined by other (clinical) factors, such as cellular populations and immune repertoire responsiveness. It is our aim to complement the current model with these factors to create a more comprehensive view of what drives vaccine response. The desirable outcome of such research is an effective multi-modal predictor ready for real-case scenarios.

Systems

Multimodal Synthetic Lethality Prediction in Cancer

Yasin Tepeli (Delft University of Technology (TUDelft)), Colm Seale (Delft University of Technology (TUDelft)) and Joana P. Gonçalves (Delft University of Technology (TUDelft)).

Abstract:

Synthetic lethality (SL) denotes the joint dependence of a cell's survival on the functions of a pair of genes, which can be exploited by targeted therapies to selectively kill tumour cells. Since exhaustive lab screening of SL gene pairs is impractical, computational prediction of promising pairs is key to guide experimental follow-up. However, current SL prediction methods: (i) rely exclusively on context-specific sources of molecular and clinical data, which may be unavailable or sparse for some (cancer) tissue types (e.g. cell line gene dependency, mutation); (ii) often make predictions driven by the topology of known SL relationships, and are thus sensitive to selection bias prevalent in these data.

We propose MMSL, a multimodal SL prediction model that overcomes these limitations by considering both context-specific features based on cancer cell line and tissue omics, and context-free gene relationships derived from amino acid sequence and protein-protein interactions. These datasets are incorporated into the MMSL model through a combination of early and late integration using regularized random forests.

Compared to competing models, MMSL performed most consistently high in single-cancer SL prediction across 8 cancer types. It was also amongst the most robust to selection bias, in a series of experiments aiming to prevent train and test samples from following the same bias. Amino acid sequence had the most influence on MMSL predictions, while other data sources were still noteworthy depending on cancer type. Finally, literature and survival analyses provided corroborating evidence for SL gene pairs that were newly predicted by MMSL.

Systems

Multi-Omics Visible Drug Activity prediction, interpreting the biological processes underlying drug sensitivity

Luigi Ferraro (Federico II, university of Naples), Giovanni Scala (Federico II, University of Naples) and Michele Ceccarelli (Federico II, University of Naples).

Abstract:

Cancer is a genetic disease resulting from the accumulation of genomics alterations in living cells. Large scale genomics studies have been instrumental to understand the recurrent somatic genetic alterations within a cell and for the characterization of their functional effects in transformed cells. One of the main challenging questions in this field is how to exploit all these molecular information to identify therapeutic targets and to develop personalized therapies, understanding which molecular features influencing sensitivity to drugs.

Machine learning models are able to exploit multi-modal screening datasets to develop predictive algorithms useful to associate omics features with response. The basic approach is to use the data from these screenings to train a machine learning “black box" model that predicts the 50% inhibitory concentration (IC50) of a drug from the multi-omics profile of a cell line, without the possibility to interpret the biological mechanisms underlying predicted outcomes and the exploitation of the unbalanced nature of the data.

In order to address these limitations we propose a Multi-Omics Visible Drug Activity prediction (MOViDA) neural network model that extends the visible network approach incorporating functional information in terms of pathway activity from gene expression and copy number data into a neural network.

We have identified which pathways and drug features are good predictors for high sensitivity of a cell line to a drug. This explanation is the basis to hypothesize drug combinations, cell editing and properties of new drugs aimed at the identification of cell vulnerabilities.

Systems

Multiscale model of the different modes of invasion

Marco Ruscone (Institut Curie), Arnau Montagud (Barcelona Supercomputing Center), Philippe Chavrier (Institut Curie), Olivier Destaing (Institute for Advanced Biosciences, University Grenoble Alpes), Andrei Zinovyev (Intitut Curie), Emmanuel Barillot (Institut Curie), Vincent Noel (Institut Curie) and Laurence Calzone (Intitut Curie).

Abstract:

Mathematical models of biological processes can be represented as complex networks of signaling pathways. They describe molecular regulations inside different cell types, such as tumor cells, T cells, or macrophages. These models mainly focus on intracellular information, however they often omit to describe the spatial organization of cells and their interactions with the microenvironment.

We have developed a model of tumor cell invasion within PhysiBoSS, a multiscale framework which combines agent-based modeling and continuous time Markov processes applied on Boolean networks. With this model, we aim at studying the different modes of cell migration through an extracellular matrix. To do so, we consider both spatial information obtained from the agent-based simulation and intracellular regulation obtained from the solutions of the Boolean model.

Each simulated cell is an agent that interacts with other cells and the microenvironment. They can move and reproduce, they have phenotype properties, a cell-cycle progression and different death models. The stimuli collected by each cell are represented as input of the intracellular signaling network, while the outputs correspond to the different behaviours that a cell can assume -such as proliferation, invasion and death.

The model integrates the impact of gene mutations with environmental conditions and allows the visualization of the results with 2D and 3D representations. The model reproduces single, collective and trail migration processes, and is validated on already published experiments on cell invasion. In silico experiments can be performed on these invasive set-ups to search for possible targets that can block the tumoral phenotypes.

Systems

Network analysis of Aicardi-Goutières syndrome-related genes - a systems biology approach

Gerda Cristal Villalba Silva (MD Anderson Cancer Center) and Shiaw-Yih Lin (MD Anderson Cancer Center).

Abstract:

Aicardi-Goutières syndrome (AGS, OMIM #225750) is an autosomal recessive inflammatory encephalopathy initiated by loss of self-tolerance resulting in the production of autoantibodies. The study of gene candidates to improve therapeutics in AGS using bioinformatics is gaining momentum. We used data from GSE193711, GSE57353, and GSE135652 from human cell lines to generate the differentially expressed gene list. We used edgeR to analyze the RNA-seq data and limma for microarray data. Enrichment pathway analysis was conducted with pathfindR, with the KEGG database. We constructed the protein interaction network with STRING. Topological analysis was performed in Cytoscape, using MCODE and Cytohubba plugins. We found 232 down-regulated genes, and 425 up-regulated genes. We identified several biological processes, like TNF pathway, Focal adhesion, MAPK, Calcium signaling, Rap1, cAMP, NF-kappa B, Cell cycle, and Toll-like receptor signaling pathways. The network was composed of 421 nodes and 1068 edges. Regarding the clustering analysis, we found 6 clusters. The top hub genes were CXCL1, CXCL8, E2F1, LEF1, PTGS2, TCF7, and VCAM1. Systems biology may help us understand the mechanisms underlying autoimmunity and the pathophysiology in the AGS.

Systems

Network medicine-based gene prioritization in Intervertebral Disc Degeneration (IDD)

Francesco Gualdi (Universitat Pompeu Fabra), Janet Piñero (Universitat Pompeu Fabra) and Baldomero Oliva (Universitat Pompeu Fabra).

Abstract:

Intervertebral disc degeneration (IDD) is a highly multifactorial disease in which environmental factors, lifestyle and genetics play a role 1. In order to interpret the complex interactions of genes and molecules that are involved in the development of complex conditions in recent years, network-based approaches have been successfully applied to integrate multi-omics data such as protein expression, protein interaction and signaling pathways. The aim of these methods is to detect functional modules that could reflect the biological underpinnings underlying the disease in order to better understand the mechanisms that lead to its development and eventually identify new biomarkers and treatments.

Here, we applied GUILD 2, a network-system approach based on guilt-by-association underlying the human interactome, to propagate the signal of IDD-associated genes. With this approach we detected a module related to IDD and prioritized new gene products that could be involved in the catabolic processes involved in IDD. By the analysis of the top 1% scoring genes we found a significantly connected module reflecting disease-relevant pathways.

In conclusion, our approach showed that in silico gene-disease prioritization methods were successfully applied to detect genes involved in biological pathways relevant to IDD that could be used as biomarkers or applied to infer potential drug targets to revert this condition.

1. Kos N, Gradisnik L, Velnar T. A Brief Review of the Degenerative Intervertebral Disc Disease. Med Arch. 2019;73(6):421-424.

2. Guney E, Oliva B. Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization. PLoS ONE. 2012;7(9).

Systems

Network-based driver identification with GoNetic

Simon Isphording (Ghent University), Giles Miclotte (Ghent University) and Kathleen Marchal (Ghent University).

Abstract:

Network-based analysis is a popular strategy to aid in data interpretation. Gene lists generated by in-house omics experiments are typically overlayed with an a priori interaction network in order to identify subnetworks of connected genes that are highlighted in the input dataset. Such overlay provides an increased mechanistic insight in the processes active in the profiled cells. In this work we present a versatile network-based data analysis method GoNetic, based on probabilistic pathfinding. In contrast to previous network-based approaches, GoNetic can explicitly exploit node and edge properties of the underlying network prior. By exploiting node properties (such as the level of confidence one has in the measurements of the node), the search space can be constrained in a biologically motivated way. GoNetic is versatile in its applicability: it has been applied succesfully to the analysis of expression data [1], mutational data and integration of mutation data with expression data in the context of gene driver identification in microbes [2] and cancer [3]. In addition because of its high performance implementation, it can handle large datasets (e.g. large tumor cohorts).

[1] Seyed Rahmani, R., et al. (2021). Genome-wide expression and network analyses of mutants in key brassinosteroid signaling genes. BMC genomics, 22(1), 1-17.

[2] Swings, T. et al. (2017). Network-based identification of adaptive pathways in evolved ethanol-tolerant bacterial populations. Molecular biology and evolution, 34(11), 2927-2943.

[3] de Schaetzen van Brienen, L., et al. (2021). Network-Based Analysis to Identify Drivers of Metastatic Prostate Cancer Using GoNetic. Cancers, 13(21), 5291

Systems

Novel systemic approach using comorbid phenotype clusters to propose disease-causing genes in undiagnosed patients

José Córdoba-Caballero (Universidad de Málaga; INiBICA), Pedro Seoane-Zonjic (Universidad de Málaga; CIBERER; IBIMA-Plataforma BIONAND), James R. Perkins (Universidad de Málaga; CIBERER; IBIMA-Plataforma BIONAND), Elena Rojano (Universidad de Málaga; IBIMA-Plataforma BIONAND) and Juan A. G. Ranea (Universidad de Málaga; CIBERER; IBIMA-Plataforma BIONAND; INB ELIXIR).

Abstract:

One of the goals of precision medicine is to accurately diagnose patients with genetic diseases. However, this is not always possible because we often do not know which genes are associated with the disease or their function. To address this problem, in this work we use a systemic approach in which we analyse information on diseases, genes and pathological phenotypes to find the genetic cause of these diseases using information from OMIM, Orphanet, MONDO and the Human Phenotype Ontology (HPO). We built different disease-phenotype networks to calculate associations between phenotypes and find which ones co-occur. Once we obtained the comorbid phenotype pairs, we grouped them into comorbid clusters. We verified these clusters by comparing them with the phenotypes of known genetic diseases. Finally, the clusters that could be verified by at least one disease were used to measure the closeness of the phenotype-associated genes to the disease-causing gene in the interaction network. The human genes were ranked using the distance in this interaction network and the ranking of the known disease-causing genes was obtained. Our preliminary results show that our methodology can rank known disease genes within the top 1-2% of the lists. With this methodology, we can propose potential disease-causing genes in undiagnosed patients with complex genetic diseases.

Systems

Numerical approaches for the rapid analysis of prophylactic efficacy against HIV

Lanxin Zhang (Robert Koch Institute), Junyu Wang (Robert Koch Institute) and Max von Kleist (Robert Koch Institute).

Abstract:

HIV remains a major public health threat. Currently, neither a cure, nor an efficient vaccine are available. However, antiretroviral drugs have been used successfully to prevent HIV infection. An important method for HIV self-protection

is pre-exposure prophylaxis (PrEP). To improve PrEP, many next-generation regimens, including long-acting formulations, are currently under investigation. However, the identification of parameters that determine prophylactic efficacy from clinical, ex vivo or in vitro data is extremely difficult. Clues about these parameters could prove essential for the design of next-generation PrEP compounds.

Mathematical models that integrate pharmacological, viral- and host factors are frequently used to complement our knowledge about prophylactic efficacy of antiviral compounds. Stochastic simulation methods are currently the gold standard for estimating prophylactic efficacy from these models. However, to obtain meaningful statistics, many stochastic simulations need to be conducted to accurately determine the sample statistics. To remedy the shortcomings of stochastic simulation, we developed a numerical method to directly compute the efficacy of arbitrary prophylactic regimen in a single run, without the need for sampling. Based on several examples with dolutegravir (DTG) -based short- and long-term PrEP, as well as post-exposure prophylaxis, the correctness of this new method and its outstanding computational performance is demonstrated. For example, a continuous 6-month prophylactic profile is computed within a few seconds on a laptop computer.

We envision that the approach can greatly expand the scope of analysis with regards to estimating prophylactic efficacy, by allowing to analyse the long-term effect of prophylaxis, as well as performing sensitivity analysis.

Systems

Omic fold changes clustering and network inference to study the radiation response of endothelial cells

Polina Arsenteva (Université de Bourgogne), Vincent Paget (IRSN), Olivier Guipaud (IRSN), Fabien Milliat (IRSN), Hervé Cardot (Université de Bourgogne) and Mohamed Amine Benadjaoud (IRSN).

Abstract:

More than 200000 patients undergo radiotherapy in France every year. Similarly to other treatments, it may induce adverse side effects for healthy tissues situated close to the irradiated tumor. It is thus of substantial importance to study and compare different modes of radiotherapy that vary in dose, volume, energy, etc. with a goal of selecting such that minimize the potential undesirable consequences. This work focuses on the response of endothelial cells, key actors in the appearance of radiation adverse effects. Specifically, we study the expression of genes originating from transcriptomic in-vitro datasets that were collected for several time points under irradiated and non-irradiated conditions. The goal is to determine a small number of the most representative behavior types among all considered genes, as well as to identify potential biological pathways linked to the response to radiotherapy. The quantity of interest is radio-induced fold change: a measure of irradiation effect represented by the difference between the two experimental conditions over time. To achieve this, we propose a new approach based on modeling fold changes as random variables, and a new distance that allows to account for uncertainties and correlations between variables. We designed a computationally efficient procedure performing simultaneous clustering and alignment of fold changes' random estimators. This procedure provides insight into regulatory pathways connecting genes with different behavior types with respect to their response. Finally, based on the obtained information, a gene network is inferred, followed by network analysis which allows to draw a comparison between different modes of radiotherapy.

Systems

Patient stratification reveals the molecular basis of comorbidities

Beatriz Urda-García (Barcelona Supercomputing Center (BSC), Barcelona, Spain), Jon Sánchez-Valle (Barcelona Supercomputing Center (BSC), Barcelona, Spain), Rosalba Lepore (Barcelona Supercomputing Center (BSC), Barcelona, Spain) and Alfonso Valencia (Barcelona Supercomputing Center (BSC), Barcelona, Spain and ICREA, Barcelona, Spain).

Abstract:

Epidemiological evidence shows that some diseases tend to co-occur; more exactly, certain groups of patients with a given disease are at a higher risk of developing a specific secondary condition. Despite the considerable interest, only a small number of connections between comorbidities and molecular processes have been identified. Indeed, previous studies analyzing disease similarities using molecular information have been able to capture interesting examples but failed to recapitulate the many comorbidities identified at the medical level, being unable to provide a general interpretation for them.

Here we present a new approach to generate a disease network that uses the accumulating RNA sequencing data on human diseases to significantly match half of the known comorbidities, providing plausible biological models for them. Of note, more than 95% of the captured comorbidities (e.g. ulcerative colitis and colorectal cancer) show immune system involvement.

Additionally, since patient-specific patterns are frequently observed at the epidemiological level, we introduce the concept of meta-patients as groups of patients from a given disease with a similar expression profile. The meta-patients reveal new disease interactions that were masked at the disease level, explaining 64% of the known comorbidities. This result points to the importance of patient stratification in the study of comorbidities and adds further evidence to the existence of molecular mechanisms behind an unprecedented proportion of known comorbidities.

(All the information is accessible at https://doi.org/10.1101/2021.07.22.21260979 and the results can be visually explored at http://disease-perception.bsc.es/rgenexcom/)

Systems

Predicting EDC mode of action from toxicogenomics data using EDTox

Arindam Ghosh (University of Eastern Finland), Amirhossein Sakhteman (University of Eastern Finland), Raghavendra Mysore (University of Eastern Finland), Einari Niskanen (University of Eastern Finland), Thomas Darde (Eurosafe), Pierre Daligaux (Eurosafe), Christophe Chesné (Eurosafe), Jorma Palvimo (University of Eastern Finland) and Vittorio Fortino (University of Eastern Finland).

Abstract:

Endocrine disrupting chemicals (EDCs) are a class of chemicals that have the potential to alter the normal functioning of the endocrine system by mimicking, blocking, or interfering with the hormones of the system. Even though many of the potential effects of exposure to these compounds are well known, their precise mode of action (MoA) is still not predictable. EDTox offers an easy-to-use web interface (http://edtox.fi/) and standalone R-Shiny application for training classifiers for prioritization of chemicals based on their endocrine disruption potential and identification of the possible MoAs. EDTox utilizes known chemical-gene interactions together with toxicogenomic data driven gene networks to train a binary classifier. The use of generalised linear models (GLM) in EDTox provides a way for identification of the features that are responsible for distinguishing the EDCs from non-EDCs and can be used for predicting the possible MoAs of the EDCs. Currently, EDTox includes classifiers based on toxicogenomics data from Open TG-Gates, DrugMatrix and LINCS that can be used for compiling the ED probability scores for chemicals by entering the list of genes that interact with the chemical. Here, we use gene expression data from HepG2 and HepaRG cell lines exposed to EDCs to train new classifiers and identify their potential MoAs. In particular, we observed that the effects of EDCs are mediated through different pathways under different conditions and that they also vary between doses.

