Bildquelle/Fotograf: Tobias Jakobi

Precision Medicine –
where Bioinformatics & Medical Informatics meet.

GERMAN CONFERENCE ON BIOINFORMATICS

16 – 19 SEPTEMBER | HEIDELBERG

German Cancer Research Center

Organized by

DECHEMA
Gesellschaft für Chemische Technik und Biotechnologie e.V.

Supported by

BioQuant
MODEL base of LIFE

dkfz.
GERMAN CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

EMBL

GBM

HEIDELBERG FACULTY OF MEDICINE

Heidelberg Institute for Theoretical Studies
HITS

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

gcb2019.de

Book of Abstracts

BOOK OF ABSTRACTS

page

Lectures

From Omics Data to Therapies - Personalizing Treatments

Oliver Kohlbacher

University of Tübingen, University Hospital Tübingen, and

Max Planck Institute for Developmental Biology, Tübingen, Germany

The massive decrease in cost due to technological advances has made large-scale high-throughput data (genomics, transcriptomics, proteomics, metabolomics, etc.) available for large patient cohorts and has enabled comprehensive analyses of these big biological data sets. In particular for cancer, these analyses enable new insights into pathogenesis, patient stratification, and – ultimately – new therapeutic options. A particular challenge lies in leveraging these insights for the treatment of each individual patient.

We present approaches for stratifying cancer therapies based on integrative analysis of omics data and describe how the analysis and interpretation of this data can be brought into a clinical setting. Molecular tumor boards (MTBs) provide an organizational framework to discuss integrated data and formulate therapeutic options. We present data analysis algorithms providing annotation and analysis of genomics and multi-omics data and discuss how these data can be brought to the clinician in a seamless manner – as well as the hurdles and pitfalls on that way.

The RNA workbench 2.0: next generation RNA data analysis

Jörg Fallmann, Bioinformatics Group, Department of Computer Science; Leipzig University, Härtelstraße 16-18, D-04107 Leipzig, Germany; Pavankumar Videm, Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg, Georges-Köhler-Allee 106, Freiburg 79110, Germany; Andrea Bagnacani, Department of Systems Biology and Bioinformatics, Institute of Computer Science, University of Rostock, Ulmenstr. 69, 18057 Rostock, Germany; Bérénice Batut, Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg, Georges-Köhler-Allee 106, Freiburg 79110, Germany; Maria A. Doyle,Research Computing Facility, Peter MacCallum Cancer Centre, Melbourne, Victoria 3000, Australia; 5 Sir Peter MacCallum Department of Oncology, The University of Melbourne,Victoria 3010, Australia; Tomas Klingstrom, SLU-Global Bioinformatics Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences Sweden; Florian Eggenhofer, Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg, Georges-Köhler-Allee 106, Freiburg 79110, Germany; Peter F. Stadler, Bioinformatics Group, Department of Computer Science; Interdisciplinary Center of Bioinformatics; German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig; Competence Center for Scalable Data Services and Solutions; and Leipzig Research Center for Civilization Diseases, Leipzig University, Härtelstraße 16-18, D-04107 Leipzig; 8 Max-Planck-Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig; Inst. f. Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria; 9 Facultad de Ciencias, Universidad Nacional de Colombia, Sede Bogotá, Colombia Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA; Rolf Backofen, Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg, Georges-Köhler-Allee 106, Freiburg 79110, Germany; Signalling Research Centres BIOSS and CIBSS, Albert-Ludwigs-University Freiburg, Schänzlestr. 18, Freiburg, Germany; Björn Grüning, Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg, Georges-Köhler-Allee 106, Freiburg 79110, Germany; Center for Biological Systems Analysis (ZBSA), University of Freiburg, Habsburgerstr. 49, 79104 Freiburg, Germany

Project: http://rna.usegalaxy.eu

Code: https://github.com/bgruening/galaxy-rna-workbench

License: MIT

RNA centric research is of growing importance for medicine and molecular biology. However, increasing amounts of data from diverse types of deep sequencing experiments create a demand for automatic analysis and interpretation solutions. To address these needs, we present the RNA Workbench 2.0, an updated online framework for RNA related data analysis.

The RNA workbench offers a wide range of tools covering classic RNA-bioinformatics as well as RNA related high-throughput data analysis. There are over 100 tools included and over 15 tutorials from the categories RNA structure prediction, RNA alignment, RNA annotation, RNA-protein interaction, RNA target prediction, RNA-seq analysis, and Ribosome profiling. RNA specific visualisation solutions for e.g. dot-bracket plots and secondary structures are an integral part of the workbench. In contrast to pre-existing solutions, our community-driven approach allows us to include classic RNA-bioinformatics tools, often with the direct support of the tool-authors. These contributions enable us to provide excellent documentation, training material and interactive tours demonstrating the functionality of the workbench.

Built on the Galaxy framework, the workbench offers sophisticated analyses to users without command line knowledge, while emphasising reproducibility, customization and effortless scale up to larger infrastructures. In addition to providing the RNA Workbench 2.0 as a portable Docker container, users can now directly use integrated tools, workflows and tutorials at the de.NBI-cloud powered European Galaxy instance - http://rna.usegalaxy.eu. This eliminates the overhead of setting up hardware or software environments and makes RNA Workbench 2.0 even easier to use, train experimentalists and bioinformaticians and to run full-fledged data analysis workflows. The tight integration into the Galaxy Training Network and the bidirectional exchange of workflows and tutorials ensure that the RNA Workbench 2.0 remains a state-of-the-art training and RNA research resource.

Kubernetes is a versatile tool to provide WebApps for data exploration on individual omics projects

Hendrik Schultheis, Bad Nauheim, Max Planck Institute for Heart and Lung Research, Germany; Carsten Kuenne; Bad Nauheim, Max Planck Institute for Heart and Lung Research, Germany; Jens Preussner, Bad Nauheim, Max Planck Institute for Heart and Lung Research, Germany; Julian Winter, Bad Nauheim, Max Planck Institute for Heart and Lung Research, Germany; Philipp Goymann, Bad Nauheim, Max Planck Institute for Heart and Lung Research, Germany; Marina Kiweler, Bad Nauheim, Max Planck Institute for Heart and Lung Research, Germany; Alexander Goessmann, Gießen, Bioinformatics and Systems Biology at the JLU, Germany*; Mario Looso; Bad Nauheim, Max Planck Institute for Heart and Lung Research, Germany*

Background

High-throughput (HT) screens of complex biological systems such as the proteome, transcriptome or genome generate extensive amounts of data. These services are typically supplied by technical facilities of research institutions in order to bundle the technical expertise and to provide access and assistance to end users. Primary results are often produced by pipelines, providing relevant features in extensive spreadsheet tables. Since interactive exploration and visualization of HT data is a key aspect for the analysis and understanding of the biological systems under investigation, we recently developed WIlsON, an interactive workbench for analysis and visualization of multi-omics data. It is primarily intended to empower screening platforms to offer access to pre-calculated HT screen results to the non-computational scientist. Facilitated by an open file format, WIlsON supports all types of omics screens, serves results via a web-based dashboard, and enables end users to perform analyses and generate publication-ready plots. However, for most bioinformatics service facilities, infrastructure aspects such as the provision of web services, or storage and computational capacities represent a major strain, binding valuable personnel. Utilization of cloud resources such as de.NBI is a versatile alternative solution for on premise IT infrastructures, allowing to reduce costs and to focus on bioinformatics aspects.

Methods and results

Here we utilize the Kubernetes open-source software provided at the de.NBI hub at the JLU Gießen. It hosts containerized services such as WIlsON in a horizontally scaling infrastructure. A tool named KASSIo (***K****ubernetes* ***Assi****stant*) was developed in-house to deploy and manage WebApp containers for individual projects automatically, triggered by analysis pipelines after completion An advanced workflow includes I) reduction of all configuration data for a whole omics project to a single JSON formatted file allowing deployment and updating of individual projects, II) utilization of an on premise S3 storage for sensitive data as permanent storage system, and III) a centralized and secure handling of user authentication and data access via randomized URLs.

Discussion

Utilizing cloud computing resources for temporary computational loads demanded by data analysis pipelines or more permanent services such as WebApps is a

worthwhile alternative to on premise IT infrastructures. However, making use of the resources provided by the de.NBI generate an urgent need to acquire the respective knowhow as most university training courses for bioinformaticians do not address this field yet.

The ELIXIR Galaxy Community update

Bérénice Batut, Bioinformatics Lab, Albert-Ludwigs-Universität, Freiburg, Germany;
Helena Rasche, Bioinformatics Lab, Albert-Ludwigs-Universität, Freiburg, Germany;
Björn Grüning, Bioinformatics Lab, Albert-Ludwigs-Universität, Freiburg, Germany;

Galaxy is an open, web-based platform for computational biomedical research. It allows researchers without programming experience to analyze their data using workflows, share their analyses, and enable others to reproduce the same analysis. Galaxy makes science reproducible and data traceable, it facilitates data and result sharing, and removes the need for users to compile and install software tools.

An increasing number of sub-communities have grown up to address specific tasks in Galaxy. To foster interaction between these data-specific communities, the ELIXIR Galaxy Community was created in 2018. In this talk we will provide an update to the ELIXIR Galaxy community.

One of the goals of the ELIXIR Galaxy community is to set up common analysis workflows and standards, and to provide complementary training, extending the Galaxy training portfolio. This training material repository contains over 120 tutorials and provides material for using and developing Galaxy. The growing amount of data generated in life science, and the growing number of users, requires increasing amounts of computing power. In 2018, the community launched usegalaxy.eu, the European Galaxy server, using the de.NBI Cloud as infrastructure. The aim now is to expand this effort into a network of Galaxy instances worldwide, guaranteeing a base level of compatibility. We also want to facilitate the usage of Galaxy across the different ELIXIR clouds for users that want to maintain their own instance, but also allow users to easily combine public and private storage as well as compute cloud services to improve the accessibility of tools and data.

Retrieving data from public databases into a Galaxy instance is the first step for most analyses, but identifying files and their URLs in order to upload these files into a computational environment is not easy for all users. The community aims to facilitate data integration into Galaxy instances from the ELIXIR Core Data Resources (ENA, ArrayExpress, etc), and also more specialised databases (Brenda, Silva, etc). We are working to create and maintain a shared storage of common reference data for genomes across Galaxy instances, based on CVMFS. This facilitates adding new genomes to any Galaxy instance, including indices and annotations.

Over the last years, the ELIXIR Galaxy community fostered exchanges between several European initiatives around Galaxy, by building a comprehensive community of the bioinformaticians, developers and administrators, and collaborate with trainers and other ELIXIR Communities and platforms.

Topological analyses of causal signaling models contextualized by gene expression data

Panuwat Trairatphisan, Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Bioquant, Heidelberg, Germany; presenting author, Enio Gjerga, Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Bioquant, Heidelberg, Germany; Anika Liu, Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Bioquant, Heidelberg, Germany; Julio Saez-Rodriguez, [1] Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Bioquant, Heidelberg, Germany, [2] Joint Research Centre for Computational Biomedicine (JRC-COMBINE), RWTH-Aachen University, Aachen, Germany.

Cellular signaling networks are often deregulated in many pathological conditions and the inference of aberrant regulatory signaling networks can generally be performed based on phosphoproteomics data. However, the availability of this data type is scarce in certain research areas such as toxicology while the measurement at the transcriptomics level downstream of signaling network is more abundant.

We introduce CARNIVAL, an R-package for causal reasoning-based network contextualisation which allows identifying a subset of regulatory signaling network from gene expression data. Transcription factors' and signaling pathways' activities can be inferred from gene expressions with our in-house tools DoRothEA and PROGENy, respectively, and they can then be integrated into network contextualization pipeline. This input information together with a signed directed protein-protein interaction network with or without potential protein drug targets is then used to derive a series of linear constraints to generate integer linear programming (ILP) problems. An ILP solver is subsequently applied to identify the sub-network topology with minimized discrepancies on fitting error and model size.

We show the value of CARNIVAL using three different case studies: deregulation of signaling in cancer cell lines, alterations in immune kidney disease, and drug-induced liver injury.

RNA modifications regulate gene expression programs

Michaela Frye

Deutsches Krebsforschungszentrum, Heidelberg/Germany

Many of the hundreds of known chemical modifications in RNA were discovered over forty years ago but then forgotten because suitable, sensitive tools to detect the modifications at high resolution were lacking. Through the development of novel biochemical, functional and genomics tools we are only now beginning to understand the whole breadth and extensive functional roles of RNA modifications in higher organisms. I will present some mechanistic examples how RNA modifications help to shape normal tissue homeostasis, and how aberrant formation of RNA modifications contributes to disease. By focusing on the roles of transfer RNA (tRNA) modification in regulating gene expression, I will discuss novel and emerging molecular functions of RNA modifications in abundant non-coding RNAs.

Together, our work demonstrates that by understanding the role of RNA modifications in physiology and pathology, novel and powerful therapeutic drug targets for human diseases and can potentially be identified and further optimized for clinical studies.

Systems biology-based investigation of cooperating miRNAs as monotherapy or adjuvant therapy in cancer

Xin Lai, Laboratory of Systems Tumor Immunology, Friedrich-Alexander-University of Erlangen-Nürnberg (FAU) and Universitätsklinikum Erlangen, Erlangen, Germany; Martin Eberhardt, Laboratory of Systems Tumor Immunology, Friedrich-Alexander-University of Erlangen-Nürnberg (FAU) and Universitätsklinikum Erlangen, Erlangen, Germany; Shailendra K Gupta, Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany; Ulf Schmitz, Gene & Stem Cell Therapy Program, Centenary Institute, University of Sydney, Camperdown, Australia; Stephan Marquardt, Institute of Experimental Gene Therapy and Cancer Research, Rostock University Medical Centre, Rostock, Germany; Susanne Knoll, Institute of Experimental Gene Therapy and Cancer Research, Rostock University Medical Centre, Rostock, Germany; Alf Spitschak, Institute of Experimental Gene Therapy and Cancer Research, Rostock University Medical Centre, Rostock, Germany; Olaf Wolkenhauer, Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany; Brigitte M Pützer, Institute of Experimental Gene Therapy and Cancer Research, Rostock University Medical Centre, Rostock; Julio Vera, Laboratory of Systems Tumor Immunology, Department of Dermatology, Friedrich-Alexander University (FAU) Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany

Abstract

High rates of lethal outcome in tumour metastasis are associated with the acquisition of invasiveness and chemoresistance. Several clinical studies indicate that E2F1 overexpression across high-grade tumours culminates in unfavourable prognosis and chemoresistance in patients. Thus, fine-tuning the expression of E2F1 could be a promising approach for treating patients showing chemoresistance.