References:

[1] Sakhteman, A et al. (2022). Bioinformatics, 38(7), pp.2066-2069.

Systems

Prediction Of Shared Intratumor Transcriptional Heterogeneity From Bulk Cancer Transcriptomic Data

Agnieszka Kraft (ETH Zurich, Switzerland; University Hospital Zurich, Switzerland; Swiss Institute of Bioinformatics (SIB), Switzerland) and Valentina Boeva (ETH Zurich, Switzerland; Swiss Institute of Bioinformatics (SIB), Switzerland; Institut Cochin, Inserm U1016, France).

Abstract:

In recent years, there has been growing evidence of the association between tumor cell composition and patients' survival. The majority of works studying the composition of tumor samples from bulk data (i.e. CIBERSORT) focus on enumerating different types of normal cells representing tumor microenvironment (TME); only recently, the BayesPrism method offered a possibility to infer malignant cell composition from bulk samples. However, BayesPrism relies on the availability of appropriate cancer single-cell references, which makes it applicable only to a limited number of cancer types.

Here, we propose a fully unsupervised method to identify and enumerate shared malignant cell subpopulations using tumor bulk RNA-seq data. It deconvoles bulk signal into respective components, identifies malignant- and TME-subpopulations, and measures shared intratumor transcriptional heterogeneity index based on Shannon entropy of malignant states' proportions. We characterize the identified malignant components by association with patients' survival, enriched pathways, and driver mutations. Our method was validated using single-cell RNA-seq data from glioblastoma, esophageal, lung, and colorectal cancer patients. We applied our method to 29 TCGA datasets. We found epithelial-to-mesenchymal transition and proliferation to be the major drivers of intratumor heterogeneity on a pan-cancer scale. In ten cancer types, our heterogeneity index showed a significant association with overall patients' survival. Moreover, we identified a number of genes (i.e. SOX2, BRAF, KRAS, and TP53) whose copy number or mutation profiles were highly correlated with the identified malignant states. Together these results validate our method and prove its feasibility to study shared intratumor transcriptional heterogeneity.

Systems

Reactome disease association overlay

Eliot Ragueneau (EMBL-EBI), Deng Chuan (Chongqing University of Posts and Telecommunications), Krishna Tiwari (EMBL-EBI), Chuqiao Gong (EMBL-EBI), Guilherme Viteri (Featurespace) and Henning Hermjakob (EMBL-EBI).

Abstract:

The Reactome Pathway Knowledgebase [https://reactome.org], an Elixir core resource, provides manually curated molecular details across a broad range of physiological and pathological biological processes in humans, including both hereditary and acquired disease processes. The processes are annotated as an ordered network of molecular transformations in a single consistent data model. Those processes are visually represented as diagrams, which can now be enriched by overlaying pairwise associations from other resources associated with their participant molecules. These associations can optionally be characterised by a score which can be used as a filter within the user interface. This functionality is currently used for Protein-Protein Interactions from IntAct[1] and Gene-Disease associations from DisGeNET[2], but can be extended to any type of binary associations involving biological molecules in Reactome, for example pathway modulators like antibodies or drug-like molecules. Sets of pairwise associations like those provided by DisGeNET can also be used as direct input to Reactome's pathway over-representation analysis, for example providing one-click enrichment analysis of the genes found to be associated with one disease: https://reactome.org/overlays/disgenet

[1] Gillespie M, Jassal B, Stephan R, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Research. 2022 Jan;50(D1):D687-D692.

[2] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020 Jan 8;48(D1):D845-D855.

Systems

RedRedundance: A web app to summarize the results of functional enrichment analyzes for one or multiple gene lists

Mireia Ferrer (Statistics and Bioinformatics Unit, Vall d'Hebron Research Institute), Esther Camacho (Statistics and Bioinformatics Unit, Vall d'Hebron Research Institute), Berta Miró (Statistics and Bioinformatics Unit, Vall d'Hebron Research Institute), Angel Blanco (Statistics and Bioinformatics Unit, Vall d'Hebron Research Institute) and Alex Sanchez (Statistics and Bioinformatics Unit, Vall d'Hebron Research Institute; GRBio).

Abstract:

Biological significance analysis (also known as functional enrichment analysis) is a useful technique for high-throughput data interpretation. Given a list of features (i.e. genes) resulting from an (omics) experiment, enrichment analysis allows identifying the functional categories that are statistically overrepresented among a list of genes. Such categories are typically derived from functional annotations (the Gene Ontology), pathway databases (KEGG or Reactome) or other resources. However, enrichment analysis often results in large lists of overrepresented sets that contain a large amount of redundancy and interdependencies between them, thus making its interpretation difficult without getting lost in the inherent noise of biological processes. Here, we present a web (shiny) app that implements state-of-the-art existing tools such as simplifyEnrichment R package [1] and EnrichmentMap Pipeline [2], to facilitate the comparison and summarization of enrichment analysis results across different databases and gene lists. With this app, the user can easily explore and filter the results of gene set enrichment analyses, choose among different algorithms for clustering and annotating the individual gene sets into larger groups and aggregate their statistics into a more meaningful unit that brings meaning back to biological significance studies.

1. Gu Z & Hübschmann D (2022). Simplify enrichment: A bioconductor package for clustering and visualizing functional enrichment results. Genomics, proteomics & Bioinformatics.

2. Reimand J et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nature protocols.

Systems

scRNA-Seq based drug repurposing for targeting alveolar regeneration in idiopathic pulmonary fibrosis

Anika Liu (University of Cambridge), Joo-Hyeon Lee (University of Cambridge), Andreas Bender (University of Cambridge) and Namshik Han (University of Cambridge).

Abstract:

Idiopathic pulmonary fibrosis (IPF) is a chronic lung disease, which affects around three million people worldwide and is associated with a poor prognosis of 2.5 to 5 years survival after diagnosis. It is characterized by impaired regeneration from recurrent injury to the alveolar epithelium resulting in progressive lung scarring. Recently, distinct intermediate cell states were identified during the regeneration process after bleomycin injury in the murine lung. Chronic inflammation was found to prevent the differentiation of AT2 cells into mature AT1 cells, leading to an accumulation of the intermediate population. Importantly, an aberrant basaloid cell population was newly identified in IPF lungs, which shares similar transcriptional signatures with these intermediate cell states. We hypothesized in this work that inducing the transition from aberrant basaloid cells to AT1 in IPF promotes lung regeneration. To this end, we characterized the intermediate population to AT1 cell transition signature using multiple recently generated single-cell RNA-Seq datasets on murine bleomycin injury and IPF patients. We then matched this signature to drug perturbation signatures retrieved from the LINCS database to identify the most suitable candidates for drug repurposing. Compound classes identified in this analysis include glucocorticoids, kinase inhibitors (e.g. targeting Src), as well as other structural classes of compounds. We plan to test the effect of selected compounds on alveolar regeneration in mouse-derived organoid models. Overall, we show how scRNA-Seq data can potentially be leveraged for drug repurposing by enabling better characterization of cell transitions which can subsequently be used for signature matching.

Systems

Simulation of ground truth interaction networks from microbial community model

Ada Rossato (Department of Information Engineering, University of Padova, Padova, Italy), Marco Cappellato (Department of Information Engineering, University of Padova, Padova, Italy), Nora Nikoloska (Department of mathematics "TULLIO LEVI-CIVITA", University of Padova, Padova, Italy), Giacomo Baruzzo (Department of Information Engineering, University of Padova, Padova, Italy) and Barbara Di Camillo (Department of Information Engineering, University of Padova, Padova, Italy).

Abstract:

A microbial community is characterised by several relationships established between microbes. Consumption and production of metabolites affect resource availability and thus determine the different microbial interactions developed in the community.

High throughput metagenome sequencing techniques have allowed an in-depth study of complex microbial community. Indeed, many bioinformatics tools have been developed to infer microbial interaction network starting from sequencing count data, such as correlation-based and regression-based tools. However, there is no consensus about the best approach to use, thus a systematic evaluation of their performance is needed.

In this work, we develop a new simulation framework that, starting from recently developed microbial population dynamics simulator [1], defines a ground truth interaction network. In particular, we derive interactions by considering resource consumption and production rates of the different microbes, thus taking into account resource competition and cross-feeding relationships. Once population absolute abundances are obtained, we simulate the sequencing process exploiting metaSPARSim [2], a sequencing count data simulator. This simulator exploits a multivariate hypergeometric model to simulate technical variability resembling sequencing count data.

Simulated data with underlying ground truth network can be given in input to network inference tools to perform a comprehensive evaluation. The overall simulation framework produces synthetic sequencing count data that reflect i) biologically realistic dependency across microbial species and ii) typical characteristics of real data. In this way we provide a useful simulation pipeline to benchmark different network inference methods on common ground truth simulated networks.

[1] Marsland et al., 2020.

[2] Patuzzi et al., 2019.

Systems

Single-cell multi-omics heterogenous multilayer network to infer gene regulatory mechanisms

Ina Maria Deutschmann (Institut de Biologie de l'Ecole Normale Supérieure (IBENS-CNRS)), Rémi Trimbour (Institut de Biologie de l'Ecole Normale Supérieure (IBENS-CNRS)) and Laura Cantini (Institut de Biologie de l'Ecole Normale Supérieure (IBENS-CNRS)).

Abstract:

About 30x1012 cells in the human body carry a copy of the genetic blueprint. Cell-specific functioning necessitates the regulation of genes mainly via transcription factors (TF) binding to promoter and enhancer regions. Usually, within single-cell investigations, gene expression (scRNA-seq) data alone is used to infer gene regulatory networks (GRN). However, their performance remains limited. In recent years, single-cell multi-omics data became increasingly available through technological advancement. More recently, methods using gene expression (scRNA-seq), genome accessibility (scATAC-seq), and prior knowledge (known TFs with their motifs) to infer GRNs have been proposed.

We hypothesized that keeping data modalities separate within one mathematical framework may yield better predictions. Thus, we developed a multilayer network framework that uses multi-omics single-cell data to predict gene regulation. Our multilayer network is heterogeneous, i.e., nodes represent different entities - genes, peaks, and transcription factors. In our application, each modality represents a layer, i.e., three layers representing genes, peaks, and regulatory entities in the form of TFs. TF-peak and peak-gene links connect the layers sequentially.

Our method outperforms existing approaches in the number of correctly predicted TF-gene links using publicly available scRNA-seq and scATAC-seq data of human and mouse embryonic stem cells. Moreover, our framework provides additional information with respect to existing GRN inference tools allowing enhancer-gene relationship prediction. Valuable to various fields, a single-cell multi-omics heterogenous multilayer network can support investigations of specific diseases such as cancer and deepen our understanding of cell development.

Systems

spongEffects: ceRNA modules offer patient-specific insights into the miRNA regulatory landscape

Fabio Boniolo (Technical University of Munich), Markus Daniel Hoffmann (Technical University Munich), Norman Roggendorf (Technical University of Munich), Bahar Tercan (Institute for Systems Biology), Jan Baumbach (University of Hamburg), Mauro Antônio Castro (UFPR), A. Gordon Robertson (BC Cancer Genome Sciences Centre), Dieter Saur (Technical University of Munich) and Markus List (Technical University of Munich).

Abstract:

Cancer is one of the leading causes of death worldwide. Despite significant improvements in prevention and treatment, mortality remains high for many cancer types. Hence, innovative methods that use molecular data to stratify patients and identify biomarkers are needed. Promising biomarkers can also be inferred from competing endogenous RNA (ceRNA) networks that capture the gene-miRNA-gene regulatory landscape. Thus far, the role of these biomarkers could only be studied globally but not in a sample-specific manner. To mitigate this, we introduce spongEffects, a novel method that infers subnetworks (or modules) from ceRNA networks and calculates patient- or sample-specific scores related to their regulatory activity. We show how spongEffects can be used for downstream interpretation and machine learning tasks such as tumor classification and for identifying subtype-specific regulatory interactions. In a concrete example of breast cancer subtype classification, we prioritize modules impacting the biology of the different subtypes. In summary, spongEffects prioritizes ceRNA modules as biomarkers and offers insights into the miRNA regulatory landscape. Notably, these module scores can be inferred from gene expression data alone and can thus be applied to cohorts where miRNA expression information is lacking.

Systems

Stochastic Model of Intra-Tumor Heterogeneity (SMITH)

Adam Streck (Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC)), Tom Kaufmann (Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC)) and Roland F. Schwarz (Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC)).

Abstract:

Cell-based simulations are a popular method for investigating intra-tumor heterogeneity and genome evolution during tumour growth. However, tracking individual cells in the size of a palpable tumour (billions of cells) is computationally expensive, and most methods thus represent groups of cells (demes, glands, patches) embedded in a lattice. This means that the models create only a simplified abstraction of the population with rigid, non-biological limitations of the lattice. We argue that the particular feature of lattice-based models is that they create implicit spatial constraints on the cell growth resulting in Darwinian selection. However, these constraints can be also expressed explicitly in terms of algebraic geometry and enforced even on non-spatial, well-mixed models.

To demonstrate this claim we have created a well-mixed, confined model of tumour growth of a solid, spherical tumour with fitness altering mutations. Our model introduces a novel mechanic, so-called confinement, that limits the cell turnover in the tumour to its outer shell of a certain width. We show that, when paired with fitness increase of mutations, confinement is sufficient to introduce the Darwinian selection to tumour growth and that different confinement values lead to different spatial dynamics, ranging from pure surface growth to full volume growth. We further show how a wide range of clonal dynamics naturally emerges from the combination of fitness increase and confinement.

Our model is implemented in the SMITH simulation tool. Due to its computational efficiency, SMITH can simulate a real-sized tumour of around ~2cm in diameter (~1 billion cells) in seconds.

Systems

STOCHASTIC SIMULATION OF SARS-COV-2 SPREADING AND EVOLUTION ACCOUNT FOR WHITHIN HOST DYNAMICS (SIMPLICITY).

Pietro Gerletti (Robert Koch Institute), Max Von Kleist (Robert Koch Institute), Sofia Paraskevopoulou (Robert Koch Institute), Matthew Huska (Robert Koch Institute) and Stephan Fuchs (Robert Koch Institute).

Abstract:

Non-pharmaceutical interventions played a crucial role in containing and slowing down the SARS-CoV-2 pandemic, especially before vaccines became available, and continue to be an important part of the efforts to curb the spread of the virus. Aside from the suffering that the pandemic is still causing, it led to unprecedented advance in biomedical knowledge on viral pathogenesis and evolutionary dynamics. For the first time, concerted, large-scale national pathogen sequencing efforts are in place and the gathered SARS-CoV-2 genomic data allows to track viral evolution at the population level. However, viral evolution not only depends on the population level transmission dynamics, but also on the within-host viral dynamics and their interplay. The approaches usually used to designate variants, reconstruct phylogenies, or to identify transmission clusters are applied to real world data, which is limited by our data-gathering possibilities. Hence, models that can generate ground truth data, which links the spreading dynamics and the evolution of the virus, are urgently needed to (but not only) provide validation datasets. In this work, we propose a susceptible-infected-recovered compartment model that accounts for intra-host viral dynamics and evolution. The approach allows to follow the evolutionary development of SARS-CoV-2 as it spreads and to generate transmission and evolution trees that can be used to validate the approaches commonly used to analyse real world data. Moreover, it can be used to test hypotheses regarding the emergence of novel variants, such as the role of (e.g. immuno-suppressed) long-shedding individuals in the emergence of these variants.

Systems

Systematic Analysis of Alternative Splicing in Time Course Data of SARS-Cov2 infection development using Spycone

Chit Tong Lio (University of Hamburg), Zakaria Louadi (University of Hamburg), Amit Fenn (University of Hamburg), Jan Baumbach (University of Hamburg), Olga Tsoy (University of Hamburg), Tim Kacprowski (TU Braunschweig and Hannover Medical School) and Markus List (TU Munich).

Abstract:

Alternative splicing (AS) drives protein and transcript diversity and is known to play a role in many diseases. The exact mechanisms controlling the AS machinery are currently insufficiently understood. During disease progression or organism development, AS may lead to isoform switches (IS) that follow temporal patterns. Several IS genes occurring at the same time point could reflect the co-regulation of AS for such genes.

We propose Spycone, a splicing-aware systematic framework for time-course data analysis. Spycone clusters genes and isoforms with similar temporal expression patterns. For isoform level analysis, we developed a novel IS detection algorithm that studies changes in total isoform abundance across time-course. Spycone couples the time-course clustering analysis with downstream analysis such as network enrichment and gene set enrichment analysis for functional interpretation. To evaluate the performance of Spycone, we implemented a novel approach for simulating time-course data. We demonstrate the performance of Spycone and TSIS using simulated and real-world RNA-seq data of SARS-Cov2 infection development (Kim et al. 2021). On the simulated data set, Spycone outperforms its closest competitor TSIS (Guo et al. 2017), which does not provide functional interpretation, in terms of precision and recall. On the real-world data set, Spycone identified gene network modules involved in cell response after SARS-Cov2 infection, uniquely highlighting changes in AS associated with the disease.

In conclusion, Spycone identifies genes with co-occurring IS in time-course RNA-seq data and allows for their functional interpretation. Spycone, thus, offers a unique systems medicine view on the temporal cellular regulation of AS.

Systems

The sample size value in Network Medicine: an application of gene co-expression networks

Joaquim Aguirre-Plans (Northeastern University), Bingsheng Chen (Northeastern University), Deisy Morselli Gysi (Northeastern University) and Albert-Laszlo Barabasi (Northeastern University).

Abstract:

The sequencing of mRNA levels (RNAseq) in specific contexts has improved our understanding of disease phenotypes and drug mechanisms. Network Medicine uses RNAseq data for inferring gene co-expression networks, which show similarity between the expression patterns of genes, unraveling potential associations between them. The influence of sample size in the predictive power of gene co-expression networks is still not understood. Previous studies suggest that sample size augments the reproducibility of networks, converging to more stable models, and improves the prediction of functional associations. However, these studies are limited to a few dozens of samples. Moreover, we still do not understand how the significant interactions in gene co-expression networks evolve with sample size, which parts of the network are more affected, nor how this could affect the outcome of a case study.

We observed that the probability of revealing new significant edges is power-law related to sample number p(N)~N^(-α), with exponent α ranging from 1.5-1.9. Based on this observation, we derived an analytical solution that explains how the number of significant interactions grows less as sample size increases. We validated our model on different independent datasets, demonstrating a good fit. The model provides an accurate estimate of the fraction of information that can be obtained from gene co-expression networks of a given number of samples, enhancing the uncertainty of Network Medicine analyses based on RNAseq data.

Systems

Topological analysis as a tool for detection of abnormalities in protein-protein interaction data

Alicja Nowakowska (Wroclaw University of Science and Technology) and Malgorzata Kotulska (Wroclaw University of Science and Technology).

Abstract:

Protein-protein interaction datasets are a valuable source of information in biological analysis. They can be modeled as networks with nodes corresponding to proteins and links to interactions between them. This representation may uncover novel potential drug targets, help to predict a therapy outcome or give insights into molecular pathways. Nevertheless, the data that constitutes such systems is frequently incomplete, error-prone and biased by scientific trends. Implementation of methods for detection of such shortcomings could improve protein-protein interaction data analysis. We performed topological analysis of three protein-protein interaction networks (PPINs) from IntAct Molecular Database, regarding cancer, Parkinson's disease (two most common subjects in PPINs analysis) and Human Reference Interactome. The data gathering procedure biased by scientific interests was found to highly impact the networks structure. It was shown that it may obscure correct systematic biological interpretation of the protein-protein interactions and limit the networks' application potential. As a solution to this problem we proposed a set of topological methods for the bias detection. Such analysis performed in the first step can give rise to more objective biological conclusions regarding protein-protein interaction data. A user-friendly tool ETNA (Extensive Tool for Network Analysis) is available on https://github.com/AlicjaNowakowska/ETNA. The software includes a graphical Colab notebook: https://githubtocolab.com/AlicjaNowakowska/ETNA/blob/main/ETNAColab.ipynb

Systems

Towards a community-driven benchmark: PerMedCoE prepares a three agent-based modelling frameworks comparison.

Thaleia Ntiniakou (Barcelona Supercomputing Center), Arnau Montagud (Barcelona Supercomputing Center) and Alfonso Valencia (Barcelona Supercomputing Center).

Abstract:

PerMedCoE is the HPC/Exascale Centre of Excellence for Personalised Medicine in Europe that aims to scale-up the essential software for the cell-level simulation to the new European HPC/Exascale systems.

Agent-based models, one of these cell-level modelling tools, have proven their usefulness in a variety of biological problems, such as modelling cancer cells and their response to drugs or studying the interplay between cancer cells and microenvironment modifications.

In PerMedCoE, we have initiated an observatory which includes different modelling pieces of software and their corresponding technical characteristics. Next, we have performed a benchmark involving three agent-based modelling frameworks as a first step to guide the procedure towards a community-driven benchmark. The process of benchmarking these tools will eventually lead to facilitating users on deciding which tool would serve them with respect to their modelling goal and preferences.

The comparisons between Timothy, Physicell, and Chaste were performed in the same cluster while utilising a common use case that was shared among the templates found for each software.

Each tool was used to simulate the growth of cell population and their response to oxygen presence. We evaluated their performance by measuring the total number of cells grown, while the simulations were executed for the same amount of time, 2 hours. Further technical metrics provided by the cluster facility were used as well to evaluate the performance of the methods, namely CPU and memory use and energy consumption. At a glance, these simulators seem similar, however, they specialise in different use cases.

Systems

Towards a data-driven network inference model of interactions between immune and cancer cells in Chronic Lymphocytic Leukaemia

Malvina Marku (INSERM, Cancer Research Center of Toulouse), Hugo Chenel (INSERM, Cancer Research Center of Toulouse), Julie Bordenave (INSERM, Cancer Research Center of Toulouse), Nina Verstraete (INSERM, Cancer Research Center of Toulouse), Leila Khajavi (Institut Universitaire du Cancer de Toulouse-Oncopole), Flavien Raynal (INSERM, Cancer Research Center of Toulouse) and Vera Pancaldi (INSERM, Cancer Research Center of Toulouse).

Abstract:

The tumour microenvironment (TME) can be seen as a complex system containing multiple cell types interacting through contact and cytokine exchanges. Particularly, immune cells play a major role in cancer development and their characterization allows a better understanding of the TME. In this context, transcriptomics time courses allow studying the gene regulatory networks and interactions between immune and cancer cells to obtain relevant information about the biology behind them, and to identify novel molecular interactions and potential drug targets.

In this project, we aim to characterise the formation of Nurse Like Cells (NLC), macrophages found in lymph nodes of Chronic Lymphocytic Leukaemia (CLL) patients, and to investigate the crosstalk between them from a network perspective. We performed GRN inference on both cell types, using advanced inference methods on a unique transcriptomics time-series on purified cell types from a 13-days co-culture. Network topology analysis allowed us to identify the main regulators of the inferred directed networks, involved in establishing the cross-talk between NLC and CLL cells. Furthermore TF enrichment analysis shed light on the processes taking place inside the two cell populations during the 13-days evolution of the co-culture. We aim to integrate these entirely data-driven results with information from databases and apply dynamical models to study the temporal behaviour of the co-culture, both at the molecular and cellular level.

Systems

Tracking pathway activity in pseudotime

Priyansh Srivastava (University of Valencia; BioBam Bioinformatics S.L.), Stefan Götz (BioBam Bioinformatics S.L.) and Ana Conesa (Institute for Integrative Systems Biology (I²SysBio), CSIC).

Abstract:

Every cell emerges from pre-existing cells by undergoing pre-determined transcriptional events. The developmental history of a differentiated cell from a progenitor cell forms cell lineages. scRNA-Seq delivers transcriptional profiling of thousands of individual cells that helps infer the underlying biology. Computational tools for scRNA-Seq data analysis enable the determination of cell fate decisions in dynamic transcriptional units called pseudotime. Pseudotime reflects how far a particular cell is from its progenitors in terms of transcriptional distances. The alignment of cells along the continuum of pseudotime is called a trajectory. Although, we can explain the cell fate decisions using genes with a significant expression in the cell lineage. However, using just an elementary gene list usually has limited biological context.

Characterizing changes in pathway activity over pseudotime can help understand the dynamic states of individual cells from a biological perspective. We propose a novel method to represent pathway activities from scRNA-seq measurements by inferring the latent space associated with each pathway and obtaining metagenes (adapted from Brunet et al., 2004) that represent the pathway. The approach returns non-negative metagene data that can be subjected to traditional trajectory analysis and visualized with standard scRNA-seq tools to identify pathway changes in pseudotime. We evaluate different dimension reduction techniques and benchmark our pipeline on various scRNA datasets with varying complexity to evaluate its reproducibility and robustness.

Systems

Tumor microenvironment evolution simulated through a hybrid Multi-Agent Spatio-Temporal model informed using sequencing data

Mikele Milia (University of Padova), Giulia Cesaro (University of Padova), Giacomo Baruzzo (University of Padova), Giovanni Finco (University of Padova), Francesco Morandini (University of Padova), Alessio Lazzarini (University of Padova), Piergiorgio Alotto (University of Padova), Noel Filipe da Cunha Carvalho de Miranda (Leiden University Medical Center), Zlatko Trajanoski (Medical University of Innsbruck), Francesca Finotello (University Innsbruck) and Barbara Di Camillo (University of Padova).

Abstract:

Several computational modeling approaches have been developed for simulating complex system, such as tumor microenvironment (TME), since they represent a mean toward better understanding of the tumor-immune dynamics that drive cancer development, and there is the unmet need for personalized simulations of specific cancer scenarios.

Here, we present MAST, a hybrid Multi-Agent Spatio-Temporal model of the interaction between immune system and tumor growth that couples a discrete and stochastic agent-based model with a continuous and deterministic partial differential equations-based model allowing to capture essential elements in the TME. Specifically, it allows to model (1) spatio-temporal dynamics of cells in response to nutrient availability, (2) both innate and adaptive immune system response, and (3) immune escape mechanisms.

The simulation of specific and unique TME scenario is performed in a data-driven way: model parameters are set from the analysis of high-throughput sequencing (HTS) TME data (e.g., bulk RNA-seq, DNA-seq and scRNA-seq), yielding knowledge on i) acquisition rate of new mutations (through the computation of tumor mutational burden from DNA-seq data); ii) cell type composition and recruitment (through the analysis of cells proportions from deconvolution of bulk RNA-seq data or cell-type annotation of scRNA-seq data); iii) loss of immunogenicity (through the identification of upregulated inhibitory immune checkpoint genes from RNA-seq or scRNA-seq data).

We inform MAST with HTS data of human colorectal cancer (CRC) and validate the model outcomes with both TCGA patient clinical data and biological knowledge on CRC Consensus Molecular Subtypes, reproducing emergent properties and predicting tumor progression consistently with literature.

Systems

Using neural networks to decipher non-equilibrium states of cell differentiation process

Susan Ghaderi (KU Leuven), Alexander Skupin (University of Luxembourg) and Yves Moreau (KU Leuven).

Abstract:

Understanding cell fate mechanisms is a fundamental problem in systems biology. While it clearly is based on a multi-stable dynamical system and differentiation is corresponding to the transition between different attractors, the underlying mechanism still is neither well-structured nor well-understood. There have been great studies on equilibrium states of this phenomenon, while cell differentiation by its nature is a non-equilibrium phenomenon.

In this research, we consider the Gibbs free energy as the appropriate thermodynamic potential of cell differentiation. For the analysis of the differentiation process, we use Kullback-Leibler distance (KL) (also referred to as relative entropy) as an indicator for the non-equilibrium state of the system where we exploit the sc-RNAseq distributions for each gene condition with a uniform reference distribution. Then, we use a self-organizing map (SOM) approach to investigate key genes of this non-equilibrium process. We applied our methodologies to sc-RNAseq data of Parkinson's disease-related iPSC differentiation into dopaminergic neurons with two treatments, mutant, and control. For the subsequent interpretation of these dynamics, we ordered genes by their non-equilibrium dynamics quantified by the KL values using SOMs. From the resulting SOM patterns, we subsequently identified the main contributing genes of the assumed attractor states using the 10\% largest KL values.

Climate Crisis and Health

A metabolomic approach to understand the metabolic link between obesity and type 2 diabetes

Qiuling Dong (Helmholtz Munich), Sapna Sharma (Helmholtz Munich) and Harald Grallert (Helmholtz Munich).

Abstract:

Introduction: The circulating metabolites are perplexed readouts of the biological processes that reflect pathophysiological events in different tissues and organs. Many metabolites have been linked to complex disorders and are also under substantial genetic control. Our aims in this study are, (a) characterise 148 small molecular signatures associated with BMI and Type 2 Diabetes (T2D) in KORA cohort N=1715. (b) identify metabolite that mediate the BMI effect on T2D and (c) using Mendelian randomization to find causal relationship of BMI and T2D with metabolites.

Results: We observed 50 metabolites were significantly associated with T2D, those comprised of branched-chain amino ketoacids, acylcarnitines, lysophospholipids, or phosphatidylcholines, are largely replicated in the previous studies. Mediation analysis with respect to BMI suggests that the effect of BMI on T2D may be mediated via replicated metabolite like SM.C16.1 and marginally by SM.C18.1. Mendelian randomization suggests a bi-directional causal relationship of BMI with SM.C16.1.

Conclusion: Our findings suggest metabolite SM.C16.1 from phosphocholines class mediate the effect of BMI on T2D, these metabolites further elucidating their role in T2D associated pathologies.

Climate Crisis and Health

APPLYING HUMAN EPIGENETIC INTER-VARIABILITY FOR PERSONALIZED NUTRITION STRATEGIES IN CANCER PATIENTS

Teresa Laguna (IMDEA Food Institute), Marco Garranzo (IMDEA Food Institute), Marta Gómez de Cedrón (IMDEA Food Institute), Ana Ramírez de Molina (IMDEA Food Institute) and Enrique Carrillo de Santa Pau (IMDEA Food Institute).

Abstract:

Food natural compounds have become of interest as modulators of cancer development, as they have shown to intervene in cellular processes such as growth and differentiation, DNA repair, programmed cell death, and oxidative stress. Here we show a workflow which compares the expression profiles of cells treated with phytochemicals and other dietary bioactives, with the expression profiles of colon tissues of colon adenocarcinoma (COAD) patients, to design specific supplementation to improve cancer treatment. For that, we established a classification of colon cancer patients based on epigenetic variability, which has been associated with cell type microenvironment proportions. We defined 6 subtypes of colon adenocarcinoma patients based on the different variability of DNA methylation (DV2,3,4,6,7,8), and characterised their expression profiles, finding DV4 and DV6 associated with immune functions. We analysed food compound transcriptomic profiles looking for the most similar or antagonic profiles with each one of the 6 subtypes. We selected 31 studies with unique bioactive concentrations and target cells investigated based on polyphenols and colon cell lines as targets, finding that 30 of these conditions were significatively similar or opposite to at least 1 group of patients. Specifically, rosemary and cocoa extracts show a general effect in all patient groups in direct and opposite direction, respectively. Carnosic acid and 6-shogaol, on the contrary, exhibit specific similarities with DV6 (immune). This workflow can be further extended to other bioactives and cancer types, laying the groundwork for designing effective food supplementation strategies for cancer patients.

Climate Crisis and Health

Automated Diagnostic System for Medical Treatment of Infectious Diseases using Causal transfer learning and biological knowledge graph embedding

Sakhaa Alsaedi (King Abdullah University of Science and Technology (KAUST)), Katsuhiko Mineta (King Abdullah University of Science and Technology (KAUST)), Xin Gao (King Abdullah University of Science and Technology (KAUST)) and Takashi Gojobori (King Abdullah University of Science and Technology (KAUST)).

Abstract:

In infectious diseases, molecular diagnostics are revolutionizing clinical practice by helping doctors understand a patient's cases caused by infection before symptoms and complications. Moreover, using machine learning algorithms to assist doctors in clinical decision-making and diagnosis is critical for patient treatment decisions and outcomes. However, current automated diagnosis systems only utilize associative deep learning methods that identify diseases strongly correlated with a patient's symptoms without considering the genetic risk factors that may cause complications. Alternatively, they could be related to other complex disorders affecting the patient's situation. In that case, understanding how different viral strains affect individual patients and, in particular, how they interact with different human host cells and immune responses is a fundamental step in order to formulate accurate treatment plans. Since the outbreak of the COVID-19 disease, host genetic variations play a significant role in the manifestation of different degrees of severity of illness among different individuals. It is crucial to use this disease as the first case study in our research. Thus, we develop a deep learning model that provides automated medical plans and predicts the severity score as well as multi-organs dysfunction scores during infection by integrating genetic and viral data with metadata and analyzing risk factors. Our preliminary result shows that our model performs better than state-of-the-art on synthetic data. The data was generated based on descriptive information that explained the severity of COVID-19 patients from scientific articles and medical reports. In addition, we test models on actual medical records of the sensitivity of obtaining medical reports. The predicted scores assist doctors in having a better understanding of the COVID-19 cases and provide an accurate treatment plan that could eventually reduce the severity and complication of infectious diseases.

Climate Crisis and Health

Characterization Of The Latent Resistome In External And Host-associated Environments

Juan Inda-Díaz (Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg), David Lund (Department of Mathematical Sciences, Chalmers University of Technology), Anna Johnning (Fraunhofer Chalmers Research Centre for Industrial Mathematics), Marcos Parras-Moltó (Department of Mathematical Sciences, Chalmers University of Technology), Johan Bengtsson-Palme (Department of Infectious Diseases, University of Gothenburg) and Erik Kristiansson (Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg).

Abstract:

Bacterial communities in the human, animal, and environmental microbiomes maintain a large collection of antibiotic resistance genes (ARGs): the resistome. Only a small part of these ARGs are known and well-established in sequence repositories, while the large majority are uncharacterized and often overlooked. Consequently, our view of the abundance and diversity of the full resistome is limited and hampers our ability to identify risk environments for the mobilization, promotion, and spread of ARGs into pathogens.

A comprehensive database with 23,092 ARGs was built, including both, computationally predicted and ResFinder genes. Latent genes, e.g., sequence similarity <90% to any gene in ResFinder, constituted 97% of them. The latent ARGs constituted a chief part of the abundance and diversity of the resistome in 10,608 metagenomic samples from 20 environments including external and host-associated environments. A strong presence of latent ARGs in commonly reoccurring genes was found in the human- and animal-associated environments, as well as in the external environments. The latent commonly reoccurring genes were, to a large extent, found in the proximity of mobile genetic elements and shared between digestive systems, suggesting that they are both, mobile and under strong selection pressures. We identified exceptionally high diversity of both established and latent ARGs in wastewater, which makes this environment a potential hot spots for the mobilization of new resistance genes into pathogens.

Future studies of the resistome should include both the established as well as the latent ARGs to provide a more complete picture of the resistome present in bacterial communities.

Climate Crisis and Health

DeclaraMID: A Declarative Framework for Modeling Infectious Diseases

Sebastiaan Weytjens (Interuniversity Institute of Biostatistics and statistical Bioinformatics, Hasselt University, Belgium), Ann Nowé (Artificial Intelligence Lab, Vrije Universiteit Brussel, Belgium), Niel Hens (Interuniversity Institute of Biostatistics and statistical Bioinformatics, Hasselt University, Belgium) and Pieter Libin (Artificial Intelligence Lab, Vrije Universiteit Brussel, Belgium).