We integrated bioinformatics, structural and kinetic modelling, and experiments to study cooperative regulation of E2F1 by microRNA (miRNA) pairs in the context of anticancer chemotherapy resistance. We showed that an enhanced E2F1 repression efficiency can be achieved in chemoresistant tumour cells through two cooperating miRNAs. Sequence and structural information were used to identify potential miRNA pairs that can form

tertiary structures with E2F1 mRNA. We then employed molecular dynamics simulations to show that among the identified triplexes, miR-205-5p and miR-342-3p can form the most stable triplex with E2F1 mRNA. A mathematical model simulating the E2F1 regulation by the cooperative miRNAs predicted enhanced E2F1 repression, a feature that was verified by in vitro experiments. Finally, we integrated this cooperative miRNA regulation into a more comprehensive network to account for E2F1-related chemoresistance in tumour cells. The network model simulations and experimental data indicate the ability of enhanced expression of both miR-205-5p and miR-342-3p to decrease tumour chemoresistance by cooperatively repressing E2F1. Our results suggest that pairs of cooperating miRNAs could be used as potential RNA therapeutics to reduce E2F1-related chemoresistance.

Key Structural Patterns for miRNA Family Reconstruction

Cristian A. Velandia-Huerto *, Universität Leipzig, Deutschland; Ali M. Yazbeck, Universität Leipzig, Deutschland; Peter F. Stadler, Universität Leipzig, Deutschland;
* cristian@bioinf.uni-leipzig.de

Background

Efforts to classify and annotate microRNAs have focused on conservation patterns of primary sequence motifs, secondary structure of the hairpin precursor, and efficient target recognition. This research interest relies on the accumulated evidence of high structural conservation levels as a requisite for correct miRNA processing. In this way, the established designation of miRNA families is mainly based on the miRBase definition, which stipulates that the family name is assigned based on the degree of conservation between the *mature* sequences, but those relations have been pragmatically chosen. An additional source is the RFAM database, which reported families based on a manually curated set of seed alignments and consequently their Covariance Models (CM) construction.

Currently, both databases are wide-used as a reference of miRNA families and/or as a resource to annotate new candidates. But in terms of family reconstruction of the miRNA genes, accurate identification of miRNAs is imperative, so additional biological criteria have to be considered to distinguish bona fide miRNAs from other small RNAs such as fragments of snoRNAs, snRNAs, tRNAs, or even coding transcripts.

Methods

To annotate new miRNA candidates based on homology searches and their classification into a corresponding family, 21 chordate genomes were subject of annotations with blast and Hidden Markov Models (HMMs). The structural folding of those candidates was assessed through a metazoan-specific, chordate specific and miRBase new CMs. Corrected mature positions from all miRBase sequences were calculated with MIRfix, these results allowed a re-definition and classification of miRNA *loci* into new defined families, that are supported by mature-fixed multiple alignments of precursors.

Results

New miRNA have been annotated on the chordate genomes, including the genomic coordinates from precursors and their correspondent mature sequences. In overall, we find that Tunicata contain a reduced repertoire of conserved miRNA. This is a odd with the common perception that miRNA families are rarely lost once they are established in the genome.

Improving accuracy of phenotype detection in whole-genome CRISPR screens

Katharina Imkeller, EMBL and DKFZ, Heidelberg, Germany; Michael Boutros, DKFZ, Heidelberg, Germany; Wolfgang Huber, EMBL, Heidelberg, Germany

High-throughput genetic screens, such as CRISPR-Cas9 knockout screens, are used to discover and characterize gene functions and genotype-phenotype relationships at a genome-wide scale. The experimental procedure is based on a library of target gene specific perturbagens (sgRNAs) that is applied to a pool of cells, which then proliferate in parallel and possibly in competition. The changes of perturbagen abundance during selection (expressed, e.g., in fold changes) are the basis for subsequent statistical data analysis to estimate the effect of each perturbagen on cell fitness. However, we find that current statistical approaches - several of which were adapted from RNA-seq data analysis - do not fit such data sufficiently, which leads to under- or overcalling of hits. Moreover, it is unclear which experimental design parameters of the labor- and cost-intensive screening experiments have essential effects on data quality.

On a dataset of multiple genome-wide CRISPR-Cas9 screens in human cell lines, we show that the distribution of fold changes is asymmetric even for perturbagens with no effect on fitness. This asymmetry is generated during the selection phase of the screen, and its strength is influenced by experimental design parameters. We present a new statistical approach that delivers improved phenotype detection and allows reduction of experiment size by up to 1/2.

circtools – a bioinformatics toolbox for circular RNAs

Tobias Jakobi, Department of Cardiology, Angiology, and Pneumology, University Hospital Heidelberg, Heidelberg, Germany, Alexey Uvarovskii, Department of Cardiology, Angiology, and Pneumology, University Hospital Heidelberg, Heidelberg, Germany, Christoph Dieterich, Department of Cardiology, Angiology, and Pneumology, University Hospital Heidelberg, Heidelberg, Germany

Background

Circular RNAs (circRNAs) originate through back-splicing events from linear primary transcripts, are resistant to exonucleases, not polyadenylated, and have been shown to be highly specific for cell type and developmental stage. Although few circRNA molecules have been shown to exhibit miRNA sponge function, for the majority of circRNAs, however, their function is yet to be determined. Prediction of circRNAs is a multi-stage bioinformatics process starting with raw RNA sequencing data and yielding, depending on tissue and condition, sets of hundreds of potential circRNA candidates that require further analyses. While a number of tools for the prediction process already exist, publicly available downstream analysis tools are rare. We aim to provide researches with a harmonized work flow that covers different stages of in silico circRNA analyses, from prediction to first functional insights.

Methods and results

Here, we present circtools (Jakobi et al. 2018), a modular, Python-based framework that unifies several *in silico* circRNA analyses in a single command line-driven toolbox. Our software includes modules for circRNA detection and reconstruction that are based on our well tested DCC (Cheng et al. 2013) and FUCHS (Metge et al. 2017) tools. Circtools however, was developed as a complete analysis work flow and therefore also contains additional modules to perform initial quality checks, test circRNAs for host gene independent expression, identify differentially spliced exons, screen circRNAs for enriched features (e.g. RBP sites), and design circRNA-specific primers for qRT-PCR verification. Circtools supports researchers with visualization options and data export into commonly used formats. We intend to add more modules in the future in order to provide a comprehensive bioinformatics toolbox for the research community and encourage users to contribute new modules.

Availability

Circtools is available under GPL-V3.0 license on GitHub via http://circ.tools; an extensive documentation is located at http://docs.circ.tools.

Big Data Meets Drug Target Discovery

Robert Gentleman, 23andMe, USA

I will discuss how we use our large genotype and phenotype database to drive the discovery of novel therapeutics. I will show some examples where we find known targets, that exemplify the approach. I will discuss how different ethnicities and founder populations can provide us with novel insights. I will discuss some of the challenges in going from a genetic signal to a drug target.

Computational methods for single-cell genetics

Dr. Oliver Stegle | *European Molecular Biology Laboratory (EMBL)* | *Germany*

Abstract not available

A Soft Alignment of Multiple Omic t-SNEs

Laleh Haghverdi, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany;

John C. Marioni, European Bioinformatics Institute (EBI-EMBL), Hinxton, UK; Cancer Research UK Cambridge Institute, Cambridge, UK;

Wolfgang Huber, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.

Biological regulation of cells identity and function runs at several levels of genetic and

epigenetic processes including gene expression, chromatin openness, DNA methylation, histone modifications, etc. To obtain an understanding of the function of each molecular level in connection to other levels, researchers are interested in integration and combined analysis of such data sets, generally known as multi-omics. Especially, recent technology developments facilitate collection of multi-omics data at single cell level, hence provide the opportunity to acquire a more holistic representation of cell heterogeneities by information coding and regulation of several molecular levels in single cells. New computational methods are thus emerging to address integration and joint analysis of multiple data modalities. Here, we present a novel cells alignment strategy which allows connecting of two omic data sets together when they are measured on different cells sampled from the same biological system.

Our method uses the similarity between the shapes of data sets (i.e. structure of cell populations with respect to each other, which is quantitatively reflected in the cells pairwise distance or similarity matrix) to map them in a common abstract latent space provided by a widely used data embedding method for single cell data known as t-distributed stochastic neighbours embedding (t-SNE). Optionally, our method can use any available information of correspondence for a number of co-assayed cells or any shared features collected for both data modes to leverage the alignment quality. We demonstrate the performance of our method on a simulated data set of four distinct cell clusters as well as a simulated data of a branching differentiation trajectory. We then apply it on a multi omics data set from mammalian germ line specification showing that the jointly aligned t-SNEs of two data modes yields a more

detailed and biologically more informative embedding of cells than the separate t-SNEs from each mode.

Extracting biological connections from single cell RNA-Seq data with deep generative models using conditional sampling

Moritz Hess[1], Stefan Lenz[1], Harald Binder[1]

1 Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg/Germany;

Background

The high dimensional gene expression profiles of single cells, inferred by deep RNA sequencing (RNA-Seq), provide information to model cellular interaction networks which is important for better understanding pathogenesis of diseases.

Methods and results

We employ generative models, specifically deep Boltzmann Machines (DBM), to investigate connections between cell identity and the underlying functional characteristics in single cell RNA-Seq data. In particular, we train DBMs to learn the joint distribution of neurotransmitter receptor gene expression and marker gene expression in different neuron types in the mouse brain. By sampling from the DBMs, conditional on the expression levels of marker genes, which indicate the cell identity, we study the connection of cell identity with a differentiation of neuronal function such as the type of neurotransmitter that is employed by the cell. For example, using the expression data of few marker genes allows us to infer the expression levels of receptor genes for GABA and glutamate signaling in the corresponding cells, indicating that the DBM learned the connection of marker genes with genes related to a differentiated cellular function.

Discussion

Here we demonstrate that learning the joint distribution of gene expression data allows to identify connections between genes that are related to cell identity and genes that are related to cell function. These findings demonstrate the potential of generative models, like DBMs, to extract biologically meaningful connections from single cell RNA-Seq data, which helps to better understand cellular networks.

Transcriptional and genomic intra-tumor heterogeneity drives subclone specific drug responses in B cell lymphoma

Tobias Roider, University Hospital Heidelberg, Heidelberg, Germany; Felix Frauhammer, Center for Molecular Biology of the University of Heidelberg, Heidelberg, Germany; Julian Seufert, German Cancer Research Center (DKFZ), Heidelberg, Germany; Marie Bordas, German Cancer Research Center (DKFZ), Heidelberg, Germany; Marta Stolarczyk, University Hospital Heidelberg, Heidelberg, Germany; Philipp Malm, German Cancer Research Center (DKFZ), Heidelberg, Germany; Sophie Rabe, University Hospital Heidelberg, Heidelberg, Germany; Peter-Martin Bruch, University Hospital Heidelberg, Heidelberg, Germany; Michael Hundemer, University Hospital Heidelberg, Heidelberg, Germany; Karsten Rippe, German Cancer Research Center (DKFZ), Heidelberg, Germany; Benjamin Goeppert, University Hospital Heidelberg, Heidelberg, Germany; Martina Seiffert, German Cancer Research Center (DKFZ), Heidelberg, Germany; Benedikt Brors, German Cancer Research Center (DKFZ), Heidelberg, Germany; Thorsten Zenz, University Hospital Zurich, Zurich, Switzerland; Matthias Schlesner, German Cancer Research Center (DKFZ), Heidelberg, Germany; Carsten Müller-Tidow, University Hospital Heidelberg, Heidelberg, Germany; Stefan Fröhling, National Center for Tumor Diseases (NCT), Heidelberg, Germany; Wolfgang Huber, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany; Simon Anders, Center for Molecular Biology of the University of Heidelberg, Heidelberg, Germany; Sascha Dietrich, University Hospital Heidelberg, Heidelberg, Germany.

Background

Despite extensive research, clonal heterogeneity of tumor diseases is still poorly understood. Recently, the potential of single cell RNA sequencing (scRNA-Seq) to dissect intra-tumor heterogeneity has been demonstrated; however, a functional characterization of the subpopulations thus identified has not yet been achieved.

Methods and Results

We assayed single cell solutions of non-Hodgkin B cell lymphoma lymph node biopsies and combined immunophenotyping, scRNA-Seq, genome sequencing and high through put drug screening to characterize tumor heterogeneity on multiple

layers. Among the malignant cells we identified transcriptionally distinct subclones and isolated them by FACS using tumor-specific surface markers inferred from the transcriptomic data. We further characterized these subclones genetically by whole-genome sequencing and functionally by high-throughput perturbation assays with a comprehensive panel of drugs. We found striking intra-tumor differences in genetic make-up and especially in drug sensitivities: Several clinically relevant drugs were effective on only one subclone, with others effective only on the other. We also assessed non-malignant bystander cells, quantitatively resolving the tumor niches' varying complements of different subtypes of B and T cells.

Discussion

Our work shows how the full potential of scRNA-Seq for understanding tumor heterogeneity can be unlocked by performing transcriptome-informed sorting and follow-up study of individual subclones. Our observations on intra-tumor differences in drug vulnerabilities open a new path to understand treatment failures and relapses.

Mono- and multi-nucleated ventricular cardiomyocytes constitute a transcriptionally homogenous cell population

Yekelchyk M[(1)], Guenther S[(1)(2)], Preussner J[(1)], Braun T[(3)(4)]

[(1)]Department of Cardiac Development and Remodeling, Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany.

[(2)]German Centre for Cardiovascular Research (DZHK), Partner Site Rhein-Main, Frankfurt am Main, Germany.

[(3)]Department of Cardiac Development and Remodeling, Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany.

Thomas.Braun@mpi-bn.mpg.de.

[(4)]German Centre for Cardiovascular Research (DZHK), Partner Site Rhein-Main, Frankfurt am Main, Germany. Thomas.Braun@mpi-bn.mpg.de.

Abstract:

Individual adult ventricular cardiomyocytes are either mono-or multi-nucleated and undergo morphological changes during cardiac hypertrophy. However, corresponding transcriptional signatures, reflecting potentially different functions or the ability for cell-cycle entry, are not known. The aim of this study was to determine the transcriptional profile of mono-and multi-nucleated adult cardiomyocytes by single-cell RNA-sequencing (scRNA-seq) and to investigate heterogeneity among cardiomyocytes under baseline conditions and in pressure-induced cardiac hypertrophy. We developed an array-based approach for scRNA-seq of rod-shaped multi-nucleated cardiomyocytes from both healthy and hypertrophic hearts. Single-cell transcriptomes of mono-or multi-nucleated cardiomyocytes were highly similar, although a certain degree of variation was noted across both populations. Non-image-based quality control allowing inclusion of damaged cardiomyocytes generated artificial cell clusters demonstrating the need for strict exclusion criteria. In contrast, cardiomyocytes isolated from hypertrophic heart after transverse aortic constriction showed heterogeneous transcriptional signatures, characteristic for hypoxia-induced responses. Immunofluorescence analysis revealed an inverse correlation between HIF1α + cells and CD31-stained vessels, suggesting that imbalanced vascular growth in the hypertrophied heart induces cellular heterogeneity. Our study demonstrates that individual

mono-and multi-nucleated cardiomyocytes express nearly identical sets of genes. Homogeneity among cardiomyocytes was lost after induction of hypertrophy due to differential HIF1α-dependent responses most likely caused by none-homogenous vessel growth.

Prioritising cancer therapeutic targets via CRISPR-Cas9 screens and multi-omics data integration

Fiona M Behan[1,2,†], Francesco Iorio[1,2,†], Gabriele Picco[1,†], Euan A. Stronach[3], Julio Saez-Rodriguez[4], Kosuke Yusa[1,2,], Mathew J Garnett[1,2,*]. [1] Wellcome Sanger Institute, UK. [2]Open Targets, UK. [3]GlaxoSmithKline Research and Development, UK. [4]Heidelberg University, Germany.*

*† Equally contributing authors, * Co-senior authors. Full list of authors in [1]*

Functional genomics approaches can overcome limitations that hamper oncology drug development such as lack of robust target identification and poor clinical efficacy. Here we performed genome-scale CRISPR-Cas9 screens in 324 human cancer cell lines from 30 cancer types and developed a data-driven framework to prioritise cancer therapeutic candidates [1]. We integrated cell fitness effects with genomic biomarkers and target tractability for drug development to systematically prioritise new targets in defined tissues and genotypes. We verified one of our most promising dependencies, Werner syndrome ATP-dependent helicase, as a synthetic-lethal target in tumours from multiple cancer types with microsatellite instability.

Our analysis provides a comprehensive resource of cancer dependencies, generates a framework to prioritise oncology targets, and nominates specific new targets. The principles described in this study can inform the initial stages of drug development by contributing a new, diverse and more effective portfolio of oncology targets.

[1] Fiona Behan et al. (2019), ***Nature*** [in press, to appear online on April 10th 2019]

DOI: 10.1038/s41586-019-1

nanotyper: an HLA genotyping algorithm for nanopore sequencing data

Steffen Klasberg[1], Kathrin Putke[1], Vineeth Surendranath[1], Alexander Schmidt[1,2], Vinzenz Lange[1], Gerhard Schöfl[1]

[1]DKMS Life Science Lab, Dresden, Germany; [2] DKMS, Tübingen, Germany

Nanopore sequencing may put sequence-based HLA genotyping in reach of even the smallest immunogenetic laboratories. Without the requirement for major capital investments, it delivers very long reads, easily covering both HLA class I and class II genes. In addition, results may be obtained very rapidly and at reasonable cost. However, these advantages come at the price of a lower per-read accuracy of only ~ 85 to 90 % compared to the high accuracy of Illumina reads. Here, we present an algorithmic approach to obtain accurate genotyping results from such noisy third-generation sequencing data. First, phase-informative polymorphic positions are identified and used to cluster reads into allele-specific read-sets. Next, reads with high rates of sequencing noise are filtered from read-sets based on cluster membership scores. Instead of assigning HLA alleles using consensus sequences, the multiple sequence alignment (MSA) derived from a read-set is used for scoring the likelihood of each allele in the reference database. To cope with incomplete reference sequence information in the IPD-IMGT/HLA database, the scoring is performed hierarchically, starting with the exons of the antigen recognition domain (ARD), followed by non-ARD exons and non-coding sequences. Partially known reference alleles are score-adjusted to avoid biased results. At each step the 10% best-scoring alleles are retained, yielding final result sets of candidate alleles in conjunction with their respective likelihoods. Importantly, this approach transparently integrates and reports ambiguities that arise both due to read quality and reference issues. We have successfully tested the nanotyper pipeline on HLA class I and class II genes using ONT MinION and PromethION data.

On the semantics of unknown DNA motifs – a machine learning approach

Theodor Sperlea, Faculty of Mathematics and Computer Science, Philipps-Universität Marburg; Lea Muth, Faculty of Mathematics and Computer Science, Philipps-Universität Marburg; Faezeh Moradi, Faculty of Mathematics and Computer Science, Philipps-Universität Marburg; Roman Martin, Faculty of Mathematics and Computer Science, Philipps-Universität Marburg; Christoph Weigel, Department of Life Science Engineering, HTW Berlin; Torsten Waldminghaus, Chromosome Biology Group, SYNMIKRO, Philipps-Universität Marburg, and Dominik Heider, Faculty of Mathematics and Computer Science, Philipps-Universität Marburg

Besides genes, DNA can contain short, reoccuring and sometimes essential patterns called motifs, which serve as protein binding sites (D'haeseleer, *Nat. Biotechnology*, 2006). While identification of these motifs can be done based on their occurence statistics, these methods cannot illuminate the function of the motifs (Das & Dai, *BMC Bioinformatics*, 2007).

In this study, we present a novel machine learning approach for the identification of taxonomically relevant motifs in bacterial origin of replication (*oriC*) sequences. Bacterial *oriC* are excellent objects for studying motifs and their interplay due to the fact that this, usually intragenic, segment of DNA contains many DNA motifs that are involved in cell cycle regulation (Marczynski, Rolain & Taylor, *Front. Microbiol.*, 2015). To this end, we created the largest database of oriC sequences to date and subsequently classified these sequences to their respective orders, families, and geni, and subsequently extracted important motifs from the best-performing models. Furthermore, using an adaptation of Word2Vec (Mikolov et al., *CoRR*, 2013) for DNA sequences, we were able to derive the semantic content of these motifs from their syntatic structure and assign functions to hitherto unknown motifs.

Accurate prediction of cell type-specific transcription factor binding

Jan Grau, Stefan Posch, Institute of Computer Science, Martin Luther University Halle-Wittenberg, Germany; Jens Keilwagen, Institute for Biosafety in Plant Biotechnology, Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, Germany

Background

Transcriptional regulation of gene expression by transcription factors (TFs) is one of the fundamental principles of gene regulation. While substantial progress has been made with regard to more accurate models of transcription factor binding and integration of diverse experimental assays of, for instance, chromatin accessibility over the past years, accurate prediction of *in-vivo* binding regions of individual TFs remains a challenging task. The ENCODE-DREAM challenge had been created to unbiasedly and comprehensively assess the current state-of-the-art in predicting *in-vivo* TF binding regions in the human system, considering different cell types for training models and for benchmarking predictions.

Methods and Results

We present our supervised approach for predicting cell type-specific *in-vivo* TF binding, which gained a shared first rank in the ENCODE-DREAM challenge. Methodology that distinguishes our approach from those previously described in the literature are i) a novel iterative training procedure successively complementing an initial set of negative regions by further "complicated" examples and ii) an ensemble approach aggregating predictions based on classifiers from individual iterations and different cell types (Keilwagen et al., 2019, Genome Biology, doi:10.1186/s13059-018-1614-y). In post-challenge studies, we discovered that chromatin accessibility and motif-based features are the most important determinants of prediction performance. This lead to a streamlined open source implementation of our approach called Catchitt (http://jstacs.de/index.php/Catchitt). Catchitt may be applied to custom data with minimal effort but also provides a low-threshold code base to extend our approach by further features or to transfer these methods to related fields.

Conclusions

Accurate prediction of *in-vivo* TF binding remains a challenging task and computational predictions are not ready, yet, to replace experimental techniques like ChIP-seq. Compared with other approaches in the ENCODE-DREAM challenge, our approach yields a shared first rank among 40 international teams. Its streamlined open-source implementation Catchitt may provide a strong basis for further improvements by us and others in the future.

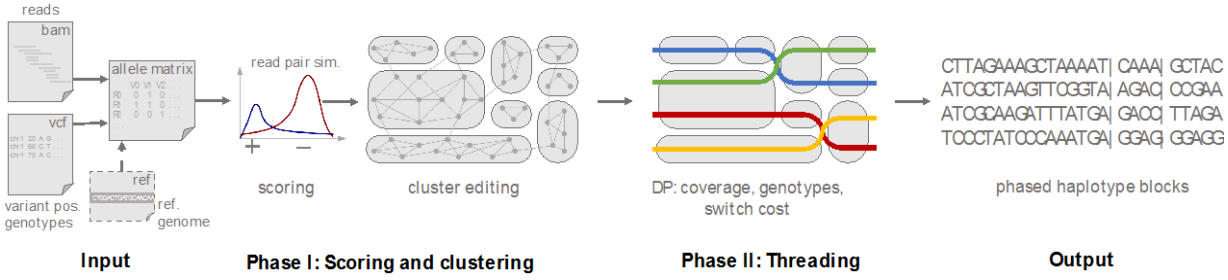
Haplotype Threading: Accurate Polyploid Phasing from Long Reads

S. Schrinner[1,★], R. Serra Mari[2,★], J. Ebler[2,3,★], G. W. Klau[1,4,☆], T. Marschall[2,3,☆]

[1]Heinrich-Heine-Universität Düsseldorf; [2]Universität des Saarlandes, Saarbrücken; [3]MPI für Informatik, Saarbrücken; [4]Cluster of Excellence on Plant Sciences (CEPLAS), Düsseldorf; ★joint first authors; ☆joint last authors

Background. The genome of many plant species, including important food crops, is polyploid. Resolving genomes at haplotype level is crucial for understanding the evolutionary history of polyploid species and for designing advanced breeding strategies. While phasing diploid genomes using long reads has become a routine step, polyploid phasing still presents considerable challenges. The Minimum Error Correction (MEC) model, the most common and successful formalization for diploid phasing, is limited in the use for polyploid phasing since it does not address regions where two or more haplotypes are identical. In addition, dynamic programming techniques solving diploid MEC become infeasible in the polyploid case.

Methods and Results. We present WHATSHAP POLYPHASE, a method for accurate polyploid phasing that overcomes these challenges by departing from the MEC model (see Figure). We propose a novel two-stage approach based on (i) clustering reads using a position-dependent scoring function and (ii) threading the haplotypes through the resulting clusters by dynamic programming. We demonstrate that our method scales to whole chromosomes and results in more accurate haplotypes than those computed by the state-of-the-art tool H-PoP. Our algorithm is implemented as part of the widely used open source tool WHATSHAP and is hence ready to be included in production settings.



Discussion. Current challenges lie in eliminating switch errors causing high Hamming rates and in scaling the algorithm to ploidies above six. We are presently exploring the ability of WHATSHAP POLYPHASE to phase polyploid plant genomes and to create maps of identical haplotype regions.

Segmenting Cells in Light and Electron Microscopy Images

Anna Kreshuk

EMBL - European Molecular Biology Laboratory, Heidelberg/Germany

Machine learning is the main driving force of the ongoing revolution in computer vision. State-of-the-art algorithms can segment multi-Terabyte microscopy volumes into single cells, detect tiny features in whole slide images and track single cells in a developing organism over many division events. I will give a brief overview of the current state of the field specifically for biological images and highlight the challenges ahead, both technical and methodological.

Segmenting Cells in Light and Electron Microscopy Images

Anna Kreshuk | *EMBL - European Molecular Biology Laboratory* | *Germany*

Abstract not available

The Influence of the Bone Marrow Niche on Drug Response Phenotypes of Blood Cancers

Sophie Rabe, University Hospital, Heidelberg; Eva Schitter, University Hospital, Heidelberg; Tobias Roider, University Hospital, Heidelberg; Carolin Kolb, University Hospital, Heidelberg; Mareike Knoll, University Hospital, Heidelberg; Junyan Lu, EMBL, Heidelberg; Peter Bruch, University Hospital, Heidelberg; Jennifer Hüllein, EMBL, Heidelberg; Marta Stolarczyk, University Hospital, Heidelberg; Christof von Kalle, NCT, Heidelberg; Christoph Lutz, University Hospital, Heidelberg; Peter Dreger, University Hospital, Heidelberg; Carsten Müller-Tidow, University Hospital, Heidelberg; Thorsten Zenz, Universitätsspital, Zürich; Wolfgang Huber, EMBL, Heidelberg; Sascha Dietrich, University Hospital, Heidelberg

Background

Signals provided by the microenvironment can modify and circumvent pathway activities that are therapeutically targeted by drugs. However, a systems-level understanding of how the microenvironment and the genetic and molecular alterations of the tumor interact with each other and contribute to drug resistance is lacking.

Methods

To address this unmet need, we established an automated microscopy-based phenotyping platform that uses co-culture conditions mimicking the bone marrow environment. We cultured primary tumor cells from more than 100 leukemia patients (CLL, AML, MCL, T-PLL, HCL) with and without bone marrow stroma cell support in DMEM and 10% human serum and treated each condition with 57 drugs in 3 concentrations. After 72h of incubation, 22 000 images per patient were acquired and processed. Our set-up allows us to increase sensitivity far beyond simple viability testing, as it reads out additional cell type specific features such as cell morphology, pathway activities and cell-cell interactions.

Results

Quality assessment revealed that in contrast to mono-culture conditions, assay plate edge effects can be avoided under stable stroma cell co-culture conditions. In the absence of their native microenvironment, primary leukemia cells undergo

spontaneous apoptosis *ex-vivo*. Viability at culture start was always >90% and dropped to a median of 51% (viability range: 17%-90%) after 72h in mono-cultures. Bone marrow stroma cell co-culture conditions protected tumor cells from spontaneous apoptosis (p=8.2e-6, paired t-test). Patient samples with a high degree of spontaneous apoptosis benefited most from co-culture conditions (p= 4.9e-12, Pearson correlation). To model interactions of stroma cell conditions and drug-induced apoptosis we established the following linear model: Viability ~ drug-effect + culture-model + drug-effect:culture-model. While activity of some drugs was significantly altered under co-culture conditions, we could also identify drugs with similar activity in mono- and co-cultures. For instance, the activity of common chemotherapeutics (fludarabine: p=0.002 at 0.6µM, cytarabine: p=0.001 at 1.5µM, ANOVA) or bromodomain inhibitors (I-BET-762: p=5.9e-5 at 4.5µM, JQ1: p=1.5e-8 at 1.5µM, ANOVA) was significantly reduced under co-culture conditions. In contrast, PI3K inhibitors idelalisib and duvelisib had a similar activity in mono-culture and stroma co-culture conditions and might represent a starting point to overcome stroma cell mediated drug resistance. A systematic comparison of *ex-vivo* drug response pattern in mono- and co-cultures across 171 drug conditions will be presented.