Abstract:

Epidemiological models aim to capture the complex dynamics of infectious diseases while they need to be computationally efficient to facilitate high-throughput simulations, which are required to fit models, simulate extensive model scenarios and optimize mitigation policies. Typically, creating a new model, that is specific to a particular pathogen and population structure, requires writing or adapting code that is non-trivial, making it prone to errors and inefficiencies. Moreover, this results in a tight coupling between the model's description (i.a., compartments, transitions, mobility) and its simulation solvers. This may render the code hard to maintain and extend, thereby impeding reuse of the model for simulating other scenarios.

To address these issues, we construct a modeling framework that enables experts to build structured metapopulation models (incorporating structures for i.a. age and vaccination status) in a declarative manner, which can be evaluated with our solver framework (i.a., ordinary differential equations, Gillespie, and custom solvers). This approach separates the model specifications from the model evaluation, thereby facilitating modularity, which enables the reuse of models and solvers.

Finally, we demonstrate the framework's potential by constructing metapopulation models for SARS-CoV-2 and the Ebola virus. We show that models can be described in significantly less code, and we establish that models can be extended easily. As such, the focus lies solely on the model description without involving technical details related to model evaluation. To conclude, our framework reduces time to construct models, enabling researchers to focus on defining model specifications and obtaining simulation results.

Climate Crisis and Health

Deep Learning to differentiate between images of burnt and normal skin

Kirsty Smith (University of Bradford), Sharmila Jivan (NHS), Ajay Mahajan (NHS), Khaled Jumah (University of Bradford) and Krzysztof Poterlowicz (University of Bradford).

Abstract:

Burn injuries are a common presentation to the accident and the emergency department. The accurate assessment of these injuries by determining area and depth can ensure the patient receives the most appropriate treatment however investigating burn injuries is not easy and the accuracy of the initial assessment can vary depending on the experience of the assessor. If these assessments are incorrect, it can result in inadequate treatment or unnecessary transfer to specialists centres causing distress. Therefore there is a clear need to develop unbiased, automated methods for a more accurate diagnosis of burn injuries that in future can be used in clinical practice to aid clinicians.

Deep Learning methods have demonstrated high accuracy and scalability in the image analysis tasks and they have been widely applied in medical image classification tasks such as computed tomography (CT) and magnetic resonance imaging (MRI)

Here we developed a convolutional neural network model to assess burn images from the study that involves a cohort of patients from the Mid Yorkshire Hospitals NHS Trust (UK).

The model showed 93% accuracy in predicting the difference between a superficial burn, a full-thickness burn and normal skin. Future work will include the implementation of the method in a clinical setting.

Climate Crisis and Health

EDAM and EDAM Geo: an ontology for data-intensive, interdisciplinary geo- and biosciences

Lucie Lamothe (IFB-core, French Institute of Bioinformatics, CNRS), Mads Kierkegaard (University of Southern Denmark, Ødense), Melissa Black (Outreachy intern (EDAM), São Paulo (at the time of contribution)), Hager Eldakroury (Outreachy intern (EDAM), Cairo (at the time of contribution)), Veit Schwämmle (University of Southern Denmark, Ødense), Hamish Struthers (Linköping University), Bryan Brancotte (Pasteur Institute and Paris Cité University), Kessy Abarenkov (University of Tartu), Olga Silantyeva (University of Oslo), Jean Iaquinta (University of Oslo), Jon Ison (IFB-core, French Institute of Bioinformatics, CNRS (at the time of contribution)), Jonathan Karr (Icahn School of Medicine at Mount Sinai, New York City), Anne Fouilloux (University of Oslo), Alban Gaignard (L'institut du thorax, University of Nantes/CNRS/INSERM), Hervé Ménager (Pasteur Institute and Paris Cité University), Matúš Kalaš (University of Bergen) and And The Edam Community (the EDAM community).

Abstract:

EDAM is an ontology of data analysis and data management, within and beyond biosciences. It comprises concepts related to analysis, modelling, optimisation, and data life-cycle. The structure of EDAM is relatively simple, divided into 4 main sections: topics, operations, data, and formats.

EDAM is used in a large number of resources, for example Bio.tools, Galaxy, CWL, Debian, BioSimulators, FAIRsharing, or the ELIXIR training portal TeSS. Thanks to the annotations with EDAM, computational tools, workflows, standards, data, and learning materials are easier to find, compare, choose, and combine. EDAM contributes to open science by supplying concepts usable for semantic annotation of research outputs (such as processed data), making them more understandable, findable, and comparable. EDAM and its applications help lower the barrier and effort for scientists, professional and "citizen", towards doing scientific research in a more open, reliable, and inclusive way.

EDAM is developed in a participatory and transparent fashion, by a diverse, growing community of contributors. A substantial community extension is EDAM Bioimaging, comprising concepts related to image analysis and machine learning.

Since 2021, the EDAM Geo community has been forming, and extending EDAM into interdisciplinary application domains. Examples include public, global, and planetary health; environmental sciences; climate and pollution; or highly interdisciplinary Earth system modelling. EDAM Geo covers also geographical data and geoinformatics data formats.

Both EDAM Geo and EDAM Bioimaging are being merged iteratively into the "main" EDAM, to smoothly serve the needs of the cutting edge, data-intensive interdisciplinary research and research-based applications.

Climate Crisis and Health

Evaluating COVID-19 vaccine allocation policies using Bayesian m-top exploration

Alexandra Cimpean (Vrije Universiteit Brussel), Lander Willem (University of Antwerp), Timothy Verstraeten (Vrije Universiteit Brussel), Niel Hens (UHasselt), Ann Nowé (Vrije Universiteit Brussel) and Pieter Libin (Vrije Universiteit Brussel).

Abstract:

Individual-based epidemiological models (IBMs) support the study of fine-grained preventive measures, such as tailored vaccine allocation policies, in silico. As IBMs are computationally intensive, it is pivotal to identify optimal strategies using a minimal amount of model evaluations. Moreover, due to the high societal impact of enforcing preventive strategies, uncertainty regarding decisions should be communicated to policy makers, which is naturally embedded in a Bayesian approach.

We contribute a novel technique to evaluate vaccine allocation strategies using a multi-armed bandit framework in combination with a Bayesian anytime m-top exploration algorithm. m-top exploration allows the algorithm to learn m policies for which it expects the highest utility, enabling the experts to inspect this small set of alternative strategies, along with their quantified uncertainty for the decision making. The anytime component provides decision makers with flexibility regarding the computation time and the desired confidence, which is important as it is difficult to make this trade-off beforehand.

We consider the Belgian COVID-19 epidemic using the STRIDE IBM, where we learn a set of optimal vaccination policies that minimize the number of hospitalisations. In this scenario, the shape of model outcome distribution cannot be assumed analytically a priori, to which end we use a Gaussian mixture to model the arms' posteriors. Through our experiments we show that our method can efficiently identify the m best policies, which is validated in a scenario where the ground truth is available. Finally, we explore how vaccination policies can best be organized under different contact reduction schemes.

Climate Crisis and Health

Exploring the Pareto front of multi-objective COVID-19 mitigation policies using reinforcement learning

Mathieu Reymond (Vrije Universiteit Brussel), Conor F Hayes (NUI Galway), Lander Willem (Universiteit Antwerpen), Roxana Radulescu (Vrije Universiteit Brussel), Steven Abrams (Universiteit Antwerpen), Diederik M. Roijers (Vrije Universiteit Brussel & HU University of Applied Sciences Utrecht), Enda Howley (NUI Galway), Patrick Mannion (National University of Ireland Galway), Niel Hens (Universiteit Hassel), Ann Nowé (Vrije Universiteit Brussel) and Pieter Libin (Vrije Universiteit Brussel).

Abstract:

Infectious disease outbreaks can have a disruptive impact on public health and societal processes. As decision making in the context of epidemic mitigation is hard, reinforcement learning provides a methodology to automatically learn prevention strategies in combination with complex epidemic models. Current research focuses on optimizing policies with respect to a single objective, such as the pathogen's attack rate. However, as the mitigation of epidemics involves distinct, and possibly conflicting, criteria (i.a., prevalence, mortality, morbidity, cost), a multi-objective decision approach is warranted to learn balanced policies. To lift this decision-making process to real-world epidemic models, we apply deep multi-objective reinforcement learning and build upon a state-of-the-art algorithm, Pareto Conditioned Networks (PCN), to learn a set of solutions that approximates the Pareto front of the decision problem. We consider the first wave of the Belgian COVID-19 epidemic, which was mitigated by a lockdown, and study different deconfinement strategies, aiming to minimize both COVID-19 cases (i.e., infections and hospitalizations) and the societal burden that is induced by the applied mitigation measures. We contribute a multi-objective Markov decision process that encapsulates the stochastic compartment model that was used to inform policy makers during the COVID-19 epidemic. We evaluate the solution set that PCN returns, and observe that it correctly learns to reduce the social burden whenever the hospitalization rates are sufficiently low. In this work, we thus demonstrate that multi-objective reinforcement learning is attainable in complex epidemiological models and provides essential insights to balance complex mitigation policies.

Climate Crisis and Health

Feed Your Model More Protein: Novel features based on viral protein composition and predicted interaction improve and slimline models of viral zoonoses

Matt Arnold (University of Glasgow), Simon Crouzet (École Normale Supérieure de Paris), Daniel Streicker (University of Glasgow Centre for Virus Research) and Simon Babayan (University of Glasgow Institute of Biodiversity, Animal Health and Comparative Medicine).

Abstract:

Increasing contact between humans, livestock, and wildlife, be it through changing land use or new invasive species as a result of climate change, exacerbates the likelihood of future pandemics. Of known viruses, over half infect multiple species and spillover events between populations can lead to outbreaks of disease, both from novel viruses and novel strains of known viruses. Tools to inform viral surveillance are therefore essential to controlling the risk of emerging viral disease, so the actual or as yet unrealised host populations for these viruses can be identified and effective control measures designed.

Using machine learning models, we have leveraged viral genome data to make predictions of the host taxon of a virus based on features calculated as biologically-informed summary metrics of its genomic information. Previous approaches by our group (Babayan et al., Science 2018) used viral phylogenetic neighbourhood and summarised genome composition in gradient-boosted tree models to achieve this end. We present here results showing improved models incorporating of novel protein structure- and amino acid sequence-derived features. Additionally, we propose a lightweight model with similar predictive capabilities using a smaller set of features designed to predict viral protein post-translational modification and protein-protein interaction.

These findings provide new insights into the genetic interactions underlying viral host species tropism and represent a step forward in the development of methods to aid field surveillance of emerging viral zoonoses.

Climate Crisis and Health

FOODRUGS: AN EXPLORATORY MINING WEB APPLICATION FOR FOOD - DRUG INTERACTIONS

Marco Garranzo (Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, 28049 Madrid, Spain), Teresa Laguna Lobo (Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, 28049 Madrid, Spain), David Pérez Serrano (Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, 28049 Madrid, Spain), Blanca Lacruz Pleguezuelos (Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, 28049 Madrid, Spain), María Isabel Espinosa (Nutritional Genomics and Health Unit, GENYAL Platform, IMDEA Food Institute, 28049 Madrid, Spain), Ana Ramírez de Molina (GENYAL Platform, IMDEA Food Institute, 28049 Madrid, Spain) and Enrique Carrillo-De Santa Pau (Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, 28049 Madrid, Spain).

Abstract:

The consumption of supplement foods with an active role to prevent non-communicable diseases have increased during the last years raising concerns over the potential interactions with drugs, particularly in patients undergoing chronic therapy. However, the current knowledge of them in clinical practice has been reported to be unsatisfactory. Therefore, it is required to develop computational tools to exploit the vast information available in dispersed databases about food-drug interactions. We present FOODRUGS, a friendly-user webapp to access and explore putative food-drug interactions. Users can explore transcriptomic relationships between food and drugs through a similarity bipartite network built with 334 food transcriptomic profiles from GEO and the CMAP database for drug profiles. Food and drug nodes are connected by a positive or negative tau score representing the similarity (positive) or dissimilarity (negative) relationship, meaning for a putative similar or opposite mechanism of action. In addition, enrichment analysis to annotate drug compounds for a food node can be performed. Lastly, more than 2.500 scientific documents from Pubmed, drugbank and drugs.com can be consulted after applying NLP techniques to extract relevant food-drug interactions. Users can access information for food transcriptomic studies, food gene sets, molecular interaction network, text information extraction and the complete database in MySQL format. FOODRUGS is accessible in http://imdeafoodcompubio.com/index.php/foodrugs/

This research was undertaken by IMDEA Food (IMDEA, ES), a beneficiary in FNS-Cloud, which has received funding from the European Union's Horizon 2020 Research and Innovation programme (H2020-EU.3.2.2.3. – A sustainable and competitive agri-food industry) under Grant Agreement No. 863059 - www.fns-cloud.eu

Climate Crisis and Health

Providing a dashboard for monitoring effects of contact behavior on the spread of SARS-CoV-2

Paul Chevelev (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Richard Schiemenz (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Ferdous Nasri (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Richard Pastor (Machine Learning Unit, Department of Engineering, NET CHECK GmbH, 10829 Berlin, Germany), Bernhard Y Renard (Data Analytics & Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam) and Sten Rüdiger (Machine Learning Unit, Department of Engineering, NET CHECK GmbH, 10829 Berlin, Germany).

Abstract:

In light of high SARS-CoV-2 infection numbers, policies to minimize its spread are necessary. Measures aiming to reduce contacts have proven to be an effective tool, but require the cooperation of the population to be effective. This support can only be achieved through clear communication.

In order to inform decision-makers and increase the public's openness to contact-reducing measures, thereby decreasing the spread of the pandemic, we have developed the following publicly available dashboard: https://contactindex.netcheck.de/.

We feature a unique metric, the contact-index, that has previously been proven to be correlated to the effective reproduction number. It is based on pings of anonymous mobile app users, which corrects the raw number of contacts by weighting in secondary contacts, akin to the so-called friendship paradox. We also feature infection data spanning the whole length of the pandemic, with many interactive tools in order to create a comprehensible overview of the progression of SARS-CoV-2. To be more educational for a general audience, we include explanations of the data and relevant scientific terms. Furthermore, we use colorblind-friendly coloring schemes throughout the dashboard to increase accessibility. Our solution is based on plotly.js in order to be highly adaptable to different types of data and regions but still maintain a high performance. By being the first German dashboard to visualize the impact of contacts on the spread of COVID-19, we believe that policy-makers can more easily see the usefulness of current interventions and make informed decisions for the future.

Climate Crisis and Health

Rooting virus phylogenies using UNREST and the general Markov model

Jose Nikhil Teja Dasari (Max Planck Institute for Molecular Genetics), Prabhav Kalaghatgi (Max Planck Institute for Molecular Genetics) and Martin Vingron (Max Planck Institute for Molecular Genetics).

Abstract:

Time-reversible models are widely used for computational ease even though it is necessary to root phylogenetic trees in order to model virus phylodynamics. Non-reversible models such as UNREST and the general Markov model (GMM) can be used to infer rooted phylogenies but have not yet been compared on SARS-CoV2 and H3N2 data. 9504 H3N2 HA gene sequences and 9499 SARS-CoV2 genomes were downloaded from GISAID and aligned using MAFFT. Structural EM was used to infer rooted trees under GMM. IQ-TREE, MAPLE and rootDigger were used to infer rooted trees under UNREST. MST-backbone was used to create the unrooted trees that are required as input by rootDigger. We assessed the location of the root in inferred trees by measuring the correlation of root-to-tip tree distances with sample collection times. We measured the inferred phylogenetic trees by calculating the Pearson's correlation coefficient (r) of sampling times. The H3N2 HA gene tree inferred using rootDigger has an r of 0.98 followed by MST-backbone under GMM with an r of 0.85. The trees inferred by IQ-TREE and MAPLE have low r-values of -0.89 and -0.07, respectively. In the case of SARS-CoV2 genomes IQ-TREE and MAPLE achieved the highest r-value of 0.91 and 0.90 followed by MST-backbone under GMM and rootDigger under UNREST each with an r of 0.62.

Climate Crisis and Health

Similarity network approach on transcriptomic data of patients with systemic infectious diseases.

Francesco Messina (National Institute of Infectious Diseases "L. Spallanzani", Rome, Italy), Carolina Venditti (National Institute of Infectious Diseases "L. Spallanzani", Rome, Italy), Carla Nisii (National Institute of Infectious Diseases "L. Spallanzani", Rome, Italy), Carla Fontana (National Institute of Infectious Diseases "L. Spallanzani", Rome, Italy) and Alessandro Capone (National Institute of Infectious Diseases "L. Spallanzani", Rome, Italy).

Abstract:

The high complexity in patients with systemic infectious diseases, due to bacterial, fungal and viral pathogens, imposed a challenge to design new strategies of clustering patients, through identification of specific molecular variation on different “omics” levels (differentially expressed genes, SNPs, ect.). Here, we propose the application of patient similarity network (PSN) in patients with systemic infectious diseases, based on single cell RNAseq in PBMC (B cells, T cells, DC, Monocytes and NK) and microbial data already published. This experience enabled to determine distinct subgroups of patients, caused by molecular variation in transcriptomic data per single cells. Such subset could be explained by known factors (e.i. ICU condition, shock septic), but also unconsidered variables.

For the first time, this analysis allowed to design Patient Similarity Network of systemic infectious diseases, defining patient clusters with different disease severity degree, featured by transcriptomic signature variations in PBMCs. This experience drives to improve molecular resolution power in systemic infectious diseases, adding other “omics” data in both human host (eQTL, whole exome, blood proteomics, etc.) and pathogen (WGS, AMR and virulence genomic profiling).

Climate Crisis and Health

TCR-epitope recognition models to untangle unique and cross-reactive T-cell immunity in COVID-19 patients

Anna Postovskaya (University of Antwerp), Alexandra Vujkovic (Institute of Tropical Medicine), Tessa de Block (Institute of Tropical Medicine), Lida van Petersen (Institute of Tropical Medicine), Maartje van Frankenhuijsen (Institute of Tropical Medicine), Isabel Brosius (Institute of Tropical Medicine), Emannuel Bottieau (Institute of Tropical Medicine), Christophe Van Dijck (Institute of Tropical Medicine), Caroline Theunissen (Institute of Tropical Medicine), Sabrina van Ierssel (Antwe

Abstract:

Efforts to unravel the cellular immune response to SARS-CoV-2 are ongoing as this knowledge may indicate how to maintain robust long-term protection against continuously emerging variants. Since the evidence on the clinical benefits of T cells cross-reactive with seasonal coronaviruses is still conflicting, we set out to characterize CD8+ T-cell response to epitopes unique to SARS-CoV-2 (SC2-unique) and shared with other coronaviruses (CoV-common) in patients with different COVID-19 severity utilizing T-cell receptor (TCR) recognition models.

Specifically, models were built for SARS-CoV-2 epitopes using the TCRex framework to predict whether an unseen TCR recognizes a given epitope. Application of those models to longitudinal (up to 8 weeks) TCR repertoires of 15 critical and 31 non-critical COVID-19 patients revealed (dis)similarities in their temporal dynamics. Despite CD8+ T-cell depletion, the frequencies of CoV-common TCRs were significantly higher than of SC2-unique TCRs during the first week of the disease in both patient groups. By the second week, only non-critical patients managed to sustain CoV-common and develop SC2-unique TCRs. In the critical group, frequencies of SC2-unique but not CoV-common TCRs took longer (at least 6 weeks) to increase accordingly. Finally, we observed that non-critical patients recognize significantly more unique and common SARS-CoV-2 epitopes and, unlike critical patients, demonstrate redundancy in CD8+ TCRs recognizing them.

Monitoring the prevalence of specific CD8+ TCRs in the repertoires offers an opportunity for timely treatment or revaccination intervention. Our complementary computational analysis can be leveraged to extract relevant information faster and with more flexibility than in vitro testing.

Applications

Alternative splicing analysis benchmark with DICAST

Amit Fenn (Technical University of Munich), Olga Tsoy (University of Hamburg), Tim Faro (Technical University of Munich), Fanny Roessler (Technical University of Munich), Alexander Dietrich (Technical University of Munich), Johannes Kersting (Technical University of Munich), Zakaria Louadi (Technical University of Munich), Chit Tong Lio (University of Hamburg), Uwe Völker (University Medicine Greifswald), Jan Baumbach (University of Hamburg), Tim Kacprowski (Technische Universität Braunschweig) and Markus List (Technical University of Munich).

Abstract:

Alternative splicing is a major contributor to transcriptome and proteome diversity in health and disease. A plethora of tools have been developed for studying alternative splicing in RNA-seq data. Previous benchmarks focused on isoform quantification and mapping. They neglected event detection tools, which arguably provide the most detailed insights into the alternative splicing process. DICAST offers a modular and extensible framework for analysing alternative splicing integrating eleven splice-aware mapping and eight event detection tools. We benchmark all tools extensively on simulated as well as whole blood RNA-seq data. STAR and HISAT2 demonstrated the best balance between performance and run time. The performance of event detection tools varies widely with no tool outperforming all others. DICAST allows researchers to employ a consensus approach to consider the most successful tools jointly for robust event detection. Furthermore, we propose the first reporting standard to unify existing formats and to guide future tool development.

Applications

Attention-based Variation Graph Autoencoder for DTA Prediction on Multiplex Heterogeneous Network

Dimitrios Papadopoulos (Dept. of Informatics, Aristotle University of Thessaloniki), Bin Liu (Key Laboratory of Data Engineering and Visual Computing, Chongqing University of Posts and Telecommunications, China), David Čechák (Panagiotis Alexiou Research Group, Centre for Molecular Medicine, CEITEC, Masaryk University,Czechia) and Grigorios Tsoumakas (Dept. of Informatics, Aristotle University of Thessaloniki).

Abstract:

A crucial step in drug discovery is hit identification, which refers to the identification of drugs that bind to biological targets to trigger or block a physiological function in order to act therapeutically to a disease. Drug-target interactions (DTIs) are validated through costly and time-consuming wet-lab experiments. Computational DTI predictions can help swiftly discover new promising drug-target interactions to narrow down the search space for wet-lab validations.

DTI task deals with the existence or non-existence of interaction between drug-target pairs (classification). Alternatively, drug-target affinity (DTA) measures interaction strength as a continuous number (regression). Training with DTA allows models to leverage fine-grained information while predicting DTA is more valuable to drug discovery.

Our work explores the DTA prediction based on the combination of currently available public data. We formulate the data as a multilayered heterogeneous attributed network. Its topology arises from the graphical representation of associations like drug-target affinity, drug-drug similarity, and other types of relationships. We augment the network by gathering the binding affinity values of the drug-target pairs. The recent advancements in graph neural networks (GNN) leverage the information from nodes (drugs/targets) and their network topology.

Additionally, we are the first to apply variational graph autoencoders with node-level attention to this problem. Variational autoencoders have been proved powerful in reconstructing the original input in several domains, including network reconstruction. Moreover, node-level multi-head self-attention provides additional learnable parameters to represent multiple levels of importance of a node's neighbours, enabling the node to attend more to some neighbours than others.

Applications

cellAssign: a small toolset to identify the cell types associated with methylation components

Reka Toth (Luxembourg Institute of Health), Pavlo Lutsik (German Cancer Research Center (DKFZ)), Christoph Plass (German Cancer Research Center (DKFZ)) and Petr V. Nazarov (Luxembourg Institute of Health (LIH)).

Abstract:

DNA methylation is highly cell type-specific; therefore, bulk DNA methylation represents signals arising from a mixture of cells. Since cell type composition is an important characteristic of biological samples, its identification is a crucial step in DNA methylation data analysis. Reference-free deconvolution of the bulk data is a powerful tool to detect cell type heterogeneity-related latent components (LCs). However, assigning these components to known cell types is difficult when reference methylomes are not available.

Here we present cellAssign, a small tool that uses cell markers identified by single-cell analysis to assign cell types to methylation components. The method is based on the assumption that the marker genes of a cell type tend to be hypermethylated in their promoter. Therefore first it calculates the correlation of gene promoter methylation with the proportion of the methylation components. Using the correlation coefficients as ranks and the marker genes as gene sets, cellAssign performs a gene set enrichment analysis (GSEA) to assess a level of significance for each cell type in each LCs. The results can be visualized on a GSEA plot. The description of the tool and a real-life example is available on GitHub (tkik/cellAssign).

Despite being a simple tool, cellAssign efficiently binds known cell types to LCs and the results help to decipher the cell type composition of samples without the need of reference methylomes, by utilizing the available single-cell RNASeq data.

Applications

Cell-type identification via functional enrichment analysis of Single-cell RNA-seq data

Marta Benegas Coll (BioBam Bioinformatics S.L.) and Stefan Götz (BioBam Bioinformatics S.L.).

Abstract:

Single-cell omics technologies have been increasing in popularity during the last years. In particular, single-cell RNA sequencing (scRNA-seq), is one of the most dominant application domains of single-cell omics and has revolutionized the field of transcriptomics.

The most common first step in any single-cell transcriptomic analysis consists of identifying the cell types present in the analyzed samples. This step requires prior knowledge of genes specific to each cell type, also known as marker genes. However, this information is rarely available, especially for non-model organisms. This work presents an initial version of an analysis pipeline that attempts to annotate the cell types through functional enrichment analysis. The presented pipeline only requires the reference genome or transcriptome as prior knowledge while the functional annotation can be generated through the process. This makes this approach especially useful for non-model organisms.

Here, we show the analysis results of the pipeline being applied to a scRNA-seq dataset consisting of Drosophila melanogaster eye disc cells. We have found evidence that cell types can be identified via biological functions. However, further validation both in silico and experimental has to be carried out to confirm the presented results.

Applications

ChIA-BERT: prediction of CTCF-mediated chromatin loops identified by Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) from DNA sequence

Mateusz Chiliński (Warsaw University of Technology) and Dariusz Plewczynski (Centre of New Technologies, University of Warsaw, Warsaw, Poland).

Abstract:

The spatial architecture of the human genome is considered to play a major role in controlling biological processes in a cell. The spatially close DNA regions, while linearly distal, can interact with each other, thus regulating the expression of genes. One of the experimental methods for the identification of statistically important 3D interactions of chromatin fiber is Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET), which observes chromatin loops mediated by CCCTC-binding factor proteins. However, not always an experimentally identified chromatin loops are possible to obtain.

That is why multiple statistical learning algorithms have been proposed to simulate in-sillico 3D genomics experiments. We have developed ChIA-BERT, a deep learning algorithm based on transformers (namely BERT-like architecture). We can predict from DNA sequence chromatin loops mediated by CTCF with an accuracy of up to 78%. The machine learning algorithm uses as input two DNA sequence segments that are interacting, and as the negative set, we use random segments of the remaining genome that are not interacting in 3D space.

Our results show clearly that the modern-day deep learning methods can predict chromatin looping from the DNA sequence. The proposed approach can have a major impact on creating in-sillico statistical models extrapolating the knowledge gathered from molecular biology experiments. DNA sequence is easily available from the modern next generation sequencing methods with the decreasing sequencing costs The improvements in the in silico predictions from DNA sequence have a major impact on functional studies allowing to predict the effect of mutations on gene expression.

Applications

Choosing variant interpretation tools for clinical applications: context matters

Xavier de la Cruz (Vall d'Hebron Institute of Research (VHIR)), Natàlia Padilla (Vall d'Hebron Institute of Research (VHIR)), Josu Aguirre (Vall d'Hebron Institute of Research (VHIR)), Selen Özkan (Vall d'Hebron Institute of Research (VHIR)), Casandra Riera (Vall d'Hebron Institute of Research (VHIR)) and Lidia Feliubadaló (Catalan Institute of Oncology (ICO), ONCOBELL-IDIBELL).

Abstract:

Our inability to solve the Variant Interpretation Problem (VIP) has become a bottleneck in the biomedical/clinical application of Next-Generation Sequencing. This situation has favored the development and use of in silico tools for the VIP. However, choosing the optimal tool for our purposes is difficult because of a fact usually ignored: the high variability of clinical context/scenarios across and within countries, and over time.

We present a computational procedure, based on the use of cost models, that allows the simultaneous comparison of an arbitrary number of tools across all possible clinical scenarios. We apply our approach to a set of pathogenicity predictors for missense variants, showing how differences in clinical context translate to differences in tool ranking.

Applications

Combined Quality Assessment of Different Long-Reads Sequencing Methods

Enrique Presa-Díez (Biobam Bioinformatics S.L), Adolfo López-Cerdán (Biobam Bioinformatics S.L) and Stefan Götz (Biobam Bioinformatics S.L).

Abstract:

Third Generation Sequencing (TGS) paradigm has achieved a notable increase in relevance by surpassing some of the limitations of the NGS technologies. All of these TGS methodologies share the capacity to produce long reads.

In the case of transcriptomics the analysis of alternative splicing and alternative polyadenylation has been limited by the length of short reads. Long reads allow direct sequencing of whole transcripts and eliminate the need for short-read assembly. Transcriptome TGS can be performed mainly using Oxford Nanopore Technologies (ONT) and Pacific Bioscience (PacBio) Systems, two technologies with differences that need independent assessment. This assessment is part of the Long-read RNA-seq Genome Annotation Assessment Project (LRGASP). The LRGASP Consortium has been gathering different long-read sequencing datasets of model and non-model organisms to rate different methods for transcript identification.

The main goal of the presented work is to characterize the strengths and potential remaining challenges in using TGS technologies to annotate and quantify the transcriptomes of both model and non-model organisms using LRGASP datasets. We propose here the combined use of the following two tools: LongQC, a quality control tool for datasets generated by TGS technologies, and SQANTI3, a tool designed to characterize a new long-read-defined transcriptome. Both implementations can be used inside OmicsBox, providing complementary information about the quality of long reads and the quality of the transcriptome that has been generated, filtering false positive transcripts. The use of these tools will allow to polish sequenced transcriptomes with improved quality and downstream analysis.

Applications

Continuous reference color scheme for genetic ancestry using CIELAB

Felix Pacheco (Novo Nordisk Center for Protein Research), Mikaela Koutrouli (Novo Nordisk Center for Protein Research), Karina Banasik (Novo Nordisk Foundation Center for Protein Research), Lars Juhl Jensen (Novo Nordisk Foundation Center for Protein Research) and Søren Brunak (Novo Nordisk Foundation Center for Protein Research).

Abstract:

Visualizing population genetic variation of individuals should ideally be category-free and continuous, as categorical labels do not represent genetic data given that individuals are a mixture of ancestries [1].

Here, we present a continuous reference color scheme to be used for visualizing genetic ancestry. Our approach overcomes the previous issues with ancestry categorization. Our reference color space was created using the Human Origins dataset. The data was first projected into three dimensions using PCA and further projected into the CIELAB color space [2].

Our three-dimensional color space can be used to visualize and understand ancestry as a label-free and continuous measure. Any dataset can be projected into this color scheme whereby individuals are assigned to a color that represents their ancestry. We illustrate our methods applicability by visualizing the impact of ancestry on laboratory clinical values. This work is a step towards an ethical representation of genetic ancestry, enhancing its inclusion in health strategies.

1. Lewis ACF, Molina SJ, Appelbaum PS, Dauda B, Di Rienzo A, Fuentes A, et al. Getting Genetic Ancestry Right for Science and Society.

2. Koutrouli M, Morris JH, Jensen LJ. U-CIE [/juː ’siː/]: Color encoding of high-dimensional data. 2021;:2021.12.02.470966.

Applications

CROSTA: CROss-Species Transmissibility Analyser for Pathogen Sequences

Shan Tharanga (Centre for Bioinformatics, School of Data Sciences, Perdana University, Kuala Lumpur, Malaysia), Eyyüb Selim Ünlü (Istanbul University), Esra Büşra Işık (Bezmialem Vakif University), Muhammad Farhan Sjaugi (Perdana University) and Mohammad Asif Khan (Bezmialem Vakif University).

Abstract:

Pathogen cross-species transmission is a major health threat. Transmission of pathogens from a natural reservoir host to a recipient host is facilitated by adaptive substitutions in protein sequence that allow pathogen entry and/or escape. Comparative analysis of sequence changes in the pathogen populations between the donor reservoir and the recipient host can enable identification of substitutions that are evolutionarily selected for the latter. CROSTA compares pathogen protein/DNA sequence of two host populations that are co-aligned and are analyzed for cross-species transmission. The comparison relies on a sliding window approach, based on a user-selected k-mer length. The distinct sequences at each aligned overlapping k-mer positions are categorized as diversity motifs, the index sequence, and its variants (major, minor, and unique) based on their incidence in the alignment. Index is the predominant sequence, major is the second most common, while minors are those in-between major and unique (observed only once). CROSTA provides users the incidences of each of the diversity motifs across the k-mer positions, with a simple interface, graphical representation of the results, and built-in analyses. These include multi-motif analysis and transmission candidates, facilitating the determination of positively selected substitutions. CROSTA also enriches the analysis with the inclusion of metadata, such as date of isolation, geographical area. Application of CROSTA is demonstrated with zoonosis analysis of SARS-CoV-2 and the tool is available at https://crosta.bioinfo.perdanauniversity.edu.my. The results of CROSTA can be used for a better understanding of substitution transmissibility from the natural reservoir to the recipient host, which can facilitate surveillance efforts.

Applications

Current activities of the ELIXIR Machine Learning Focus Group

Fotis Psomopoulos (CERTH), Emidio Capriotti (University of Bologna), Núria Queralt Rosinach (Leiden University), Machine Learning Focus Group Tasks Participants (ELIXIR), Leyla Jael Castro (Germany ZB MED – Information Centre for Life Sciences) and Silvio Tostto (University of Padova).

Abstract:

Machine Learning (ML) has emerged as a discipline that enables computers to assist humans in making sense of large and complex data sets. With the drop in the cost of high-throughput technologies, large amounts of omics data are being generated and made accessible to researchers. Analyzing these complex high-volume data is not trivial, and the use of classical statistics cannot explore their full potential. Machine Learning can thus be very useful in mining large omics datasets to uncover new insights that can consequently lead to the advancement of Life Sciences.

The ELIXIR Machine Learning Focus Group was initiated in October 2019, in order to capture the emerging need in Machine Learning expertise across the network. In addition to producing the DOME Recommendations, a set of community-wide recommendations for reporting supervised machine learning–based analyses applied to biological studies, the Focus Group is currently working on three main activities; (1) using the DOME recommendations to annotate relevant literature in order to gain insights into the level of adherence to DOME, (2) evaluating the gold standard datasets widely used in ML process in order to define and describe the aspects of a gold standard with particular focus on human data, and (3) reviewing the efforts around synthetic data in order to establish a set of best practices for their use and application in ML.

Applications

DA4LT: an efficient computational method for cross-species label transfer from single cell data

Rawan Olayan (Harvard University), Sergio Picart-Armada (Boehringer Ingelheim Pharma GmbH & Co. KG), Gregorio Alanis-Lobato (Boehringer Ingelheim Pharma GmbH & Co. KG), Francesc Fernandez-Albert (Boehringer Ingelheim Pharma GmbH & Co. KG), Timothy B Sackton (Harvard University) and Stefano Patassini (Organization:Boehringer Ingelheim Pharma GmbH & Co. KG).