Discussion

Our results suggest that high throughput co-culture drug testing can be robustly performed and provide an unprecedented understanding of how the stroma cell microenvironment and the genetic make-up of tumor cells contribute to drug resistance and sensitivity.

Analyzing and modelling cell type-specific function with virtual cells

Martin Eberhardt, Laboratory of Systems Tumor Immunology, Hautklinik, Universitätsklinikum Erlangen and Friedrich-Alexander University (FAU) Erlangen-Nürnberg, Erlangen, Germany; Pia Wentker, Laboratory of Systems Tumor Immunology, Hautklinik, Universitätsklinikum Erlangen and Friedrich-Alexander University (FAU) Erlangen-Nürnberg, Erlangen, Germany; Florian S. Dreyer, Laboratory of Systems Tumor Immunology, Hautklinik, Universitätsklinikum Erlangen and Friedrich-Alexander University (FAU) Erlangen-Nürnberg, Erlangen, Germany; Martina Cantone, Laboratory of Systems Tumor Immunology, Hautklinik, Universitätsklinikum Erlangen and Friedrich-Alexander University (FAU) Erlangen-Nürnberg, Erlangen, Germany and Faculty of Mechanical Engineering, Specialty Division for Systems Biotechnology, Technische Universität München, Munich, Germany; Faiz Khan, Department of Systems Biology and Bioinformatics (SBI), University of Rostock, Rostock, Germany; Philipp Junk, Laboratory of Systems Tumor Immunology, Hautklinik, Universitätsklinikum Erlangen and Friedrich-Alexander University (FAU) Erlangen-Nürnberg, Erlangen, Germany; Olaf Wolkenhauer, Department of Systems Biology and Bioinformatics (SBI), University of Rostock, Rostock, Germany; Julio Vera, Laboratory of Systems Tumor Immunology, Hautklinik, Universitätsklinikum Erlangen and Friedrich-Alexander University (FAU) Erlangen-Nürnberg, Erlangen, Germany

Background

Cells process information through interactions of the molecules within and between them. Due to the sheer number of those interactions, cells are extremely complex molecular machines whose function is determined by non-linear regulatory processes. The challenge that molecular biologists and clinicians face is integrating cell function from the effects of molecular interactions.

Methods and Results

We produce and publish fully annotated, machine-readable, cell type-specific regulatory networks to analyze and model cell function. In case studies, we have applied graph theory-based analysis methods on networks enriched with high-throughput data to

identify drugs with a putative effect on pneumonia and metastatic melanoma. In the context of brain development, we built a mathematical model that is sufficient to explain the spatiotemporal progression of oligodendroglial differentiation. Our networks and results are available on https://vcells.net/.

Discussion

It is the combination of network biology, quantitative measurements and mathematical modelling that enables predictions on biological systems whose complexity is too vast to be understood intuitively. We borrow concepts and methods from these three fields to address specific questions in biomedicine, guiding experimental studies or generating hypotheses on pre-existing observations.

Spatial heterogeneity of cell response to cancer drugs

France Rose; Auguste Genovesio. Bio-Imagerie Computationnelle et Bioinformatique, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.

Robotics and automated fluorescence microscopes have promoted high-content cell-based screenings (HCS): cellular probes which target DNA or other major components are used to image hundreds of thousands of cells under many different conditions. Cell-based assays have proven to be efficient at discovering first-in-class therapeutic drugs, i.e. drugs acting on a new target [1]. We study spatial heterogeneity in a HCS image dataset [2] of cultured cells. We found that neighboring cells influence one another and display a similar phenotype more frequently than expected at random. This result, assessed across hundred of tested treatments, shows that, even for cultured cells with homogeneous genetic background, cells are not all alike and independent, but create spatial heterogeneity via cell lineage and interaction. This result can be put in parallel with recent approaches [3,4] which, by including cell neighborhood information, improved classification into mechanisms of action (MOA), gene's pathways, or cell types.
Once this primary observation of spatial heterogeneity was made, we wondered if we could use it to predict the MOA of a drug. We used graph kernels to compare images to images directly, integrating the cell subpopulations at the same time as their positions. We show, under some conditions, that including the positional information with the phenotypic heterogeneity improves the MOA classification.

[1] Swinney & Anthony (2011). How were new medicines discovered? Nat. Rev. Drug Discov.
[2] Ljosa et al. (2013). Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen*.
[3] Toth et al. (2018). Environmental properties of cells improve machine learning-based phenotype recognition accuracy. *Scientific Reports*.
[4] Rohban et al. (2018). Capturing single-cell heterogeneity via data fusion improves image-based profiling. *bioRxiv Bioinformatics*.

Dynamic optimization of the immune response during aspergillosis

Jan Ewald[1], Axel A. Brakhage[2,3], Christoph Kaleta[4], Stefan Schuster[1]

1: Department of Bioinformatics, Friedrich-Schiller-Universit Jena , Germany

2: Department of Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology (HKI), Jena, Germany

3: Institute of Microbiology, Friedrich-Schiller-Universität Jena , Germany

4: Medical Systems Biology, Christian-Albrechts-University Kiel, Germany

Background

The mold *Aspergillus fumigatus* is an opportunistic fungal pathogen and its spores (conidia) are ubiquitous and constantly inhaled by humans. While in the immunocompetent host immune cells are able to clear the pathogenic spores completely, in immunocompromised patients germinating conidia can develop filaments (hyphae) which damage the epithelial cell barrier of the lung alveoli. The resulting invasive aspergillosis is associated with very high mortality rates and incidence has been increasing over the years due to the higher number of immunocompromised patients in medical care.

Model and Results

Since containment of fungal growth in the lung is a race between the innate immune system and the development of *A. fumigatus*, a dynamic model is well suited to elucidate which key factors are decisive to prevent an invasive fungal infection. Unlike other models of invasive aspergillosis, we use dynamic optimization to gather insights in the recruitment and depletion of alveolar macrophages and neutrophils. Further, we model the different fungal growth states of resting and swollen conidia as well as hyphae to perform a detailed study of the importance of fungal parameters and host immune parameters for the outcome of infection. Our dynamic optimization approach determines the time optimal recruitment strategy of immune cells based on the objective that the number of pathogenic cells and tissue damage is minimized.

Starting from parameter values gathered from existing literature and experimental data we performed a parameter sensitivity analysis. Initial results show that a rapid immune response is crucial and neutrophils are the most important phagocytes. Further, our model is able to disclose the contribution of macrophages and epithelial cells to inflammation and supports experimental findings about the underestimated role of epithelial cells.

In conclusion our model reflects well experimental observations and provides new insights in the process of invasive aspergillosis. Beyond that the model can be utilized in the future to find better treatment protocols by focusing on the most critical components of the immune response or fungal virulence.

Identifying the unknowns with combinatorial and machine learning methods

K. Dührkop[1]; [1] Friedrich-Schiller-University Jena/D

Background Identification of small molecules remains a central question in analytical chemistry, in particular for natural product research, metabolomics, environmental research, and biomarker discovery. Mass spectrometry is the predominant technique for high-throughput analysis of small molecules. But it reveals only information about the mass of molecules and, by using tandem mass spectrometry, about the mass of molecular fragments. Automated interpretation of mass spectra is often limited to searching in spectral libraries, such that we can only dereplicate molecules for which we already have recorded reference mass spectra.

Methods and Results During my thesis we developed two methods that answer two important questions in untargeted metabolomics: What is the molecular formula of the measured ion and what is its molecular structure? SIRIUS is a combinatorial optimization method for annotating a spectrum and identifying the ion's molecular formula by computing hypothetical fragmentation trees. We developed a statistical model that describes the fragmentation tree computation as a maximum a posteriori estimator. This allows us to learn parameters and hyperparameters of the scoring directly from data.

CSI:FingerID is a method for predicting a molecular fingerprint from a tandem mass spectrum using kernel support vector machines. The predicted fingerprint can be searched in a structure database to identify the molecular structure. CSI:FingerID is based on FingerID, that uses probability product kernels on mass spectra for this task. We describe several novel kernels for comparing fragmentation trees instead of spectra. These kernels are combined using multiple kernel learning.

We evaluate the molecular formula identification rate of SIRIUS and the structure identification rate of CSI:FingerID on several datasets.

Discussion We demonstrate that the statistical model of SIRIUS, which was fitted on a small dataset, generalizes well across many different datasets and mass spectrometry instruments and significantly improves on the task of molecular formula assignment. In contrast, the prediction performance for CSI:FingerID drops significantly for compounds which are dissimilar to the training structures. However, the fragmentation tree kernels and the multiple kernel learning improves the generalization capability. We show on several datasets that CSI:FingerID outperforms competing methods for structural elucidation.

While machine learning methods show impressive results on image and speech recognition tasks, adapting them to new domains is highly non-trivial. We demonstrate how we can improve on the task of structural elucidation, by first transforming spectra into fragmentation trees. Both methods, SIRIUS and CSI:FingerID, are implemented in the freely available SIRIUS 4 software framework.

Bioinformatics meets Medical Informatics: Data requirements for Precision Medicine

Sabine Koch | Karolinska Institutet | Sweden

Since two decades several initiatives have described the conversion of Medical Informatics (MI) and Bioinformatics (BI) towards Biomedical Informatics (BMI) as a new discipline bridging informatics methods across the spectrum from genomic research to personalised medicine.
Personalised medicine is mainly associated with combining genomic with phenotypic data from electronic health records to provide tailored approaches to therapy. Precision medicine evolved from personalized medicine as an approach to disease management that considers individual variability in genes, environment, and lifestyle, focusing on understanding individual variability in disease prevention, care, and treatment.
The use of new data types and data sources by precision medicine thus poses new challenges to biomedical informatics that will require new solutions. This keynote will detail these challenges and discuss the relevance and impact of different data sources for population health and clinical care.

The Portal of Medical Data Models – how can it support Bioinformatics research?

Martin Dugas[a,1], *Sarah Riepenhausen*[a], *Julian Varghese*[a], *Philipp Neuhaus*[a], *Cornelia Mertens*[a], *Alexandra Meidt*[a] *and Stefan Hegselmann*[a]

[a] *Institute of Medical Informatics, University of Münster, Germany*

[1] *Corresponding author: dugas@uni-muenster.de*

Background

Due to the complexity of medical terminology (>800.000 terms in SNOMED) independent creation of compatible data models is improbable: Sharing is necessary for harmonized data collection. The Portal of Medical Data Models (https://medical-data-models.org) promotes transparency, compatibility and harmonization in medical information systems. It is essential for the analysis of big data sources like next-generation sequencing (NGS) datasets to have precise information on the clinical phenotype (e.g. exact diagnosis and disease course). Often these clinical data are the bottleneck of medical research. Semantic annotation of data elements can foster re-use of data in various settings. Researchers, clinicians and the patient may benefit from the resource-saving re-use of data and data models.

Methods and results

The portal's data models are created with ODMEdit in CDISC ODM format. ODMedit enables re-use of items, itemgroups and UMLS codes supporting uniform semantic coding. MeSH keywords facilitate search functions.

In March 2019 the portal had >1,200 registered users and >20,300 data models. To our knowledge, it constitutes Europe's largest collection of medical forms. 88% of >480,000 available items were UMLS annotated. "Clinical Trial" (>15,000) was the most frequent keyword. The most frequent disease area was "Neoplasms" (>7,700). 1,268 data models belonged to routine documentation.

Discussion

To properly interpret results from biological samples, bioinformatics research needs high-quality, harmonized patient data regarding medical history, clinical and apparative diagnostics, comorbidities, treatment and disease course. The Portal of Medical Data Models provides transparency about data structure of existing data sources and can foster harmonized data collection through reuse of tried and tested data models. These show a broad coverage, especially in clinical trials and oncology.

The need for an automated whole exome sequencing analysis pipeline to support a molecular tumor board

Patrick Metzger[1,2,3], Maria Hess[1,4], Victor Jaravine[2], Martin Boeker[2,4], Jan Christoph[5], Philipp Büchner[5], Philipp Unberath[5], Hauke Busch[6], Geoffroy Andrieux[1,7] and Melanie Boerries[1,4,7]

[1]Department for Biometry, Epidemiology and Medical Bioinformatics, Medical Center, Faculty of Medicine, University Freiburg, Germany; [2]Institute for Medical Biometry and Statistics, Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany; [3]Facutly of Biology, University of Freiburg, Freiburg, Germany; [4]Comprehensive Cancer Center Freiburg (CCCF), University Medical Center, Faculty of Medicine, University of Freiburg, Germany; [5]Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany; [6]Luebeck Institute of Experimental Dermatology and Institute of Cardiogenetics, University of Luebeck, Luebeck, Germany; [7]German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), partner site Freiburg, Freiburg, Germany

Background: Whole exome sequencing (WES) of patients with advanced tumors is becoming an established method in medical centers. However, somatic variant calling and interpretation as well as report creation and case presentation require both in-depth knowledge in bioinformatics and oncology. In addition to the analysis and interpretation of the data, the exchange of results and recommendations of a molecular tumor board (MTB) is also of vital importance. A prerequisite to achieve this is data harmonization and integration. For this reason, the German Ministry of Education and Research (BMBF) has launched the Medical Informatics Initiative (MI-I), which aims to integrate all universities and university clinics in Germany through several consortia. MIRACUM (Medical Informatics for Research And Care in University Medicine), one of the four consortia funded by the German government, is rolling out three use cases in ten university hospitals throughout Germany. Among others the support of interdisciplinary molecular tumor boards, which combine extensive molecular diagnostics with state-of-the-art sequencing and bioinformatic analysis for personalized recommendation.