Abstract:

Single-cell omics have shed light on the cell type diversity of human and animal tissues. This advanced our understanding of the developmental and differentiation processes of complex biological systems. One of its most challenging tasks is the cell-type annotation aiming to identify cell types for a newly generated un-labeled dataset from cell types in an annotated reference dataset. It is also key to transferring knowledge from model organisms to humans. Despite the remarkable progress toward the development of label transfer methods, current approaches show limited performance when applied to cell types from different species. We present DA4LT, an efficient computational method to perform cross-species cell-type label transfer from single cell data, at low cost with reasonable accuracy. Here, we focus on the comparison between major neuronal cell types in the mouse and human cortex. Our methodology starts with generating pathway-based informative features that determine the specific characteristic of each cell type. Then, we align cell types from different species datasets into a shared space using canonical correlation analysis. Over this shared space, we identify aligned cell types with similar pathway activation signatures. Finally, we provide and rank potential lists of pathways and pathway-enriched genes in which cell types can be compared across different species datasets. By benchmarking with other methods, we show that DA4LT improves the performance of the cell-type annotation than when annotating cells based on the gene-level expression. This suggests that DA4LT would serve as a useful method to build well-annotated datasets and develop more accurate disease models.

Applications

fCAT - Assessing gene set completeness using domain-architecture aware targeted ortholog searches

Vinh Tran (Goethe University) and Ingo Ebersberger (Goethe University).

Abstract:

The assessment of gene set completeness is a routine task in genome analysis. The standard workflow starts with the identification of a set of single copy core genes for the taxonomic group the newly sequenced species, the target, is part of. The fraction of missing core genes serves then as a proxy of the target gene set completeness. Genes that are represented but differ significantly in length from the expectation provide information about the fragmentation status of the predicted gene models, and ultimately gene duplication levels can be assessed. Though well established, this approach comes along with a number of restrictions of which the focus on single copy genes, the use of an error-prone unidirectional ortholog search, and the application of a simple length cutoff criterion are the most prominent ones.

Here we present fCAT, a novel algorithm for assessing gene set completeness using a targeted and domain architecture-aware ortholog search. As a consequence, fCAT's core sets are not limited to single-copy orthologs by that providing a comprehensive overview of the core gene set. Next to the conventional length difference assessment, fCAT identifies target genes that significantly differ in their domain architectures from the core genes allowing an alternative view on the accuracy of target gene models. Phylogenetic profiles resulting from the analysis can be visualized and explored in the context of the entire orthologous groups which provides the necessary information to identify and ultimately correct erroneous gene annotations.

Applications

FrankenSeq: A user-friendly and modular machine learning package for Single-cell RNA-Seq Cluster Analysis

Edward Agboraw (University of Glasgow) and Quan Gu (University of Glasgow).

Abstract:

The advent of single-cell RNA sequencing (scRNA-seq) and next-generation sequencing (NGS) technologies has led to a massive increase in detailed transcriptomic data, enabling researchers to analyze the transcriptomes of individual cells and interrogate heterogeneous gene expression patterns within single tissues. In scRNA-seq analysis, unsupervised clustering is the process of dividing cells into different subpopulations based on gene expression patterns, in data where true cell labels are unknown. Here we present FrankenSeq, a modular R-Shiny-based application that provides a user-friendly interface for single-cell cluster analysis, requires minimal background knowledge of programming or machine learning to use and allows users to visualize how different feature selection, dimension reduction, and clustering algorithms (both traditional and deep learning) impact final cluster assignments.

Applications

Functional Enrichment Analysis of Regulatory Elements with GeneCodis4

Adrian Garcia-Moreno (Centre for Genomics and Oncological Research (GENyO)), Raul Lopez-Dominguez (Centre for Genomics and Oncological Research (GENyO) & Dpt. Statistics and Operational Research University of Granada), Juan Antonio Villatoro-Garcia (Centre for Genomics and Oncological Research (GENyO) & Dpt. Statistics and Operational Research University of Granada), Samuel Perez-Fernandez (Centre for Genomics and Oncological Research (GENyO)) and Pedro Carmona-Saez (Centre for Genomics and Oncological Research (GENyO) & Dpt. Statistics and Operational Research University of Granada).

Abstract:

Functional enrichment analyses are used to extract biological information from omics experiments. These are widely used to examine gene and protein lists, however the development of high-throughput technologies for regulatory elements demands new bioinformatics approaches. On most occasions, given our current knowledge, the functional characterisation of regulation networks require to be inferred via the participant genes. This standard approach, based on the hypergeometric distribution, has been proven to lead to biassed results that limit the biological interpretation. Here we present an update of GeneCodis with novel and state of art statistical methods, based on the Wallenius distribution, to analyse lists of regulatory elements such as microRNAs, TFs and CpGs, besides, genes, proteins. Additionally, microRNAs annotation databases are included to support the standard approach. As a result, GeneCodis4 becomes a functional enrichment web tool that allows the integration of heterogeneous information, discovering co-annotations and being able to study different biological entities in a single tool.

Applications

GeneHTracker: improving reproducibility and reusability of datasets based on gene identifiers.

Hugo Guillen-Ramirez (University of Bern; University College Dublin), Daniel Sanchez-Taltavull (University of Bern) and Rory Johnson (University of Bern; University College Dublin).

Abstract:

Gene identifiers are fundamental for downstream analysis in a variety of bioinformatics subfields. Particularly, ENSEMBL/GENCODE IDs are widely accepted for the community to share results. However, although these IDs are stable, they are still susceptible to be retired or deprecated, the gene type could be changed, or even a new ID can be assigned to a given locus. These problems limit the reproducibility and reusability of published gene lists. In order to mitigate the impact of ID changes, we developed the Python module GeneHTracker. GeneHTracker is able to 1) retrieve the latest coordinates for a list of gene IDs; 2) find the complete (GENCODE) annotation history of a gene ID or gene symbol; 4) generate putative mappings for deprecated/retired IDs based on overlapping annotations curated since Gencode version 5/Ensembl version 60, representing almost 12 years of annotations; and 5) build a local index of custom annotations. GeneHTracker is available at GitHub.

Applications

Implications of phenotypic and functional heterogeneity in cancer cells

Asadullah (IIT BOMBAY) and Shamik Sen (IIT Bombay).

Abstract:

Phenotypic heterogeneity in cancer is increasingly appreciated to play important roles in cancer invasion and drug resistance. Given the difficulty in quantifying phenotypic heterogeneity in vivo, most studies have focused on quantifying genetic heterogeneity. Nevertheless, the extent of variation in physical properties of tumour cells and the presence of different population of cells (functionally heterogeneous) in the environment and the importance of this variability vis-a-vis invasiveness remain largely unknown.

Recently, we had shown that combined heterogeneity in cell size and deformability enhances cancer invasion, with gradual enrichment of small cells at the invasive front [Asadullah et al., Journal of Cell science (JCS), 2021].

Therefore, presently we are trying to address this question by first documenting variability in cell spread area and deformability (phenotypic heterogeneity) in proteolytic invasive MDA-MB-231 breast cancer cells and non-proteolytic MDA-MB-231 (MMP [Matrix Metalloproteinase]-Knockdown) cells. The interacting group of MMP-secreting and non-MMP secreting cells in a tumour environment forms the functional heterogeneity. A computational model (using Compucell3D) is generated with heterogeneous group of cells, comprised of the above mentioned types in different ratios and then their translocation efficiency and percentage enrichment is studied. The preliminary analysis suggests that there is co-operative behaviour among the different cell types that is non-proteolytic cells can disseminate from its site of origin when in conjunction with even lower count of proteolytic cells (i.e., 29% of the total population), without undergoing any clonal selection or phenotype switching.

Computational tools presented here will open new avenues for experimental studies pertaining to cancer invasion.

Applications

Mathematica as a platform to develop tools for education, data representation, and analysis in epidemiology

Rui Alves (Universitat de Lleida), Ester Vilaprinyo (Universitat de Lleida), Alberto Marin-Sanguino (Universitat de Lleida) and Albert Sorribas (Universitat de Lleida).

Abstract:

The past two years saw a rise in the public profile of epidemiology as a science, because of the COVID19 pandemic. Many political and policy decisions were based on the use of pandemic data to fit mathematical models that enabled understanding and predicting the consequences of those decisions on the dynamics of the pandemic. This created many needs, at different levels: a) to educate a general audience about mathematical concepts in epidemiology, b) to quickly represent and compare data about the pandemic, and c) to have easy ways to calculate epidemiological parameters from the data.

Typically, we use different computational platforms to meet each need. Still, because of their interconnectedness, using a single platform to address them all is desirable for developers.

Here we present three tools that address the three needs described above using the Mathematica platform. EpiPlay is an application that uses gaming to explain key concepts in epidemiology from a quantitative perspective and using mathematical models as a motor for the game [1]. COVIDWORLD downloads, graphically represents, and compares live data about the state of the pandemic in different country, continents and territories [2]. Finally, EpiFit is a Mathematica notebook that allows users to import epidemiological data, fit it to different epidemiological models and calculate the best guess parameters for that data.

References

1 – Alves et al. (2022) COVIDWORLD app: current data and visualizations for SARS-CoV2 pandemic. https://community.wolfram.com/groups/-/m/t/2473065.

2 – Alves et al. (2022) EpiPlay: using Wolfram Language to gamify education in epidemiology. https://community.wolfram.com/groups/-/m/t/2535927

Applications

MicroNAP-web: A novel framework to characterize genomic and transcriptomic changes in engineered microbial strains based on Next Generation Sequencing data.

Veronika Schusterbauer (Graz University of Technology), Daniel Degreif (Microbial Platform USP Development, Sanofi-Aventis Deutschland GmbH), Christoph Schiklenk (Microbial Platform USP Development, Sanofi-Aventis Deutschland GmbH), Anton Glieder (bisy GmbH) and Gerhard Thallinger (Graz University of Technology).

Abstract:

Engineering of microbial cell factories is a tedious process in which the reason of success or failure often remains hidden. Whole genome sequencing and RNA-seq are the perfect tools to reveal off-targeting and integration locus effects. Nowadays the costs and time for data analysis, however, surpass the resources required for sequencing itself.

MicroNAP-web aims to close that gap. By offering a simple user interface MicroNAP-web allows biotechnicians without special bio-IT skills to analyze their own next generation sequencing data and to document all relevant metadata. The MicroNAP-web framework offers two analysis pipelines, MicroNAP and MicroRAP, facilitating the analysis of short-read whole genome sequencing and RNA-seq data, respectively. The MicroNAP pipeline especially improved the calling and visualization of complex structural variants compared to available tools. It was tested on Illumina sequencing data of Komagataella phaffii strains harboring 15 gene knockouts and 10 random vector insertions. MicroNAP was able to detect and resolve the exact sequence for 12 out of 15 knockouts and detected 8 out of 10 insertion sites. Performance on an in-silico test set of simulated WGS reads of mutated K. phaffii genomes, which included all common types of simple and complex SVs, was equally successful. With sufficient sequencing depth and library fragment size, 75% of SVs could be fully reconstructed.

All pipeline results including small and large variants or differentially expressed genes and pathways, contamination analysis and quality controls are presented in concise and appealing human readable PDF reports, making interpretation, and sharing of results as easy as possible.

Applications

MinSizeML: An R Package to Estimate the Minimum Sample Size in Supervised Learning for Classification

Guillermo Prol Castelo (Universitat Oberta de Catalunya - UOC) and Jose Luis Mosquera (Universitat Oberta de Catalunya - UOC and Bioinformatics Unit, Institut d'Investigació Biomèdica de Bellvitge - IDIBELL).

Abstract:

Knowing how to estimate the minimum samples required in biomedical studies is a crucial issue. This estimate is not straight-forward when using machine learning approaches. To tackle this challenge on supervised learning methods (e.g. kNN, logistic regression, naive Bayes, and random forest) we have developed an algorithm based on already-existing methodologies that fits an inverse power law to the learning curve and proves effective in estimating minimum sample size in the training step for some datasets and learning algorithms. This fitted curve suggests a "sweet spot" where increasing the training sample size does not improve predictions further. Moreover, we also propose some fitting alternatives to improve the decrease in accuracy and Cohen's kappa as the sample size surpasses a threshold. In this work, we present our algorithm and an open-source R package aiming to ease the sample size estimate in supervised learning for classification even for researchers who are not specialists.

Applications

Multi-Omics Analysis and Metabolic Network Construction in suberinization of cork oak.

Nuria Mauri (Centre for Research in Agricultural Genomics), Maria Verdum (Institut Català del Suro), Patrícia Jové (Institut Català del Suro), Ignacio Ontañon (Universidad de Zaragoza), Jordi Roselló (Francisco Oller S.A.), Vicente Ferreira (Universidad de Zaragoza) and David Caparrós-Ruíz (Centre for Research in Agricultural Genomics).

Abstract:

Suberin is a specialized polymer present in the cell wall of dermal tissues growing above and below ground, which provides a hydrophobic layer to prevent water loss and pathogen infection. Wounding experiments in potato tubers and waterlogging in Arabidopsis roots have been largely used as a model system to study suberin deposition. However, the periderm of cork oak presents an specific composition and structure that represents an opportunity to advance further in the knowledge on this polymer.

The main aim of this study is to investigate the metabolic networks involved in Quercus suber suberinization with particular emphasis on the phenylpropanoid biosynthesis by combining multi-omics data analysis with metabolic network construction methods to predict novel reactions and interactions. For this purpose, two different populations belonging to 'suber' and 'ilex' chlorotype lineages with different mechanical and quality cork properties were studied.

Transcriptomics and metabolomic results expands the understanding of the cork oak biology in some aspects such as the participation of cell wall peroxidases and esterases in the assembly of the suberin polymer, or the role of phytohormones in its regulation. In addition, the genetic variation described in this work may contribute to breeding and conservation programs for these species, for instance, in responding to drought-stressed conditions.

Applications

New algorithms for accurate and efficient de-novo genome assembly from long DNA sequencing reads

David Guevara-Barrientos (Universidad de los Andes), Laura Natalia González García (Universidad de los Andes), Daniela Lozano (Universidad de los Andes), Juanita Gil (Universidad de Los Andes), Maria Camila Hoyos (Universidad de los Andes), Christian Chavarro (Universidad de los Andes), Natalia Guayazan (Universidad de los Andes), Luis Chica (Universidad de los Andes), Maria Camila Buitrago (Universidad de los Andes), Edwin Bautista (Universidad de los Andes), Juan Camilo Bojacá (Universidad de los Andes), Miller Andrés Trujillo Achury (Los Andes) and Jorge Duitama (Universidad de los Andes).

Abstract:

Producing high-quality de-novo genome assemblies for complex genomes is possible thanks to the development of long read DNA sequencing technologies. We present here new algorithms for assembly of haploid and diploid samples from long DNA sequencing reads. Our algorithm builds an undirected graph having two vertices for each read. A minimizers table is constructed from the reads to identify overlaps between the longest sequences in linear time. K-mer hash codes are calculated based on rankings relative to the mode of the k-mer counts distribution. Statistics collected during this process include an overlap estimation, a coverage of shared k-mers (CSK), and the percentage of the predicted overlap supported by k-mers. These statistics are used as features to build layout paths following a Naive Bayesian machine learning approach in which selecting path edges can be thought of as a binary classification problem. For diploid samples, we integrated a reimplementation of the ReFHap algorithm to perform molecular phasing. The phasing procedure is used to remove edges connecting reads assigned to different haplotypes and obtain a phased assembly running the layout algorithm on the filtered graph. We ran the implemented algorithms on PacBio HiFi and Nanopore sequencing data taken from bacteria, Drosophila, rice, maize and human samples. Our algorithms showed competitive contiguity and efficiency, as well as superior accuracy in some cases, compared to other currently used software. We expect that this new development will be useful for researchers currently building genome assemblies for different species.

Applications

Organisation-wide data discovery at the European Bioinformatics Institute

Matthew Pearce (EMBL), Prasad Basutkar (EMBL), Ossama Edbali (EMBL), Anton Kolesnikov (EMBL), Henning Hermjakob (EMBL) and Rodrigo Lopez (EBI).

Abstract:

The EMBL-EBI Search application (https://www.ebi.ac.uk/ebisearch/) provides a full-text search engine across nearly 5 billion entries drawn from more than 150 sources across EMBL-EBI and selected external resources. It allows users to search metadata from across these resources using a user-friendly web application, as well as via a RESTful API. Data is categorised into hierarchical domains, updated nightly, and can be retrieved in a number of formats, including XML, JSON, and CSV. The option to stream full result sets has recently been added for some domains, allowing retrieval beyond the normal 1-million result limit (with some limitations). The search engine uses Lucene Core technology, providing functionality including multi-level faceting, bi-directional cross-references to other data sources, and searches against sequence tool analysis results. It provides the primary search results for a number of EMBL-EBI websites, including the COVID-19 Data Portal.

Applications

PaintOmics 4: new tools for the integrative analysis of multi-omics datasets supported by multiple pathway databases

Tianyuan Liu (Cardiff University), Pedro Salguero (Universitat Politècnica de València), Marko Petek (National Institute of Biology), Carlos Martinez-Mira (Biobam Bioinformatics), Leandro Balzano-Nogueira (University of Florida), Živa Ramšak (National Institute of Biology), Lauren McIntyre (University of Florida), Kristina Gruden (National Institute of Biology), Sonia Tarazona (Universitat Politècnica de València) and Ana Conesa (Spanish National Research Council).