Methods and Results: In order to offer an adequate solution, we have developed a fully automated WES analysis pipeline (MIRACUM-Pipe), which only requires the

sequencing files (fastq files) and the patient's gender as input, in order to finally present the results in an automatically generated report. The pipeline consists of three parts: (i) Alignment and quality control, (ii) analysis and annotation which is further divided in coverage, variant calling and copy number variations and (iii) final results reporting. It incorporates a bash script running the WES data processing and basic annotation as well as R scripts for a complex functional annotation and downstream analysis of the results. The called somatic variants are annotated using ANNOVAR and are further classified into tumor suppressors (TSG) or oncogenes (OG) according to OncoKB, cancer hotspot mutations are marked based on cancerhotspots.org, while possible therapy options are identified from OncoKB, TARGET and the drug-gene-interaction databases (DGIdb). For further biological interpretation of the called variants a functional enrichment is performed based on Gene Ontology (GO) terms, Reactome and ConsensusPathDB. In addition, COSMIC mutational signatures are annotated. The results are finally summarized in a PDF report and are also made available in tabular form.

Discussion: The MIRACUM-Pipe is currently used for the MTB at the University Medical Center Freiburg and has been further implemented at the MIRACUM partner sites Mainz and Gießen. With this approach we will support existing and newly established MTBs at the MIRCUM partner sites with the aim to define a standardized procedure for analysis and annotation of WES for molecular tumor board’s patients. Of course, the general discussion about the definition of cutoff of e.g. variant allele frequency has to be considered, but this approach will give us a unique opportunity to considerably enhance the comparability of these data/results at several partner sites simultaneously in order to make any necessary adjustments.

Acknowledgements: MIRACUM is funded by the German Ministry for Education and Research (BMBF) within the "Medical Informatics Funding Scheme" (FKZ 01ZZ1801B).

GestaltMatcher: Finding the second patient of a novel syndrome with next-generation phenotyping

Tzung-Chien Hsieh, University of Bonn, Bonn, Germany; Aviram Bar Haim, FDNA, Boston, United States; T.J. Pantel, Charité Universitätsmedizin Berlin, Berlin, Germany; Y. Gurovich, FDNA, Boston, United States; Y. Hanani, FDNA, Boston, United States; T. Kamphans, GeneTalk, Bonn, Germany; P. Krawitz, University of Bonn, Bonn, Germany

Introduction: Recent advances in next-generation phenotyping (NGP) tools for syndromology such as DeepGestalt have learned phenotype representations of multiple disorders by training on thousands of patient photos. However, many of Mendelian syndromes are still not represented by existing NGP tools as only handful of patients were diagnosed or represented in the databases. We therefore propose a facial gestalt matching approach to quantify the similarity among patients and further explore the uncovered phenotypic space.

Methods and results: We compiled a dataset consisting of 15000 patients with 500 different monogenic disorders. For each individual a frontal photo and the molecularly confirmed diagnosis were available. We utilized deep convolutional neural network to train model on patient's frontal photos, and facial embeddings were further extracted to construct a Clinical Face Phenotype Space (CFPS). We used Joint Baysian to quantify the similarity among patients and correct the confounding effects of age and ethnicity in CFPS. Given only the facial phenotypes as input we could visualize molecular interactions of genes by their close proximity in the CFPS. A prominent cluster was formed by the genes of the MAPKinase pathway that result in disorders such as Noonan syndrome. Beyond this proof of concept, we were able to match patients with the same facial dysmorphic phenotype, of syndromes that are not represented by previous work such as DeepGestalt.

Discussion: GestaltMatch is able to match the patients with the similar phenotype. It opens the door to extend the coverage of phenotypes and to enable further exploration of the unknown phenotype-genotype associations.

Deep Learning on Chaos Game Representation for Proteins

Hannah F. Löchel, Dominic Eger, Theodor Sperlea, Dominik Heider

Department of Bioinformatics, Philipps-University of Marburg, Marburg, Germany

Background

Machine learning has been widely used it protein classification. However, before classifying protein sequences by machine learning techniques, they need to be made machine-readable. To this end, different encodings exist, which are typically based on physical or chemical properties. The Chaos Game Representation (CGR) is a promising alternative, which transforms sequences into fractal images [1]. While CGR has been used mainly for DNA, e.g., for genomic sequences [2], we propose a novel frequency matrix Chaos Game Representation (FCGR) for encoding of protein sequences, which is based on n-flakes fractals.

Methods and Results

We implemented the R package *kaos* [3] for encoding sequences as FCGR-images. The package can be used for DNA as well as for protein sequences, leading to fractal structures that can be used in subsequent image classification studies, e.g., based on deep neural networks (DNNs). As a proof of concept, we encoded HIV-1 protease and reverse transcriptase with known resistance against 14 different drugs with FCGR. We subsequently trained DNNs, support vector machines (SVMs), and random forests (RFs) on these data. We could demonstrate that all models show promising results compared to the state-of-the-art methods and that the DNNs outperform the other methods.

Discussion

FCGR is a promising new encoding method for protein sequences, which has been demonstrated on HIV-1 drug resistance benchmark datasets. In the future, we will analyze the dependencies between sequence lengths and classification performance of DNNs trained on FCGRs.

References

[1] Jeffrey HJ (1990): Chaos game representation of gene structure. Nucleic Acids Research, 18(8): 2163-70.

[2] Wang Y, Hill K, Singh S, Kari L (2005): The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene, 346: 173-85.

[3] Loechel HF, Eger D, Sperlea T, Heider D (2019): Deep Learning on Chaos Game Representation for Proteins. bioRxiv, 575324

Learning Embeddings from a Biomedical Knowledge Graph for Predicting Novel Relations

Timo Sztyler, Brandon Malone

NEC Laboratories Europe, Heidelberg, Germany

1 Background

Biologists, physicians, clinical researchers, and others have spent countless hours studying the regulatory behavior among genes, the effects of medications on diseases, the correlation of therapy outcomes among patients, and knowledge about many other relationships. These successful discoveries improve our quality of life, but their discovery is time-consuming and expensive. Therefore, we propose a knowledge graph-based approach that predicts novel relationship by automatically augmenting the known relationships with predictions from a graph neural network.

2 Methods and Results

	disease-protein		polypharmacy	
	AuROC	Avg. Prec.	AuROC	P@50
Baseline	71%	69%	39%	53%
KBLRN	86%	81%	77%	83%

We construct a knowledge graph (KG) which describes known associations, similarities and interactions among diseases, proteins and drugs. These relationships are extracted from publicly-available resources. More precisely, we predict novel *disease-gene associations* as well as *polypharmacy side effects*, that is, unexpected side effects due to drug-drug interactions. For this, we apply the recently-proposed KBLRN to our KG. KBLRN complements a standard neural relational approach with *path features* based on the KG in a mixture-of-experts model. Quantitatively, our approach outperforms strong baselines on both prediction tasks (see Table 1). The path features offer explanations for our predictions. As a concrete example: It is known that the proteins LPAR1 and MMP2 interact with each other and this information is captured in our KG. Similarly, the KG is aware that the drug paliperidone affects LPAR2, whereas the drug calcium affects MMP2. These relationships are captured as path features and used during learning. Our model predicts that taking both drugs in combination leads to the previously unknown side effect ``inflammation'', while explainable AI techniques reveal that the path feature is important for that prediction. Indeed, a manual literature review confirms this effect. Thus, our approach can also facilitate scientific hypothesis generation due to the explainability of the path features.

3 Discussions

Our results show that modern neural, yet explainable, approaches can improve on standard approaches to automatically augment knowledge bases represented as KGs. While our KG is constructed from public data sources based on the primary literature, such as PubMed abstracts, we do not yet directly incorporate such resources into our model. Thus, a next step in our work is to incorporate natural language processing models into our mixture of experts.

Posters

Inference of DNA copy number alterations from single-cell RNA sequencing data using transcriptional regulatory networks

Ronja Johnen, CECAD, Cologne, Germany; Luise Nagel, CECAD, Cologne, Germany; Ana Carolina Leote, CECAD, Cologne, Germany; Manuel Lentzen, CECAD, Cologne, Germany; Andreas Beyer, CECAD, Cologne, Germany, CMMC, University of Cologne, Cologne, Germany, CSCB, University of Cologne, Cologne, Germany

Somatic copy number alterations (CNAs) are an important contributor to diseases such as cancer and it potentially contributes to the emergence of age-associated phenotypes such as accumulating senescent cells. The detection of rarely occurring early-stage CNAs requires single cell measurements.

Here, we present a comparison of single-cell CNA detection methods using single cell RNA-seq (scRNA-seq) data. scRNA-seq data is often more readily available than single cell DNA measurements, however, expression variation complicates the detection of CNAs based on RNA-measurements. In order to address this issue, we developed a new method based on a genome-wide transcriptional regulatory network, which enables us to more accurately detect altered gene expression due to CNAs. The relative copy number state of a gene is estimated from the deviation between the measured expression and the predicted expression based on genes in the relevant regulatory subnetwork. Using a large reference dataset based on different scRNA-seq protocols, we show that our network approach outperforms published methods in identifying CNAs.

In further studies, this improved method can be used to infer CNVs in aging cells and to investigate their region- and tissue-specific accumulation with age.

Gene expression and TMB of MSI-H and POLE mutation by colorectal cancer

Klaus Kluck, Moritz von Winterfeld, Martina Kirchner, Peter Schirmacher, Albrecht Stenzinger, Jan Budczies; Institute of Pathology, University Hospital Heidelberg, 69120 Heidelberg, Germany

Background: POLE mutation (POLE-mut) and mismatch repair deficiency resulting in microsatellite instability (MSI-H) represent separate causes of hypermutation in colorectal cancer (CRC). Hypermutated cancer is known to respond well to immune checkpoint blockade exemplified by the histology-agnostic FDA approval of pembrolizumab for MSI-H cancer in 2017. To gain inside in the immune biology of hypermutated CRC, we analyzed mRNA expression and tumor mutational burden (TMB) of POLE-mut, MSI-H and microsatellite stable (MSS) CRC.

Methods and Results: Our finding cohort consisted of three groups of in house CRC samples: POLE-mut (n=21), MSI-H (n=7) and the control group MSS (n=18). Our validation cohort consisted of CRC samples from the cancer genome atlas (26 POLE-mut, 49 MSI-H, 313 MSS). We analyzed the mRNA expression of the 770 genes in the PanCancer IO 360 panel from NanoString Technologies with the R package DESeq2 to find significant differentially expressed genes. For validation we checked if fold change was significant ($p<0.05$, one-sided t-test) and had the same sign. We found 132 significant differentially expressed genes between POLE-mut and MSS and 149 significant differentially expressed genes between MSI-H and MSS. We observed a significant overlap of 39 genes between these analyses ($p=0.0023$, Fisher's exact test). 24 of the 132 (17.8%) detected genes (POLE-mut vs. MSS) and 45 of the 149 (30.2%) detected genes (MSI-H vs. MSS) could be validated in the TCGA data. In the CRC TCGA data, we analyzed TMB and the abundance of specific mutation types. POLE-mut samples had significantly more missense, silent and nonsense mutations, but less indels than MSI-H samples. For all these mutation types, TMB was lower in MSS tumors compared to POLE-mut and to the MSI-H tumors.

Discussion: We uncovered and validated gene expression changes between hypermutated and normal-mutated CRC. For biological interpretation of the detected genes, the complex tumor microenvironment of CRC including not only tumor cells, but also immune cells, fibroblasts and other cells needs to be taken into account.

SPLICE-q: a Python tool for Genome-Wide Determination of Splicing Efficiency

Veronica Melo Costa, Max Planck Institute for Molecular Genetics and Freie Universität Berlin, Berlin, Germany; Julianus Pfeuffer, Universität Tübingen, Tübingen, Germany; Evgena Ntini, Max Planck Institute for Molecular Genetics, Berlin, Germany; Ulf Oerom, Aarhus Universitet, Aarhus C, Denmark; Rosario Piro, Freie Universität Berlin and Charité - Universitätsmedizin Berlin, Berlin, Germany.

Eukaryotic genes are mostly composed of a number of exons intercalated by introns that are generally removed from pre-mRNAs to form mature RNA molecules. This post-transcriptional process is called splicing which consists basically of a series of hydrolysis and ligation reactions lead by the spliceosome. Splicing is dynamic and can occur right after the transcription of a complete intron. An efficient pre-mRNA splicing is essential and its mis-regulation is related to numerous human diseases. To better understand the dynamics of this process and the perturbations that might be caused by aberrant transcript processing it is important to quantify its efficiency. Splicing efficiency (SE) is commonly calculated with the use of RT-qPCR with primers that span exon-exon and exon-intron boundaries. Yet, this methodology can only investigate a limited number of genes. By contrast, RNA-seq allows these analyses from a genome-wide point of view. There are frameworks for calculating SE from RNA-seq data, but the bioinformatics steps involved might be challenging, especially for experimental biologists. Thus, we presented a complete, up-to-date and easy-to-use python tool for genome-wide quantification of SEs from total RNA-seq data. The tool quantifies SEs for each intron using a coverage based method. For each intron, it takes aligned reads and an annotation file as input, calculates coverage on the splice junctions (SJ), extracts split and non-split reads and estimates a Splicing Efficiency Score (SES) as a combination of both in the SJs pair. SPLICE-q allows the user to select different levels of restrictiveness for filtering, including: (I) select only introns that do not overlap with any exon of the same or different gene; (ii) select only introns whose SJs do not overlap any exon; (iii) keep all introns in the gemone regardless of overlaps with other genomic elements. Other filtering can also be set up according to minimum coverage, read quality, and minimum intron length. We will describe the main architecture of SPLICE-q and show examples of its usage utilizing time series nascent RNA-Seq.

Defined subsets of functional endogenous LINE-1 loci that mobilize in human tumors are activated and mobilized in human embryonic and induced pluripotent stem cell lines

Gerald G. Schumann, Paul Ehrlich Institute, Germany; Julien Duc, École Polytechnique Fédérale de Lausanne, Switzerland; Marc Friedli, École Polytechnique Fédérale de Lausanne, Switzerland; Liam Childs, Paul Ehrlich Institute, Germany; Ulrike Held, Paul Ehrlich Institute, Germany; Tanja Kearns, Paul Ehrlich Institute, Germany; Tiffany Kaul, Tulane Cancer Center, US; Prescott Deininger Tulane Cancer Center, USA; Didier Trono, École Polytechnique Fédérale de Lausanne

Background

Human induced pluripotent stem cells (hiPSCs) are used for disease modeling, the study of cell development and function and hold substantial promise for cell therapies. However, reprogramming and subsequent cultivation of hiPSCs can induce genetic and epigenetic abnormalities that could elevate the risk of tumorigenesis. We have recently demonstrated that reprogramming-induced epigenome remodeling in hiPSCs results in mobilization of the endogenous retrotransposons LINE-1 [L1], *Alu* and SVA, and that intronic L1 *de novo* insertions occur during reprogramming, which can interfere with host gene expression and corrupt the biosafety of hiPSCs and/or their differentiated derivatives (1).

Here, we set out to identify which loci, among the $5x10^5$ L1 insertions in the genome of human pluripotent stem cells (hiPSCs), encode retrotransposition-competent (RC) L1 elements that are both expressed and mobilized. We investigate if the retrotransposition activity observed in hiPSCs and human embryonic stem cells (hESCs) results from genome-wide activation of all full-length L1 copies, or from the activation of specific subsets of functional L1 loci.

Methods and results

We applied RNA-seq procedures and a method coupling 5'-RACE with PacBio sequencing on total RNA preparations of 2 hESC, 6 hiPSC and two parental cell lines to identify which L1 loci are expressed in hiPSC lines in which L1-mediated mobilization was demonstrated to occur. After alignment, presence/absence calls were made for all functional, full-length L1 elements in addition to identifying the significantly differentially expressed elements. The subset of all known

retrotransposition-competent elements was curated from published research and examined for evidence of retrotransposition activity in the hiPSCs.

Of the 174 retrotransposition-competent L1 loci identified in the analyzed cell lines, 102 were polymorphic. By combining RNA-seq data and 5'-RACE/PacBio-data, we found a specific subset of only 32 RC-L1 loci that are part of the reference genome, including 4 out of 10 highly active L1 elements, were responsible for the majority of retrotransposition activity in humans. Furthermore, an additional 13 polymorphic L1 loci were shown to be transcribed in hiPSC and hESC lines. Three L1 source elements causing most of the transposition events in epithelial cancers were found to be highly expressed in hiPSC lines. Finally, we identified the individual L1 source loci causing individual mutagenic L1 *de novo* insertions that occurred during the reprogramming process of hiPSC lines.

Discussion

The presented data underscore the significant potential of expressed functional L1 elements as endogenous mutagen in hiPSCs. We show that not only are L1 elements expressed in hiPSCs but they are also capable of retrotransposition and introduce a random mutagenic factor in the cell. Furthermore, hiPSC lines derived from different parental cell types show differential upregulation of L1 elements demonstrating the need to carefully select the necessary parental cell type depending on the application. In the presented data, the set of expressed, functional L1 loci identified in hiPSC lines derived from cord blood-derived endothelial cells showed the greatest similarity to the set of L1 loci expressed in hESCs.

Identifying gene signatures: purely data-driven approaches in analysis of the transcriptomic data and its implementation in immunological analysis

Bin Liu, German Center for Lung Research, partner site Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), Hannover, Germany; Institute of Technical Chemistry, Leibniz University of Hannover, Hannover, Germany; Patrick Lindner, Institute of Technical Chemistry, Leibniz University of Hannover, Hannover, Germany; Adan Chari Jirmo, Department of Pediatric Pneumology, Allergology and Neonatology, Hannover Medical School, Hannover, Germany; German Center for Lung Research, partner site Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), Hannover, Germany; Ulrich Maus, Division of Experimental Pneumology, Hannover Medical School, Hannover, Germany; Thomas Illig, German Center for Lung Research, partner site Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), Hannover, Germany; Hannover Unified Biobank, Hannover Medical School, Hannover, Germany; and David S. DeLuca, German Center for Lung Research, partner site Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), Hannover, Germany;*

Background:

Despite the significant contribution of transcriptomics to the fields of biological and biomedical research, interpreting large lists of significantly differentially expressed genes remains a challenging step in the analysis process. This paper focuses on a data-derived signature approach of testing differential expression in groups of functionally related genes and shows the possibilities of interpreting the transcriptomic data using previous data sets. Particularly, the paper shows the feasibility of this approach in analyzing immune-related processes, which are complicated in their nature but play an important role in the medical researches.

Methods and results:

We demonstrated four different methods in this paper that can be potentially applied in the data-derived signature analysis on new data sets. In addition to the widely

utilized Mann-Whitney-Wilcoxon Test, we also reformed the Fisher's Exact Test into a semi-quantitative method based on regulation directions and suggested two permutation-based methods in the detection of data-derived signatures. We applied all permutation-based methods on a self-built immune-related database and compared the performance of these methods on this database. The result shows that all the four methods can potentially be applied in the detection of data-derived signatures, with the signature concordance showing the highest sensitivity and the Fisher's Exact Test showing best overall performance. We also compared the data-derived signature to the publicly available pre-defined gene sets to compare the sensitivity and specificity of each approach.

Conclusions:

Compared to ubiquitously utilized approaches based on human, machine or literature parsed gene sets, the data-derived signature approach can fully reflect the complex nature of immunological processes and is robust against technical noise. Given the large volume of previously generated transcriptome profiles, the data-derived signatures can be generated in the cases that well-defined gene sets are lacking from pathway databases, particularly in immunological researches. All these advantages make the approach a good candidate for the interpretation of transcriptomic data. Overall in this paper, we propose a data-derived approach and demonstrate four methods potentially feasible to the application of this approach.

Decoding Nanopore Signals with Neural Networks for Cell Type-Specific Expression Profiling

T. Baar, Institute of Medical Statistics and Computational Biology IMSB, Faculty of Medicine, University of Cologne, Germany; M. Ignarski, Dept. II of Internal Medicine and Center for Molecular Medicine Cologne, University of Cologne, Germany; S. Dümcke, Institute of Medical Statistics and Computational Biology IMSB, Faculty of Medicine, University of Cologne, Germany; I. Helmuth, Institute of Pharmacy and Biochemistry, Johannes Gutenberg-University Mainz, Germany; J. Hertler, Institute of Pharmacy and Biochemistry, Johannes Gutenberg-University Mainz, Germany; Mark Helm, Institute of Pharmacy and Biochemistry, Johannes Gutenberg-University Mainz, Germany; R. Müller, Dept. II of Internal Medicine and Center for Molecular Medicine Cologne, University of Cologne, Germany; A. Tresch, Institute of Medical Statistics and Computational Biology IMSB, Faculty of Medicine, University of Cologne, Germany

Transcriptome analysis has come a long way, starting with microarrays, followed by short-read sequencing, and, recently, the development of single-molecule sequencing methods, such as Oxford Nanopore. Nanopore sequencing utilises a microfluidic device for direct (i.e., without cDNA conversion), full-length RNA sequencing. The transcript of interest is threaded through a protein pore. Nucleotides residing inside the pore influence an ionic current which flows through it. The current is then measured and analysed to identify the nucleotide sequence.

Using Nanopore sequencing, we aim to generate *in vivo*, cell type-specific expression profiles. We achieve this through metabolic labelling of nascent RNA by 5-ethynyl uridine (EdU), a uridine (U) analogue. Thus, our approach avoids several drawbacks of previous methods resulting from RNA purification, reverse transcription and amplification.

Here, we demonstrate the feasibility of the approach. We have synthesised RNA in which U was either replaced entirely by EdU or not at all. Using neural network classifiers, consecutive current signal chunks were analysed to discriminate labelled from unlabelled transcripts. We benchmarked the performance of different neural network architectures (CNNs and LSTMs) and show that these methods outperform classical machine learning approaches, such as Gaussian mixture clustering and support vector machines.

Ribosome Profiling with Bayesian Predictions to Survey the Cardiac Landscape of Translation

E. Boileau[1,2,3], I. Atanassov[4], E. Riechert[2,3], S. van Heesch[5,3], C. Hofmann[2,3], S. Mayer[2,3], A.A. Gorska[2,3], S. Kreher[6,3], A. Konzer[6,3], F. Leuschner[2,3], A. Schneider[6,3], J. Graumann[7,3], T. Braun[6,3], S. Doroudgar[2,3], H.A. Katus[2,3], N. Hübner[5,3], M. Völkers[2,3], C. Dieterich[1,2,3]

1 Section of Bioinformatics and Systems Cardiology, Klaus Tschira Institute for Integrative Computational Cardiology, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany and 2 Department of Internal Medicine III (Cardiology, Angiology, and Pneumology), University Hospital Heidelberg, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany and 3 DZHK (German Centre for Cardiovascular Research) and 4 Proteomics Core Facility, Max Planck Institute for Biology of Ageing, Joseph-Stelzmann-Straße 9B, 50931 Cologne, Germany and 5Genetics and Genomics of Cardiovascular Disease, Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Straße 10, 13125 Berlin, Germany and 6 Department I Cardiac Development and Remodelling, Max Planck Institute for Heart and Lung Research, Ludwigstraße 43, 61231 Bad Nauheim, Germany and 7Biomolecular Mass Spectrometry, Max Planck Institute for Heart and Lung Research, W.G. Kerckhoff Institute, Ludwigstraße 43, 61231 Bad Nauheim, Germany

Recent studies have revealed a previously unknown complexity of RNA translation across different species. Taking advantage of the resolution of Ribo-seq, we present a comprehensive catalog of *in vivo* translation events derived from the cardiac translatome of 35 murine hearts. We used the Ribo-seq methodology with an *in vivo* ribosome-tagging approach (RiboTag) to target translation from cardiomyocytes [1,2].

Our bioinformatics workflow is based on an unsupervised Bayesian approach, Rp-Bp, to predict translated open reading frames (ORFs) from ribosome profiles [3]. Rp-Bp incorporates Markov Chain Monte Carlo techniques to estimate posterior distributions of the likelihood of translation. The automatic Bayesian selection of read lengths and ribosome P-site offsets (BPPS) method allows to select and shift aligned reads to account for the P-site of the ribosome. Our workflow is complemented with evolutionary sequence conservation, characterization using protein signatures and evidence from mass spectrometry. Our results confirm the potential role of small non-canonical translation events in the heart and show that translation is pervasive upstream of annotated coding sequences and modulated under stress condition.

References

[1] V Kmietczyk, E Riechert, L Kalinski, E Boileau, E Malovrh, B Malone, A Gorska, C Hofmann, E Varma, L J”urgensen, V Kamuf-Schenk, J Altmüller, R Tappu, M Busch,

P Most, HA Katus, C Dieterich, and M Völkers. m6a-mrna methylation regulates cardiac gene expression and cellular growth. Life Sci Alliance, 2(2):e201800233, 2019.

[2] S Doroudgar, C Hofmann, E Boileau, B Malone, E Riechert, AA Gorska, T Jakobi, C Sandmann, L Jürgensen, V Kmietczyk, E Malovrh, J Burghaus, M Rettel, F Stein, F Younesi, U Friedrich, V Mauz, J Backs, G Kramer, HA Katus, C Dieterich, M Völkers. Monitoring Cell-Type-Specific Gene Expression Using Ribosome Profiling In Vivo During Cardiac Hemodynamic Stress. Circ Res, 125:431-448, 2019.

[3] B Malone, I Atanassov, F Aeschimann, X Li, H Grosshans, and C Dieterich. Bayesian prediction of rna translation from ribosome profiling. Nucleic Acids Res, 45(6):2960-2972, 2017.

circtools – a bioinformatics toolbox for circular RNAs

Tobias Jakobi, Department of Cardiology, Angiology, and Pneumology, University Hospital Heidelberg, Heidelberg, Germany, Alexey Uvarovskii, Department of Cardiology, Angiology, and Pneumology, University Hospital Heidelberg, Heidelberg, Germany, Christoph Dieterich, Department of Cardiology, Angiology, and Pneumology, University Hospital Heidelberg, Heidelberg, Germany

Background

Circular RNAs (circRNAs) originate through back-splicing events from linear primary transcripts, are resistant to exonucleases, not polyadenylated, and have been shown to be highly specific for cell type and developmental stage. Although few circRNA molecules have been shown to exhibit miRNA sponge function, for the majority of circRNAs, however, their function is yet to be determined. Prediction of circRNAs is a multi-stage bioinformatics process starting with raw RNA sequencing data and yielding, depending on tissue and condition, sets of hundreds of potential circRNA candidates that require further analyses. While a number of tools for the prediction process already exist, publicly available downstream analysis tools are rare. We aim to provide researches with a harmonized work flow that covers different stages of in silico circRNA analyses, from prediction to first functional insights.

Methods and results

Here, we present circtools (Jakobi et al. 2018), a modular, Python-based framework that unifies several *in silico* circRNA analyses in a single command line-driven toolbox. Our software includes modules for circRNA detection and reconstruction that are based on our well tested DCC (Cheng et al. 2013) and FUCHS (Metge et al. 2017) tools. Circtools however, was developed as a complete analysis work flow and therefore also contains additional modules to perform initial quality checks, test circRNAs for host gene independent expression, identify differentially spliced exons, screen circRNAs for enriched features (e.g. RBP sites), and design circRNA-specific primers for qRT-PCR verification. Circtools supports researchers with visualization options and data export into commonly used formats. We intend to add more modules in the future in order to provide a comprehensive bioinformatics toolbox for the research community and encourage users to contribute new modules.

Availability

Circtools is available under GPL-V3.0 license on GitHub via http://circ.tools; an extensive documentation is located at http://docs.circ.tools.

A combined computational pipeline to detect circular RNAs in human cancer cells under hypoxic stress

Antonella Di Liddo, Buchmann Institute for Molecular Life Sciences and Institute of Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany; Camila de Oliveira Freitas Machado, Institute of Cell Biology and Neuroscience and Institute for Cardiovascular Regeneration, Goethe University, Frankfurt am Main, Germany; Sandra Fischer, Department of Biology, Technical University Darmstadt, Germany; Stefanie Ebersberger, Institute of Molecular Biology (IMB), Mainz, Germany; Andreas W. Heumüller, Institute for Cardiovascular Regeneration, Goethe University, Frankfurt am Main, Germany; Julia E. Weigand, Department of Biology, Technical University Darmstadt, Germany; Michaela Müller-McNicoll, Institute of Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany; Kathi Zarnack, Buchmann Institute for Molecular Life Sciences, Goethe University, Frankfurt am Main, Germany

Hypoxia is associated with several diseases, including cancer. Cells that are deprived of adequate oxygen supply trigger transcriptional and posttranscriptional responses, which control cellular pathways like angiogenesis, proliferation and metabolic adaptation. Circular RNAs (circRNAs) are a novel class of mainly non-coding RNAs that have been implicated in multiple cancers and attract increasing attention as potential biomarkers.

Here, we characterize the circRNA signatures of three different cancer cell lines from cervical (HeLa), breast (MCF-7) and lung cancer (A549) under hypoxia. In order to reliably detect circRNAs, we integrate available tools with custom approaches for quantification and statistical analysis. Using this consolidated computational pipeline, we identify about 12,000 circRNAs in the three cancer cell lines. Their molecular characteristics point to an involvement of complementary RNA sequences as well as *trans*-acting factors in circRNA biogenesis. Notably, we detect a number of circRNAs that are more abundant than their linear counterparts. In addition, 64 circRNAs significantly change in abundance upon hypoxia, in

most cases in a cell type-specific manner.