Abstract:

PaintOmics is a web server for the integrative analysis and visualisation of multi-omics datasets using biological pathway maps. PaintOmics 4 has several notable updates that improve and extend analyses. Three pathway databases are now supported: KEGG, Reactome and MapMan, providing more comprehensive pathway knowledge for animals and plants. New metabolite analysis methods fill gaps in traditional pathway-based enrichment methods. The metabolite hub analysis selects compounds with a high number of significant genes in their neighbouring network, suggesting regulation by gene expression changes. The metabolite class activity analysis tests the hypothesis that a metabolic class has a higher-than-expected proportion of significant elements, indicating that these compounds are regulated in the experiment. Finally, PaintOmics 4 includes a regulatory omics module to analyse the contribution of trans-regulatory layers (microRNA and transcription factors, RNA-binding proteins) to regulate pathways. We show the performance of PaintOmics 4 on both mouse and plant data to highlight how these new analysis features provide novel insights into regulatory biology.

Applications

Predicting clinical response to immunotherapy in advanced melanoma

Matthieu Genais (CRCT), Bruno Ségui (CRCT), Anne Montfort (CRCT) and Vera Pancaldi (CRCT).

Abstract:

Immune checkpoint inhibitors (ICI) such as anti-PD-1 act on exhausted/inactivated T cells to restore their ability to kill cancer cells. In that context our team has identified a mechanism of resistance to immunotherapy that depends on the production of TNF, which acts as a brake on the immune response against tumours. Bulk RNA-seq on tumour samples from TNF knock-out (KO) mice and wild type mice were analysed by deconvolution of cell types and inference of transcription factor activities. We showed that knocking out TNF-alpha seems to have an inconsistent impact on mice. Indeed, computing differential transcription factor (TF) activities between TNF KO mice and WT mice with a list of up-activated and down-activated TFs. TF activities of the first group of mice correlated positively with immune scores and infiltrated deconvolved immune cells (CD8) and the opposite for the second one. We then classified mice samples into hot and cold tumours, respectively.

Finally, using this TNF KO TF activity signature in human melanoma patients (pre-treatment to anti-PD1 therapy), we distinguished the same two profiles that we had seen in mice by creating a TF activities based score (TF score). We saw an enrichment of responders to anti-PD1 treatment for high score patients compared to low score patients. High score patients in the pre-treatment condition seem to exhibit a similar profile to TNF KO mice, suggesting that TNF might not be essential for hot tumour patterns in patients who are the most susceptible to respond to immunotherapy, as observed in mice.

Applications

Secured and annotated execution of workflows with WfExS-backend

José M. Fernández (Barcelona Supercomputing Center, INB/ELIXIR-ES), Laura Rodríguez-Navas (Barcelona Supercomputing Center, INB/ELIXIR-ES) and Salvador Capella-Gutierrez (Barcelona Supercomputing Center, INB/ELIXIR-ES).

Abstract:

WfExS-backend (https://github.com/inab/WfExS-backend) is a high-level workflow execution command line program, which fetches and materialize all the elements needed to run a workflow: the workflow itself, the engine, the needed software containers and the inputs. The program both creates and consumes RO-Crates, focusing on the interconnection of content-sensitive research infrastructures for handling sensitive human data analysis scenarios. WfExS-backend delegates workflow execution on existing workflow engines, e.g. Nextflow or cwltool, and it is designed to facilitate secure and reproducible workflow executions to promote analysis reproducibility and replicability using sensitive data. Secure executions are achieved using FUSE encrypted directories for non-disclosable inputs, intermediate workflow execution results and output files.

RO-Crate representations of those executions are, indeed, an element of knowledge transfer between repeated workflow executions. WfExS-backend stores all the gathered execution details, output metadata and execution provenance in the output RO-Crate to achieve future reproducible executions. Final execution results can be encrypted with GA4GH crypt4gh standard using the public keys of the target researchers or destination so that the results can be safely moved outside the execution environments through unsecured networks and storage.

Future developments are focused on ironing secure data export procedures and supporting other workflow engines, like SnakeMake or Galaxy.

Applications

SMASCH: Facilitating multi-appointment scheduling in longitudinal clinical research studies and care programs.

Carlos Vega (Luxembourg Centre for Systems Biomedicine), Piotr Gawron (Luxembourg Centre for Systems Biomedicine), Jacek Lebioda (Luxembourg Centre for Systems Biomedicine), Valentin Grouès (Luxembourg Centre for Systems Biomedicine), Rejko Krüger (Luxembourg Centre for Systems Biomedicine), Reinhard Schneider (Luxembourg Centre for Systems Biomedicine) and Venkata Satagopam (Luxembourg Centre for Systems Biomedicine).

Abstract:

Longitudinal clinical research studies require conducting various assessments over long periods of time. Such assessments comprise numerous stages, requiring different resources defined by multidisciplinary research staff and aligned with available infrastructure and equipment, altogether constrained by time. While it is possible to manage the allocation of resources manually, it is complex and error-prone. Efficient multi-appointment scheduling is essential to assist clinical teams, ensuring high participant retention and producing successful clinical studies, directly impacting patient throughput and satisfaction.

We present Smart Scheduling (SMASCH) system [1], a web application for multi-appointment scheduling management aiming to reduce times, optimise resources and secure personal identifiable information.

SMASCH facilitates clinical research and integrated care programs in Luxembourg, providing features to better manage multi-appointment scheduling problems (MASPs) characteristic of longitudinal clinical research studies and speed up management tasks. It is present in multiple clinical research and integrated care programs in Luxembourg since 2017, including Dementia Prevention Program, the study for Mild Cognitive Impairment and gut microbiome, and the National Centre of Excellence in Research on Parkinson's disease [2] which encompasses the study for REM sleep behaviour disorder and the Luxembourg Parkinson's Study.

SMASCH is a free and open-source solution available both as a Linux package and Docker image.

[1] Vega, C. et al. Smart Scheduling (SMASCH): multi-appointment scheduling system for longitudinal clinical research studies. JAMIA Open, 2022.

[2] Hipp, G., et al. The Luxembourg Parkinson's study: a comprehensive approach for stratification and early diagnosis. Frontiers in aging neuroscience, 2018, p. 326.

Applications

The Bioinfo4Women Research Working Group on Sex and Gender Bias in Healthcare and AI

Atia Cortés Martínez (Barcelona Supercomputing Center), Davide Cirillo (Barcelona Supercomputing Center), Maria José Rementeria (Barcelona Supercomputing Center) and Nataly Buslon (Barcelona Supercomputing Center).

Abstract:

Historically, scientific research has been led by and focused on males, and in the case of research with human beings, these were traditionally white men. The introduction of Artificial Intelligence (AI) in precision medicine has proven that these practices were discriminating for minorities, and more specifically for women. Sex and gender biases are found at all stages of the AI lifecylce, from design to data acquisition, deployment and use. To move towards a fair development of new technologies, it is essential to include sex and gender diversity both at workplace and in research practices. In this context, the Bioinfo4women (B4W) program of the Barcelona Supercomputing Center works in three axes: (i) promote the participation of women scientists by improving their visibility, (ii) foster international collaborations between institutions and programs and (iii) advance research on sex and gender bias in AI and health.

This poster introduces the objectives of the B4W research working group and the first two actions taken in 2021: (i) Raise awareness about sex and gender differences in healthcare and biases in biomedical research and AI through a series of conferences and open debate between scientific community and society; (ii) Build best practices for biomedical dataset creation and use as a result of a qualitative and quantitative analysis done at the Elixir Biohackathon 2021. These initial steps are fundamental to guarantee the scientific quality of the results and impact in society and to develop a science of excellence.

Applications

The ELIXIR::GA4GH Cloud

Alexander Kanitz (Biozentrum, University of Basel, Basel, Switzerland & Swiss Institute of Bioinformatics, Lausanne, Switzerland), Justin Clark-Casey (European Bioinformatics Institute, EMBL-EBI, Hinxton, United Kingdom), Michael Crusoe (Vrije Universiteit Amsterdam, Amsterdam, The Netherlands), Gavin Farrell (ELIXIR Hub, ELIXIR, Hinxton, United Kingdom), Alvaro Gonzalez (CSC, Finnish IT Center for Science, Espoo, Finland), Thanasis Vergoulis ("Athena" Research Center, Athens, Greece) and Jonathan Tedds (ELIXIR Hub, ELIXIR, Hinxton, United Kingdom).

Abstract:

The Global Alliance for Genomics and Health (GA4GH) is an international organization bringing together opinion leaders in the life sciences and the biomedical sector to define open community standards and guidelines for the responsible sharing of omics data, and the provision of FAIR software and infrastructure for their analysis.

The ELIXIR Cloud and Authentication and Authorization Infrastructure (AAI) Initiative is developing the ELIXIR::GA4GH Cloud (EGC), a federated analysis platform composed of standalone data access, storage and processing services implementing dedicated, GA4GH-defined API specifications and common authentication/authorization guidelines. This design ensures a high degree of interoperability, agnosticism with regard to underlying technologies (e.g., workflow engines, compute solutions) and extensibility. The EGC is under heavy development, but an incremental rollout to end users is planned for the end of 2022.

Are you a systems administrator and would you like to make EGC services available to your institution, or would you like to make your institutional resources available to a wider community? Are you a developer of a relevant tool or service and are you interested in implementing GA4GH Cloud API specifications to reach a bigger audience and make your solution interoperate with related services? Are you part of a large-scale data analysis project and are you interested to learn how cloud computing via the EGC might benefit you, or are you interested in driving its further development? If you have answered "Yes!" to any of these questions, please do not hesitate to reach out to us.

Applications

The Job Dispatcher sequence analysis tools services from EMBL-EBI in 2022

Fábio Madeira (European Bioinformatics Institute (EMBL-EBI)), Joon Lee (European Bioinformatics Institute (EMBL-EBI)), Adrian Tivey (European Bioinformatics Institute (EMBL-EBI)), Nandana Madhusoodanan (European Bioinformatics Institute (EMBL-EBI)), Sarah Butcher (European Bioinformatics Institute (EMBL-EBI)) and Rodrigo Lopez (European Bioinformatics Institute (EMBL-EBI)).

Abstract:

The Job Dispatcher (JD) tools framework [1] provides free access to popular bioinformatics sequence analysis applications and sequence libraries. The services can be accessed via user-friendly web interfaces and via established RESTful and SOAP Web Services Application Programming Interfaces (APIs), allowing them to be integrated into third-party systems. In fact, the services reported here are integrated into many popular data resources provided at the EMBL-EBI, including for example UniProtKB, ENA, Ensembl Genomes and InterPro. Tools running under JD comprise popular bioinformatics applications (e.g. Clustal Omega, InterProScan, FASTA, NCBI BLAST+, and many more), which combined with nucleotide and protein sequence databases (e.g. UniProtKB and ENA) represent powerful tools for SARS-CoV-2 and COVID-19 research. Countless jobs have been performed on EMBL-EBI's high-performance computing clusters, with a noticeable surge of more than 60 million jobs per month performed during the COVID-19 outbreak months of April and May 2020. Nearly 500 million analyses were executed during 2021 by over 700,000 unique users.

Here, we would like to describe the latest improvements made to the framework, focusing on the biological databases indexing pipelines. We would like to provide an update on the datasets and tools provided in JD and describe our current and future goals.

[1] Fábio Madeira, Matt Pearce, Adrian R N Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov and Rodrigo Lopez. 2022. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research, 2022, gkac240, https://doi.org/10.1093/nar/gkac240

Applications

The Service Collections of the ELIXIR Rare Disease Community

Emidio Capriotti (University of Bologna), David Salgado (French Bioinformatics Institute), Friederike Ehrhart (Maastricht University.), Allegra Via (CNR), Claudio Carta (ISS), Marko Vidak (University of Ljubljana), Rare Disease Community Contributors And Work Package Leaders (ELIXIR), Marco Roos (Leiden University) and Sergi Beltran (CRG).

Abstract:

During the last years, the ELIXIR Rare Disease (RD) community has been supporting research in the field by connecting international research institutions and projects with the ELIXIR infrastructure. To support such initiative and facilitate data interoperability and analysis, the RD community promoted the assembly of three service collections relevant for RD research which seek interactions with other ELIXIR Human Data Communities to adapt and integrate resources, proof of concept and pilot implementations of standards and methodologies.

The service collections supported by the RD community are: (1) The Assessing Molecular Pathogenicity for Rare Diseases (AMP4RD) which focuses on the integration of the available resources and tools for the assessment of molecular pathogenicity of genetic variants and aims to identify and collect the procedures and guidelines for the annotation and interpretation of genetic variants in the human genome; (2) The Systems Biology Service Bundle for Rare Diseases which defines concept maps enabling the comprehensive combination of possible tools and methods for systems biology/omics data analysis for rare diseases, to create and curate rare disease pathways and networks of disease associated genes, variants and drugs; (3) The Rare Disease Training Platform which collects the needs of the RD community, promote training events to support the data FAIRification and develop e-learning platform for sharing materials, tools and services.

Applications

VaRHC: an R package for semi-automate ACMG/AMP variant classification on hereditary cancer genes according to gene-specific ClinGen guidelines

Elisabet Munté (IDIBELL), Lidia Feliubadaló (Institut Català d'Oncologia (ICO)), Marta Pineda (Institut Català d'Oncologia (ICO)), Eva Tornero (Institut Català d'Oncologia (ICO)), Maria Isabel González-Acosta (Institut Català d'Oncologia (ICO)), Carla Roca (Institut investigació Biomèdica de Bellvitge (IDIBELL)), Jose Marcos Moreno-Cabrera (Institut investigació Biomèdica de Bellvitge (IDIBELL)), Concepción Lázaro (Institut Català d'Oncologia (ICO)) and Jose Luis Mosquera (Institut investigació Biomèdica de Bellvitge (IDIBELL)).

Abstract:

Variant classification is a major challenge. Only an accurate classification allows proper genetic counseling, personalized risk estimation and subsequent clinical management. Classifying a variant is a manual, time-consuming, iterative and tedious process that combines information of distinct nature and must follow published updated guidelines. In this context, the automation of this task can be very helpful to speed up the classification, search comprehensively through available databases and avoid manual errors.

We present vaRHC, an R package that gathers information from diverse databases, applies up-to-date ACMG/AMP guidelines considering gene specificities and generates a final classification. It automates criteria related to mutation type, population frequencies and in silico predictors, and provides information from clinical databases to help calculate the remaining criteria within 30-100 seconds. The package implements gene-specific guidelines for ATM, CDH1, CHEK2, MLH1, MSH2, MSH6, PMS2, PTEN and TP53 and the updated general ACMG/AMP rules for the remaining cancer susceptibility genes. vaRHC also provides a user-friendly report to examine and store results for non-bioinformatics users.

A validation process was conducted with 659 manually classified variants, downloaded from ClinGen repository, Vargas-Parra[1] and an in-house dataset to assess its robustness and accuracy. vaRHC has allowed us to identify 47 manual errors. The performance has also been compared with cancer SIGVAR[2] . Cohen's Kappa test revealed significant differences in favour of vaRHC.

REFERENCES.

1. Vargas-Parra, G. et al. (2020) doi:10.1002/HUMU.24110

2. Li, H. et al. (2021) doi:10.1002/humu.24177

Applications

Whole-genome sequencing analysis of food enzyme products reveals contaminations with genetically modified microorganism of related origin

Jolien D'Aes (Sciensano), Marie-Alice Fraiture (Sciensano), Bert Bogaerts (Sciensano), Sigrid C.J. De Keersmaecker (Sciensano), Nancy H.C. Roosens (Sciensano) and Kevin Vanneste (Sciensano).

Abstract:

Genetically modified microorganisms (GMM) are frequently employed for manufacturing of microbial fermentation products, such as vitamins and enzymes. Although the presence of the GMM in the final product destined for consumption is prohibited, GMM contaminations have repeatedly been reported.

The aim of this study was to perform genomic characterization and phylogenomic comparison of viable Bacillus velezensis GMM strains carrying a transgenic construct with a protease encoding gene, isolated from commercial food enzyme (FE) products from different brands.

The isolates were subjected to both short-read Illumina and long-read Oxford Nanopore Technology (ONT) whole-genome sequencing to employ a de novo hybrid assembly strategy and additional long-read bioinformatics analysis. To investigate the relationship between the GMM isolates from different samples, short-read based Single Nucleotide Polymorphism (SNP)-phylogenomic and SNP typing analyses were conducted.

We found that the GMM primarily harbour the transgenic construct on a free high-copy pUB110-derived plasmid carrying antimicrobial resistance (AMR) genes, raising serious food safety and public health concerns. The complete genetic makeup of this GMM could only be resolved by combining both short- and long reads, highlighting the added value of this approach.

SNP-phylogenetic analysis demonstrated that the isolates cluster together monophyletically, and that they differ from each other by at most 21 SNPs. Taken together, these results indicate that the GMM isolates are genetically almost identical, and probably originate from the same parental GMM strain. To our knowledge, this study is the first to demonstrate the potential of a SNP-phylogenetic approach for source-tracking of GMM.

Reference: https://doi.org/10.3390/foods10112637

Applications

WiNGS: Widely integrated NGS platform for federated genome analysis

Haleh Chizari (Katholieke Universiteit Leuven), Nasim Lalani Shabani (KU Leuven), Nishkala Sattanathan (uantwerpen), Geert Vandeweyer (uantwerpen) and Yves Moreau (KU Leuven).

Abstract:

Over 350 million people suffer from rare diseases which remain undiagnosed or misdiagnosed for years. Although most rare diseases (80%) are genetic in nature, there are various challenges associated with the diagnosis of rare diseases. In this regard, we want to tackle two complementary objectives: (1) genetically diagnose patients for which the genetic basis of the disorder is known and (2) discover the genetic basis of clinical phenotypes for which the causative basis is unknown or incomplete.