In summary, we present a comparative circRNA profiling in human cancer cell lines, which promises novel insights into the biogenesis and function of circRNAs under hypoxic stress.

Improving Subtype Classification of Non-Small Cell Lung Cancer by Independent Component Analysis

Thomas Eveno, Luxembourg Institute of Health, Strassen, Luxembourg; Maryna Chepeleva, Luxembourg Institute of Health, Strassen, Luxembourg; Tony Kaoma Luxembourg Institute of Health, Strassen, Luxembourg; Arnaud Muller, Luxembourg Institute of Health, Strassen, Luxembourg; Francisco Azuaje Luxembourg Institute of Health, Strassen, Luxembourg; Petr Nazarov, Luxembourg Institute of Health, Strassen, Luxembourg

Background

Characterization of lung cancer patients using gene expression data from large publicly available datasets, such as the Cancer Genome Atlas (TCGA), shows promising results in finding new patterns for drug response, sub-type classification or survival analysis and paves the way for precision medicine. However, due to significant biases across platforms (batch effect) researchers often fail to obtain comparable results on small independent cohorts. To tackle this issue, we propose a data-driven deconvolution method: Independent Component Analysis (ICA). ICA extracts independent source signals from our merged dataset —i.e. reference and independent cohort— and allows us to do analysis only on signals not associated with batch effects.

Method

We analyse gene expression data from patients having Non-Small Cell Lung Cancer disease (NSCLC) and classify them into one of the two main groups corresponding to sub-types: adenocarcinoma (AD) or squamous cell of the lung (SQ). Classifiers are trained and evaluated by cross-validation on a large reference dataset. We then tested classifiers on many small independent cohorts. Pre-processing step is achieved using our consensus ICA algorithm (https://gitlab.com/biomodlih/consica) on the merged dataset. It maps gene expression features into relevant biological signals cleared from batch effect. Our method is compared with other pre-processing or feature selection techniques: PCA and gene ranking via limma-based differential expression.

Results

TCGA lung cancer database is used as a reference dataset for model training whereas test cohorts (five to ten samples) are randomly sampled from an independent public dataset. We repeat sampling and classification task for 1000 times and compare average accuracy on the test sets using different algorithms for classifications : Random Forest (RF), Naïve Bayes Classifier (NB), support vector Machine (SVM) and lasso logistic regression. ICA pre-processing step provides performance slightly better than performance when using limma gene selection —for all classifier except SVM— but PCA preprocessing fails on test dataset even though expected performances obtained by cross validation are high.

The HEK-SeqOme – A comparison of multiple RNA-Seq libraries

Arnaud Muller, Quantitative Biology Unit, Luxembourg Institute of Health, Strassen, Luxembourg; Nathalie Nicot; Elise Mommaerts; Francisco Azuaje; Gunnar Dittmar

Background

Sequencing technologies using *Sequencing By Synthesis* are becoming routine in multiple research projects leading to the expansion of the library preparation ecosystem. These biochemical technologies allows us to capture signal closer to the biological context, especially in the transcriptomic field.

Methods and results

Here we benchmark the results obtained from 3 different types of RNA library preparations; regular mRNA sequencing (TruSeq Stranded total mRNA), Low Input mRNA sequencing (Nugen Trio RNA-Seq) and 3' end mRNA sequencing (QuantSeq 3' mRNA seq with UMI). These libraries preparations have been applied to an identical RNA extract corresponding to HEK (Human Embryonic Kidney) cells that have been either treated with MG132 (carbobenzoxy-Leu-Leu-leucinal), a synthetic peptide inhibiting the proteasome complex responsible for the polyubiquitinated protein degradation, or non-treated.

The collected sequencing data are investigated at each step of the analysis starting from the *fastq* files through mapping, counting, gene expression significance, specific biotypes investigation and functional enrichment.

Because of the specificity of each library preparation, the initial results show discrepancy at the level of the read assignment and quantification over the genome. Moreover, these differences are challenging to mitigate even after expression significance analysis.

Nevertheless, by considering results at the level of gene sets or functions, the 3 methods converge to similar signals, including those linked to *protein degradation* among others.

Discussion

Our comparative analysis extends our understanding of a regular RNA-Seq experiment by confining over interpretation due to artefacts. Moreover, application of specific library methods could lead to an accurate evaluation of the global transcriptomic landscape in addition to the specificity of the considered protocol.

SilentMutations (SIM): a tool for analyzing long-range RNA-RNA interactions in viral genomes and structured RNAs

Daniel Desirò, Friedrich Schiller University, Jena, Germany; Martin Hölzer, Friedrich Schiller University, Jena, Germany; Bashar Ibrahim, Friedrich Schiller University, Jena, Germany; Manja Marz, Friedrich Schiller University, Jena, Germany

Long-range RNA-RNA interactions (LRIs) are already known to directly activate or inhibit viral replication and translation [1]. Technically, such interactions can be disrupted *in vitro* by the removal or mutation of interacting regions [2]. These modifications can result in secondary structure changes and finally manifest different viral titers. However, introducing only disruptive alteration will not exclude possible effects on other sequence parts.

To focus solely on the LRI, we can modify both interacting RNA segments simultaneously while maintaining synonymous codons. Additionally, we aim for the combination of both mutated segments to result in a similar interaction strengths in comparison to the interaction between the wild-type (WT) sequences. A single mutated sequence segment in combination with the respective WT sequence should destroy the interac-tion. Such a technique was recently used to verify a possible LRI between two in-fluenza A virus (IAV) segments [3].

Here, we developed SilentMutations (SIM), a tool that can simulate disruptive and compensatory mutants of two interacting single-stranded RNAs. This allows a fast and accurate assessment of key regions potentially involved in functional LRIs and will help RNA-experts to design appropriate experiments. We applied our tool on two experimentally validated IAV and hepatitis C virus interactions and we were able to predict potential double mutants for in vitro validation experiments. The tool is available at github.com/desiro/silentMutations.

[1] Nicholson BL, et al. (2014). Functional long-range RNA-RNA interactions in positive-strand RNA viruses. Nature reviews. Microbiology.

[2] Friebe P, et al. (2005). Kissing-loop interaction in the 3' end of the hepatitis C virus genome essential for RNA replication. Journal of virology.

[3] Gavazzi C, et al. (2013). A functional sequence-specific interaction between influenza A virus genomic RNA segments. Proc Natl Acad Sci USA.

Baltica: Benchmarking alternative splicing identification

Thiago Brito-Borges[1, 2], Volker Boehm[3], Jennifer V. Gerbracht[3], Niels H. Gehring[3], Christoph Dieterich[1,2]

[1] *Section of Bioinformatics and Systems Cardiology, Department of Internal Medicine III and Klaus Tschira Institute for Integrative Computational Cardiology, University of Heidelberg, 69120 Heidelberg, Germany*
[2] *DZHK (German Centre for Cardiovascular Research), Partner site Heidelberg/Mannheim, 69120 Heidelberg, Germany*
[3] *Institute for Genetics, University of Cologne, 50674 Cologne, Germany*

Background

Alternative Splicing (AS) event identification has helped to understand the molecular basis of human disease. Advances in RNA-Seq library preparation and sequencing technologies, notably the increase on the sequenced read-length, lead to a rapid increase in the number of observed exon-exon junctions, sequencing reads that span one or more exons. However, methods to identify and quantify alternative splicing produce results with low agreement.

Methods and results

To understand this disagreement, we reviewed several methods for AS event detection and developed Baltica, a framework to compute and analyze the results for multiple differential splicing methods, focusing on methods on from the latest generation. Baltica enables the interpretation of alternative splicing changes detected on RNA-Seq datasets. It combines exon-exon junctions' differential usage, derived from the intersection of the results of three independent tools, JunctionSeq, Majiq and Leafcutter, with transcript abundance estimates, obtained after de novo transcriptome assembly. Baltica uses Snakemake as a workflow manager, which facilitates reproducible and portable research. Baltica also produces reports from the analysis, that including information on the splice sites and the examination of the potential consequence to the gene function, such as losing a protein domain.

Discussion

We are benchmarking JunctionSeq, Majiq and LeafCutter using a transcript isoform base-truth dataset, the SIRVome from Lexogen, and aim to extend the framework with a third tool, Whippet. We plan to identify the origin of the referred disagreement amount the tools so we can extract the most biological relevant events from differential splicing methods.

Efficient algorithms for decoding whole-transcriptome MERFISH experiments

Till Hartmann, Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, University Hospital Essen, Essen, Germany; Sven Rahmann, Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, University Hospital Essen, Essen, Germany

With the Multiplexed Error Robust Fluorescence In Situ Hybridisation (MERFISH) protocol, thousands of single-cell transcriptomes can be analysed simultaneously, providing both RNA transcript counts and spatial locations.

However, the number of transcripts that can be targeted has been limited so far due to high observational error rates, conservative design decisions and lack of efficiently scaling decoding algorithms.

Background

In MERFISH, for each species of RNA (e.g. mRNA transcript under investigation), a binary barcode sequence and corresponding fluorescent probes are designed such that in the i-th round of smFISH imaging a fluorescent signal is emitted iff the barcode's i-th entry is 1 (and no signal if the entry is 0). By observing the bit pattern at each imaged pixel, it can be inferred which transcript is present at the corresponding location.

Due to background noise and biological as well as chemical issues, real-world data does not exhibit a one-to-one correspondence between designed barcode and observed bit pattern: Some true signals may not be recorded (false negatives), and similarly some spurious signals may be observed (false positives).

Experience shows that false negatives happen more frequently than false positives.

Method

Given observed transcript abundances y (as a result of a MERFISH experiment) and false negative/positive error rates e, “true” transcript abundances x can be calculated. Similarly, given observed abundances y and true abundances x, error rates e can be calculated. However, neither true abundances nor error rates are known beforehand, suggesting an iterative scheme which alternatingly estimates true abundances and error rates, depending on some initial guess for either of those.

In practice, we formulate the problem as a system of linear equations, so that true abundances can be calculated by successive-over-relaxation and error rates can be estimated by minimising the system in terms of its residuals via gradient descent.

Results

With the approach presented here, the number of targets can be increased by one to two orders of magnitude while achieving reasonable transcript quantification accuracy and lifting restrictions on protocol design.

While a mathematical sufficient condition guarantees that our approach works for low observational error rates (up to 4%), we empirically find that the approach also works for a larger range of error levels, depending on encoding scheme.

A Comprehensive Analysis of Antisense Transcription Across Tissue and Cancer Types

Julia Feichtinger, Division of Cell Biology, Histology and Embryology, Gottfried Schatz Research Center for Cell Signaling, Metabolism and Aging, Medical University of Graz, Graz, Austria; Ramsay J. McFarlane, North West Cancer Research Institute, School of Medical Sciences, Bangor University, Bangor, United Kingdom;

Background

Identification of cancer-restricted biomarkers is fundamental to the development of novel cancer therapies and diagnostic/prognostic stratification tools. Antisense transcription has not been extensively studied to date, although it is incredibly widespread and could exhibit high potential for such applications. The term antisense transcript usually refers to a non-coding RNA, which is at least partially complementary to a corresponding, mainly protein-coding transcript. These transcripts can have diverse functional roles in gene regulation, such as modifying epigenetic marks.

Methods and Results

By making use of the current wealth of transcriptomics data provided in the constantly growing public repositories, strand-specific RNA-seq data of a large panel of healthy controls as well as of selected cancer samples were curated, downloaded and aligned to the human genome using HISAT2. In addition to known non-coding RNAs, antisense transcripts were assembled using StringTie, and a comprehensive picture of antisense transcription across numerous normal tissues and selected cancer types was constructed.

Discussion

The importance of antisense regulation has been overlooked largely due to difficulties in identification, functional characterization and validation, leading currently to profound gaps in the analysis of these regulatory transcripts. Here, antisense transcription was investigated in a large-scale study, which can not only provide new insights into gene regulation but may also lead to novel diagnostic/prognostic marker and drug target candidates for clinical applications.

A comprehensive method protocol for annotation and integrated functional understanding of ncRNAs

Maximilian Fuchs, Functional Genomics and Systems Biology Group, Department of Bioinformatics, University of Würzburg, Germany

Meik Kunz, Chair of Medical Informatics, Friedrich-Alexander University of Erlangen-Nürnberg, Erlangen, Germany;

Thomas Dandekar, Functional Genomics and Systems Biology Group, Department of Bioinformatics, University of Würzburg, Germany

Abstract

Non-coding RNAs such as micro RNAs (miRNAs) and long non-coding RNAs (lncRNAs) have emerged as fundamental biological regulators. Recent studies reported that ncRNAs are involved in disease development and progression. Moreover, they appear as new promising non-invasive biomarkers for diagnosis and prognosis [1]. However, their functional role is often unclear or loosely defined as experimental characterization is challenging and bioinformatics methods are limited [2].

We developed a novel integrated method protocol for the annotation and detailed functional characterization of ncRNAs within the genome. It combines annotation, normalization and gene expression analysis with sequence-structure conservation, functional interactome and promoter analysis. Our protocol allows an analysis based on the tissue and biological context and is powerful in functional characterization of experimental and clinical RNA-Seq datasets including existing ncRNAs. This is demonstrated on several lncRNAs in cardiac disease (GATA6-AS1, Chast) and miRNAs in lung cancer [2-4]. In conclusion, our method protocol is powerful in functional characterization of RNA-seq data and opens new windows for an effective analysis of data in precision medicine.

References

1. Kunz, M., et al., *Non-Coding RNAs in Lung Cancer: Contribution of Bioinformatics Analysis to the Development of Non-Invasive Diagnostic Tools.* Genes (Basel), 2016. **8**(1).
2. Kunz, M., B. Wolf, and M. Fuchs, *A comprehensive method protocol for annotation and integrated functional understanding of lncRNAs.* Briefings in Bioinformatics, 2019.
3. Kunz, M., et al., *MicroRNA-21 versus microRNA-34: Lung cancer promoting and inhibitory microRNAs analysed in silico and in vitro and their clinical impact.* Tumour Biol, 2017. **39**(7): p. 1010428317706430.
4. Viereck, J., et al., *Long noncoding RNA Chast promotes cardiac remodeling.* Sci Transl Med, 2016. **8**(326): p. 326ra22.