The actual disease-causing/contributing genetic variants need to be identified among numerous candidate variants. Aso, careful and extensive phenotyping and genotype-phenotype correlation are needed. Eventually, collaboration among physicians, geneticists, and bioinformaticians plays an important role. Finding similar cases require access to large amounts of information, the confidentiality of the data and the protection of the privacy of the patients becomes an important concern. Therefore, federated analysis is a key strategy to manage the risks inherent to sharing individual genomics data.

We have developed the Widely Integrated NGS (WiNGS) platform which is a fast, fully interactive, and open-source web-based platform to analyze DNA sequencing data for research and diagnosis. Its architecture is in a federated setup and a privacy-controlled manner among specific centers. It offers a rich user interface that allows annotation, query, filtering, and prioritization of the variants. Our tool explores variants among different centers to provide statistical results based on the associated phenotype. Moreover, variant discovery is a solid feature of WiNGS, set in different layers of non-sensitive data for those who are not eligible to access the complete data and disclosure of more details for those with full access rights.

Training

Bioinformatics core facility management training - availability and challenges

Eva Alloza (The Spanish National Bioinformatics Institute (INB) / Barcelona Supercomputing Center (BSC)), Ezgi Karaca (Izmir Biomedicine and Genome Center, Dokuz Eylul University), Salvador Capella-Gutiérrez (The Spanish National Bioinformatics Institute (INB) / Barcelona Supercomputing Center (BSC)), Cath Brooksbank (EMBL-EBI) and Patricia Carvajal-López (EMBL-EBI).

Abstract:

Bioinformatics core facilities play an essential role in enabling research in life sciences. As deep learning and high-throughput sequencing methodologies are increasingly applied to analyse molecular data, there is a growing need for highly specialised services from bioinformatics core facilities and their supporting teams. Added to these challenges, the rapidly changing necessities of biological data management and analysis outgrow the capabilities of these facilities in terms of software or hardware usage and human resources.

Bioinformatics core facility's scientists face many career recognition and progression challenges, often hampered by the lack of formal education in their transition from a research-focused to a service-focused and management role.

Several efforts have emerged to strengthen management-related competencies for bioinformatics core facilities. Since 2017 EMBL-EBI has run a yearly Bioinformatics Core Facility Management course aimed at experienced core facility managers. Biannually since 2019, an EMBO Research to Service practical course focuses on aspiring managers and those transitioning from a research-focused role. These events have reached over 100 participants with more than 200 applicants. Training contents evolve according to emerging needs like core facility strengths and limitations, financial sustainability, project management, human/computing resources estimation, sensitive data management, and engagement with their users and other core facilities. The participants enthusiastically apply the gained knowledge with their teams and beyond, some becoming core management trainers themselves.

Diverse efforts are needed to support the bioinformatics core facilities community of practice. The training targeted toward this community is essential to continue enabling research and development within the life sciences.

Training

Community building and training in the EMBL Bio-IT project

Renato Alves (European Molecular Biology Laboratory) and Lisanna Paladin (European Molecular Biology Laboratory).

Abstract:

The Bio-IT project (bio-it.embl.de) is a community initiative to support computational activities at the European Molecular Biology Laboratory (EMBL). Born as a way to coordinate voluntary efforts such as the organisation of internal training, it is currently active in four areas: delivering training in computational research skills; creating consulting opportunities and connections between community members; developing and maintaining resources and supporting infrastructure; and disseminating relevant information throughout the community.

Bio-IT acts in synergy with the EMBL Centres (cross-departmental structures that support researchers in specific areas of expertise) and with larger initiatives EMBL, such as the research programme trask forces and the open science actors. In light of its involvement in such large-scale initiatives, and in the context of the COVID-19 pandemic - that increased the demand of computational training and support to facilitate remote work and collaboration - Bio-IT restructured some of the community building and training activities, aiming at providing a unique platform to integrate many aspects of the research community interactions.

Bio-IT is essentially aiming at connecting EMBL's computational biology expertise (past and present), by making the information about skills and knowledge at the Laboratory as available as the information about the technical infrastructure. We aim at sharing our reflections on this process, including the lessons learned and the opportunities for growth, as well as the challenges we faced and the open problems.

Training

ELIXIR-hCNV: Galaxy workflows and training

Khaled Jumah (University of Bradford), Katarzyna Kamienicka (University of Bradford), Wolfgang Maier (University of Freiburg), David Salgado (Aix-Marseille University - INSERM), Christophe Béroud (Aix-Marseille University - INSERM), Michael Baudis (University of Zurich & SIB), Steven Laurie (CNAG), Tim Beck (University of Leicester), Salvador Capella-Gutierrez (BSC), Björn Grüning (university of Freiburg) and Krzysztof Poterlowicz (University of Bradford).

Abstract:

Human Copy Number Variations (hCNVs) represent the outcome of structural genomic rearrangements that result in the duplication or deletion of DNA segments, including homozygous deletions or amplifications leading to tens or even hundreds of copies of a sizable genomic region. These imbalanced alleles significantly contribute to human genetic variability, genetic diseases and somatic genome variations in Cancer and other conditions. hCNVs can be routinely investigated by genomic hybridisation and sequencing technologies, and a range of software tools can be used to identify and quantify hCNVs. However, so far hCNVs lack standardised formats for data representation and exchange. Moreover, sensitivity, specificity, reproducibility and reusability of research software for CNV detection and analysis are variable. Therefore, the area of genomic copy number analysis shows a need for the adoption of community-developed standards for data discovery and exchange, e.g. ELIXIR Beacon protocol, GA4GH standards; as well as mechanisms for annotating, benchmarking and making reproducible and sharable tools and workflows, e.g. WorkflowHub, Galaxy, OpenEBench; and, especially, accessible training resources and infrastructure.

To address the above, the ELIXIR human CNV Community (hCNV) developed hCNV-X and hCNV-bundles implementation Studies (2021-2023). Here, we would like to present our progress in integrating some structural genomic variant calling tools in Galaxy through the biocontainers registry and through the development of a tutorial on CNV data analysis.

Training

ELIXIR-UK DaSH: A Fellowship of data stewards

Xenia Perez Sitja (University of Bradford), Robert Andrews (Cardiff University), Sara Khalil (University of Bradford), Branka Franicevic (University of Bradford), Catherine Knox (Earlham Institute / ELIXIR-UK), Munazah Andrabi (University of Manchester), Shoaib Sufi (University of Manchester), Neil Hall (Earlham Institute), Susanna-Assunta Sansone (University of Oxford), Carole Goble (University of Manchester) and Krzysztof Poterlowicz (University of Bradford).

Abstract:

In 2021, ELIXIR-UK was granted a UKRI award for the ELIXIR-UK DaSH Project. The University of Bradford leads it, along with a delivery team at Cardiff University, Earlham Institute, University of Manchester and University of Oxford.

The project aims to produce and deliver training in FAIR data stewardship and Research Data Management (RDM), using ELIXIR-UK knowledge, experts and resources, such as the RDMkit and the FAIR Cookbook.

It focuses on building capacity and professionalising the role of data stewards at a local level with a Fellowship. Using the local expertise from diverse UK universities and research institutes, the Fellowship aims to generate a culture change from within the institutions.

The project is looking to recruit a total of 20 Data Steward Fellows, who will act as RDM and FAIR ambassadors to work on six main goals with open and reusable outputs:

1. Creating short training videos – RDMbites

2. Developing training courses

3. Delivering training locally

4. Seeding a local community of practice

5. Enhancing training skills and competencies

6. Contributing to international projects

The leading and delivery teams of the grant will simultaneously focus on building a community of practice, overseeing and coordinating the training curriculum, and ensuring the sustainability of the Fellowship beyond the project.

Training

GOBLET: Unite, inspire and equip Bioinformatics Trainers worldwide

Javier De Las Rivas (Global Organization for Bioinformatics Learning, Education & Training (GOBLET) & SoIBio-RIABIO), Eija Korpelainen (Global Organization for Bioinformatics Learning, Education & Training (GOBLET)), Annette McGrath (Global Organization for Bioinformatics Learning, Education & Training (GOBLET)), Asif M. Khan (Global Organization for Bioinformatics Learning, Education & Training (GOBLET)) and Celia W.G. van Gelder (Global Organization for Bioinformatics Learning, Education & Training (GOBLET)).

Abstract:

The Global Organization for Bioinformatics Learning, Education & Training (https://www.mygoblet.org/) is a not-for-profit Foundation, established in 2012 to harmonize bioinformatics training activities worldwide. Its members (ranging from national and international societies, networks and research institutes, to individual researchers, instructors, and students) work together to cultivate the global bioinformatics trainer community, set standards, share best practices, and provide high-quality resources to support learning, education and training. GOBLET's vision is to unite, inspire and equip bioinformatics trainers worldwide. GOBLET activities are undertaken by GOBLET members together with other members of the global bioinformatics trainer community. Examples of these activities are: (i) The publication of Professional Guides, Practical Guides and Critical Guides: short, comprehensive guides, freely available for all (https://www.mygoblet.org/publications/); (ii) Co-organising the global Bioinformatics Education Summit (4th edition was held in 2022), that is a working meeting over several time zones, where the trainer community works together in progressing the field of bioinformatics training and education (https://sites.google.com/view/educationsummit2022/home); (iii) Organizing a yearly Annual General Meeting (AGM), always combined with a Symposium and/or flanking workshops relevant for bioinformatics trainers and training organizers; and (iv) Maintaining a community-curated compendium of Resources and Documentation relevant for bioinformatics trainers and training organizers that can be accessed on the GOBLET website (https://www.mygoblet.org/training-portal/trainer-resources/).

Training

Investigating Student Sense of Belonging in Biology and Computer Science

Shamima Runa (University College Dublin), Brett Becker (University College Dublin) and Catherine Mooney (University College Dublin).

Abstract:

Students' academic sense of belonging (SoB) is important and has been associated with motivation and persistence. However, SoB varies according to factors such as race/ethnicity and gender. We examined the SoB of undergraduate biology (N=100) and computing students (N=71). In previous work, we found statistically significant differences in SoB of computing students identifying as women and as part of a minority [1]. During COVID-19 we observed a reduction in SoB of students identifying as men, and those not identifying as being part of a minority [2].

We extended our prior work to include biology students and found a statistically significant lower SoB in students identifying as a minority in biology versus those who do not ($p<0.01$). Our results showed that 23% of women in computing and 21% of women in biology identified as a minority, but for different reasons (biology;computing): gender (29%;61%), sexual identity (29%;17%), race/nationality (33%;44%), disability (14%;6%), and socioeconomic status (10%;11%).

These results provide insight that may help improve the SoB of our undergraduate students and ensure that we create inclusive learning environments for all students.

1. Mooney and Becker, 2020. Sense of Belonging: The Intersectionality of Self-Identified Minority Status and Gender in Undergraduate Computer Science Students. https://doi.org/10.1145/3416465.3416476

2. Mooney and Becker, 2021. Investigating the Impact of the COVID-19 Pandemic on Computing Students' Sense of Belonging. https://doi.org/10.1145/3463408

Training

The Bioinfo4Women International Pilot Mentoring Programme for Young Scientists

Alba Jene-Sanz (Barcelona Supercomputing Center), Olfat Khannous-Lleiffe (Barcelona Supercomputing Center), Othmane Hayoun-Mya (Barcelona Supercomputing Center), Maria Sopena-Rios (Barcelona Supercomputing Center), Mireia Codina-Tobías (Barcelona Supercomputing Center), Eva Alloza (Barcelona Supercomputing Center, Spanish National Bioinformatics Institute (INB/ELIXIR-ES)), Àtia Cortés (Barcelona Supercomputing Center) and Maria José Rementeria (Barcelona Supercomputing Center).

Abstract:

The Bioinfo4Women programme (B4W) is an initiative that started in 2018 to promote the research done by women in computational biology, and it supports researchers by promoting the exchange of knowledge and experience of outstanding women researchers through activities such as seminars, conferences, training and mentorships. B4W has particular focus on the areas of personalised medicine, bioinformatics and HPC, and ultimately aims at building a more collaborative, supportive, and equal scientific community.

The B4W International Mentoring Programme for Young Scientists kicked off as a pilot in June 2022 for one year and connects international, accomplished scientists as Mentors with junior researcher Mentees, providing role models for young researchers to help them build their scientific career. It has been designed with a gender equality perspective and with professional support, including the development of the Guidelines of the programme. Criteria for Mentors and Mentees selection were defined and training sessions took place in May and June 2022, with feedback gathered. Such training has been very well received and supports both groups in the planning and implementation of the mentoring sessions, which will be planned at the Mentee's chosen pace.

Here, we report on the design and implementation of the pilot mentoring programme and the feedbacks received from the training sessions and initial Mentor-Mentee meetings. The aim is to complete the pilot successfully by May 2023, and set the ground for more ambitious implementations, and expand towards providing guidance to postdoctoral researchers in their transition to independent researcher roles.

Training

The EOSC-Life Training programme as a tool for bringing together a community of practice

Daniel Thomas-Lopez (EMBL-EBI), Rebecca Ludwig (EATRIS) and Vera Matser (EMBL-EBI).

Abstract:

Communities of Practice (CoPs) are networks of professionals in a field that share information and best practices, contributing to the general development of the domain.

Within EOSC-Life, a project that aims to set up an open collaborative digital space for European life sciences, the training work package has been key to establishing a CoP that continuously exchanges knowledge about the organisation of training and events in various formats such as courses, hackathons, and self-learning tutorials.

A major community-building action of the network has been to create a long-lasting remote training series that periodically shares experience about training users remotely and transitioning face-to-face events to virtual (and hybrid) format. Additionally, the EOSC-Life CoP engages with project partners and external institutions who organise training funded through the project Open Calls, and EOSC-Life also has collaborations with other initiatives, including LifeScience-RI and EOSC Future, to deliver training on skills such as engagement, moderation and facilitation. These activities contribute to the expansion and strengthening of the CoP.

Finally, many members of the training community participate as translators and take part in world cafés sessions, two project initiatives that aim to bring together the technical experts with those working on training, dissemination and sustainability to build a joint understanding of EOSC-Life and its activities.

Overall, developing a versatile training plan that includes diverse learning interventions and has been flexible to adapt to the COVID-19 pandemic, has proven essential to deliver EOSC-Life products as well as to consolidate the communities and reach to others.

Training

The new Community of Practice for Data Management/Data Stewardship Training

Helena Schnitzer (FZ Jülich / ELIXIR-DE), Nils-Christian Lübke (FZ Jülich / ELIXIR-DE / de.NBI), Irena Maus (FZ Jülich / ELIXIR-DE), Tanja Dammann-Kalinowski (FZ Jülich / de.NBI), Robert Andrews (Cardiff University / ELIXIR-UK), Alexia Cardona (University of Cambridge / ELIXIR-UK), Fatima Nazeefa (Nordic Computational Biology / ELIXIR-NO), Celia van Gelder (DTL / ELIXIR-NL), Mijke Jetten (DTL / ELIXIR-NL), Brane L. Leskosek (University of Ljubljana / ELIXIR-SI), Jessica Lindvall (SciLifeL

Abstract:

Research Data Management – including all aspects of FAIR and Open Science – is an essential part of good scientific practice. To ensure that all relevant components of data management are considered from the very beginning, more and more research organisations and institutes require their researchers to develop data management plans. Especially the adherence to FAIR data principles and promotion of continuity of research projects due to data reproducibility, sharing, and reuse represents crucial benefits and motivations for good data management.

The scientific community's rising demand to provide good research data management induced the idea to establish a Community of Practice for Data Management/Data Stewardship (DM/DS) Trainers. This Community of Practice (CoP) is related to life science data and is intended for all life scientists who are interested and/or involved in DM/DS training. Principle aims of this community are to create more awareness for the necessity of data management, specify common standards in data management training, and to set a universal base level of knowledge with regards to best practices through training in data management.

In the framework of the European project ELIXIR-CONVERGE, a first CoP kick-off took place in March 2022, with the aim to discuss and shape the structure of this community.

Future regular CoP meetings and additional community events (e.g. hackathons, webinars) will serve to share information, exchange knowledge and expertise, and encourage active discussions.

Training

The new de.NBI / ELIXIR-DE training platform in 2022

Daniel Wibberg (FZ Jülich - ELIXIR-DE), Nils Christian Lübke (FZ Jülich - ELIXIR-DE), Andreas Tauch (FZ Jülich - ELIXIR-DE) and Helena Schnitzer (FZ Jülich - ELIXIR-DE).

Abstract:

de.NBI / ELIXIR-DE has established an extensive and diverse training program for the services offered over the past few years. For instance, between 2015 and 2022, de.NBI and ELIXIR-DE organized more than 400 training courses with more than 8000 participants. Since the beginning of 2022, Forschungszentrum Jülich GmbH as a member of the Helmholtz Association has been entrusted with the consolidation of ELIXIR Germany that goes hand in hand with the reorganisation of the whole network.

Due to the reorganisation of de.NBI and ELIXIR Germany, its training platform is also in a restructuring phase. The training platform will be more focusing on strategic aspects in the training field, training collaborations in Germany (e.g. with NFDI) and Europe, developing new data management training courses and training material in the context of ELIXIR-CONVERGE and on trainer community building.

Based on the restructuring, the de.NBI / ELIXIR-DE training platform will be better aligned to the challenges of bioinformatics training in Germany and Europe of the next few years.

Technical Secretariat

Pl. Europa, 17-19 1st floor
08908 L'Hospitalet de Llobregat Barcelona, Spain
Ph: +34 93 882 38 78
eccb2022@bcocongresos.com