Non-coding RNAs in podocyte disorders – taking glomerular disease beyond the coding genome

Talyan S*[1,2], Ignarski M[3], Müller RU[3], Dieterich C[1,2]

1 Section of Bioinformatics and Systems Cardiology, Department of Internal Medicine III and Klaus Tschira Institute for Integrative Computational Cardiology, University of Heidelberg, 69120 Heidelberg, Germany
2 DZHK (German Centre for Cardiovascular Research), Partner site Heidelberg/Mannheim, 69120 Heidelberg, Germany
3 Cologne Excellence Cluster on Cellular Stress Responses in Aging-associated Diseases (CECAD), University of Cologne, Cologne, Germany

Focal segmental glomerulosclerosis (FSGS) is a glomerular diseases caused by any infection, drug, or a disease like diabetes. Glomerular diseases are the most frequent cause of chronic kidney disease (CKD) or Kidney failure and contribute significantly to the rising prevalence of CKD. The glomerular filtration unit consists of three different cells – endothelium, mesenglial and podocytes. Podocytes enwrap the entire surface of the glomerular capillaries and leave the slit between them. This complex architecture of podocytes is important for maintaining the integrity of filtration unit any injury in podocytes leads to proteinuric kidney disease. However, this injury can be caused by many different diseases and deeper understanding of its basic pathophysiology is needed to investigate underlying alterations in podocytes biology. By using, high-throughput transcriptomic data of two well established FSGS mouse models, we study the role of long non-coding RNAs (lncRNAs). LncRNAs are RNA molecules known to be involved in mammalian disease and development. However, very little is known on the functional involvement of lncRNAs in glomerular diseases, including FSGS.

Here, we present a workflow for the identification of lncRNAs, although many lncRNA identification algorithm have been published which uses different criteria and cutoff for defining lncRNAs, but there is still scope for improvement. With our new pipeline, we identified 4178 known and novel lncRNAs which are expressed in podocytes based on FACS sorted podocytes dataset. Out of these 1057 are differentially regulated in either of two FSGS mouse model. Furthe, the relation of mouse FSGS lncRNA is also explored in human with respect to their conservation and expression. Our final set have 18 Podocytes specific lncRNA which are differentially regulated in both FSGS mouse model and also conserved and expressed in human kidney control datasets from GTEx and TCGA.

The lncRNAs identified by our approach will have to be further studied in detail using the current *state-of-the-art* RNA molecular biology to gain more knowledge about their function as well as knockout mouse models to confirm their pathophysiological role in FSGS.

Disentangling transcription factor binding site complexity

Ralf Eggeling, University of Tübingen, Germany

Background

The binding motifs of many transcription factors comprise a higher degree of complexity than a single position weight matrix model permits. Additional complexity is typically taken into account either as intra-motif dependencies via more sophisticated probabilistic models or as heterogeneities via multiple weight matrices. However, both orthogonal approaches have limitations when learning from in vivo data where binding sites of other factors in close proximity can interfere with motif discovery for the protein of interest.

Methods and results

In this work, we demonstrate how intra-motif complexity can, purely by analyzing the statistical properties of a given set of transcription factor binding sites, be distinguished from complexity arising from an intermix with motifs of co-binding transcription factors or other artifacts. In addition, we study the related question whether intra-motif complexity is represented more effectively by dependencies, heterogeneities, or variants in between. Benchmarks demonstrate the effectiveness of both methods for their respective tasks and application on motif discovery output from recent tools detects and corrects many undesirable artifacts.

Discussion

These results suggest on the one hand that the prevalence of intra-motif complexity may have been overestimated in previous studies that learn dependency models from in vivo data and should thus be reassessed. On the other hand, we also find evidence that the orthogonal approach of estimating an optimal number of PWM-based motifs from ChIP-seq data may underestimate intra-motif complexity.

Full paper: https://doi.org/10.1093/nar/gky683

Multi-Omics Factor Analysis (MOFA): an unsupervised statistical framework to disentangle patient heterogeneity in multi-omics studies

Ricard Argelaguet, European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom, presenting author; Britta Velten, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany; Damien Arnol, European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom; Sascha Dietrich, Heidelberg University Hospital, Heidelberg, Germany; Thorsten Zenz, Heidelberg University Hospital, Heidelberg, Germany; John C Marioni, Cancer Research UK Cambridge Institute, Cambridge, United Kingdom; Florian Buettner, Helmholtz Zentrum München-German Research Center for Environmental Health, Munich, Germany; Wolfgang Huber, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany; Oliver Stegle, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.

Multi-omics studies promise the improved characterization of biological processes across molecular layers. However, methods for the unsupervised integration of the resulting heterogeneous data sets are lacking. We present Multi-Omics Factor Analysis (MOFA), a computational method for discovering the principal sources of variation in multi-omics data sets. MOFA infers a set of (hidden) factors that capture biological and technical sources of variability. It disentangles axes of heterogeneity that are shared across multiple modalities and those specific to individual data modalities. The learnt factors enable a variety of downstream analyses, including identification of sample subgroups, data imputation and the detection of outlier samples. We applied MOFA to a cohort of 200 patient samples of chronic lymphocytic leukaemia, profiled for somatic mutations, RNA expression, DNA methylation and ex vivo drug responses. MOFA identified major dimensions of disease heterogeneity, including immunoglobulin heavy-chain variable region status, trisomy of chromosome 12 and previously underappreciated drivers, such as response to oxidative stress. Taken together, MOFA captured the key drivers of inter-patient variability, thereby enhancing data interpretation and facilitating the definition of predictive models for clinical outcomes.

Bioinformatics pipeline to screen metagenomic data for novel natural products and quantify biosynthesis potential

Shrikant Mantri, University of Tübingen, Tübingen, Germany
Helena Sales-Ortells, University of Tübingen, Tübingen, Germany
Timo Negri, University of Tübingen, Tübingen, Germany
Daniel Huson, University of Tübingen, Tübingen, Germany
Nadine Ziemert, University of Tübingen, Tübingen, Germany

Background:

With the increasing threat of antibiotic resistant pathogens, reemerging infective diseases and high cancer rates, there is an urgent need for new therapeutics. The majority of drugs has been, and continues to be, developed from chemical scaffolds produced by living organisms, so called natural products. A large portion of these natural products is produced as secondary metabolites by microbes. Next generation sequencing methods and the enormous amount of available DNA data has shifted drug discovery efforts from traditional bioactivity guided screening methods towards genome-based approaches. Genome mining, heterologous expression, and genetic engineering offer the unique opportunity to discover the huge untapped potential hidden in environmental data. Shotgun metagenomic DNA sequencing and meta-barcoding approaches have revealed the expansive biodiversity of bacteria and their secondary metabolites that have been missed by traditional culture-based drug discovery methods. However, the complex nature of metagenomic data and the highly repetitive structure of natural product biosynthetic pathways makes the analysis challenging.

Methods and Results:

In this project we have used the sequence tag-based domain profiling approach and developed a user-friendly bioinformatic pipeline that automatically detects and classifies natural product sequence tags in different environmental metagenomic datasets. It allows easy screening of the shotgun and amplicon metagenomic data for known and novel polyketides, RiPPs, terpenes and NRPs. This minimal pipeline and representative operational biosynthetic units database helps researchers to explore the biosynthetic potential of diverse ecosystem metagenomes.

Discussion:

Quick large-scale screening of diverse ecosystem metagenomes can be performed with our pipeline to discover novel natural product biosynthesis domains and quantify biosynthesis potential.

Statistical Problems in Gene-Based Testing and Possible Solutions

Ozan Cinar, Maastricht University, Maastricht, the Netherlands; Wolfgang Viechtbauer, Maastricht University, Maastricht, the Netherlands;

Background: Gene-based testing is an extension to genome-wide association studies (gwas) that examines the contribution of genetic variants in the etiology of a disease at the gene-level. By aggregating the signals from individual genetic markers, gene-based testing can explain the genetic architecture of complex diseases that cannot be examined with conventional gwas. Furthermore, gene-based testing brings several advantages to the analysis by reducing the number of Type I error and increased power of the study.

Methods and results: Gene-based testing yields a single p-value for each gene based on a combination of the p-values of the single-nucleotide polymorphisms (SNPs) that belong to those genes. There are a variety of well-known methods for combining p-values such as Fisher's method. However, these methods assume independence among the p-values which is violated in gwas due to linkage disequilibrium (LD), that is, non-random associations among the SNPs. Although the current methods can be modified to account for dependence with a variety of techniques such as effective number of tests or by calculating the test statistic under dependence, these modifications cover only a fraction of possible data dynamics. For example, they assume that the correlations are positive or they are applicable to only one-sided p-values. Furthermore, these modifications require the correlations among the p-values which are not known and instead need to be estimated based on the degree of LD among the SNPs. However, this estimation process has not been studied in the literature so far and requires further attention. Our studies show that the current modifications do not guarantee a nominal rejection rate. Moreover, estimating the correlations among the p-values is a complex process which depends on the type of the test and hypothesis yielding those p-values.

Discussion: In this talk, we discuss the current status of methods for combining p-values and describe their shortcomings by showing applications on real data sets. We

propose a number of possible solutions to these problems and discuss the improvements users can expect by incorporating these methods into their workflow.

The definition of open reading frame revisited

Patricia Sieber, Department of Bioinformatics, Friedrich Schiller University Jena, Jena, Germany and Research Group PiDOMICS, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute (HKI), Jena, Germany; Matthias Platzer, Leibniz Institute on Aging – Fritz Lipmann Institute (FLI), Jena, Germany; Stefan Schuster, Department of Bioinformatics, Friedrich Schiller University Jena, Jena, Germany

After genomic sequencing, an important step in deciphering information from sequences is the detection of open reading frames (ORFs). The concept of ORFs is of importance to identify potential protein-coding genes. The term is used frequently in biology and, in particular, bioinformatics.
However, in many textbooks, not much effort is spent on defining the term, or it is not perfectly clear-cut. Most commonly, an ORF is defined as a sequence stretch that is bounded by a start and stop codon and not interrupted by internal stop codons in the considered reading frame. Often, an ORF is considered equal to the corresponding coding sequence, and this definition already reaches its limits when introns are involved. Surprisingly, there is no unique agreed definition of that term and at least three definitions are in use with different specified boundaries. This demonstrates that it is worth questioning the established ORF definition. We present several molecular biological and bioinformatics aspects and discuss advantages and disadvantages of the different definitions. In the end, we recommend using the definition in which an ORF starts and ends with a stop codon each.

P. Sieber, M. Platzer and S. Schuster: The definition of open reading frame revisited. Trends in Genetics 2018 34(3):167-170

Optimizing panel sequencing estimates for tumor mutational burden (TMB) measurement

Jan Budczies, Eugen Rempel, Klaus Kluck, Daniel Kazdal, Volker Endris, Peter Schirmacher, Albrecht Stenzinger

Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany

Introduction: Tumor mutational burden (TMB), defined as the total number of missense mutations in a tumor tissue sample, is an emerging biomarker for cancer immune therapy. Due to high costs of whole exome sequencing (WES) and limited availability fresh-frozen tissues as well as paired germline DNA, TMB measurement is likely to be implemented in routine diagnostic by panel sequencing (PS) using panels typically covering 0.5-2 Mbp. So far, the accuracy PS TMB estimates in comparison to WES TMB has not been investigated comprehensively.

Methods and Results: Using a random mutation model and properties of the binomial distribution, we derived a mathematical formula for the coefficient of variation (CV) of PS TMB estimates: The CV of PS TMB turned out to be inversely proportional to the square root of the panel size and inversely proportional to the square root of the TMB level. This mathematical law could be validated by simulations of PS in 10,000 cancer exomes of the TCGA. In the real-word data simulations, we observed a slightly higher CV (about 15-30%) compared to the random mutation model, while the structure of the mathematical formula was preserved. Furthermore, we simulated PS in a recently published data set of WES data of immune checkpoint inhibition (ICI) treated patients (Miao et al., Nat Genet. 2018) with therapy response data available. The simulations were carried out for five of the largest currently commercial available tumor sequencing panels covering 0.24-1.34 Mbp of the coding region. Analyzing the response of lung cancer to ICI in ROC curve analyses, we observed a significantly or border-line significantly inferior performance of PS compared to WES for all five panels covering. Finally, we sequenced an in-house cohort of sixteen lung carcinoma with existing WES data using the TSO500 panel (Illumina Inc.). Typically (cohort median), the derivation of the lab PS TMB from the WES TMB was 2.7-times higher than the derivation of the simulated PS TMB from the WES TMB. This result demonstrates that panel size is an important contributor to the accuracy of PS TMB estimates.

Discussion: Combinatorial calculations and simulations of PS allow a transparent analysis of the accuracy of PS TMB estimates and contribute to the optimization of PS approaches for TMB measurement in clinical practice.

Galaxy for long-read ONT data analysis and public education

Bérénice Batut, Bioinformatics Lab, Albert-Ludwigs-Universität, Freiburg, Germany; Milad Miladi, Bioinformatics Lab, Albert-Ludwigs-Universität, Freiburg, Germany; Saskia Hiltemann, Erasmus MC, Rotterdam, Netherlands; Willem de Koning, Erasmus MC, Rotterdam, Netherlands; Street Science Community, Freiburg, Germany; Björn Grüning, Bioinformatics Lab, Albert-Ludwigs-Universität, Freiburg, Germany;

Thanks to the Oxford Nanopore Technologies (ONT), rapid long-read sequencing have been made accessible for a much broader range of applications and end-users. Two trends can be observed with this rapidly evolving third generation sequencing technology. On one hand, Nanopore sequencing is becoming more accessible, easier/simpler and yielding higher throughputs. On the other hand, related algorithms and downstream software pipelines are under constant development and improvement.

Bioinformatics analysis was already a bottleneck with the second generation of sequencing technologies, but even more so with long-read generation. Nanopore-based sequencing technologies are so much more accessible and rapidly produce so much more data that the data analysis challenges is getting fundamental. Community-driven solutions to democratize data analysis is crucial in the same way Oxford Nanopore is democratizing sequencing.

Galaxy has been shown to be a successful option for the 2nd-generation sequencing, but we think it has advantages that can shine even more in the era of long-read sequencing. Firstly, the user-friendly web interface allows it to be used with limited computational skills, making it ideally suited for educational projects. Secondly, the pipelines and software can be seamlessly upgraded at the server side while maintaining 100% reproducibility of the results. Thirdly, the backend infrastructure has support from personal computers to cluster grids and the cloud.

Within this project, we have integrated ONT-related tools deeply into Galaxy. We have developed a collection of the best practice workflows for genome assembly within Galaxy. Our work is available for everyone at the European Galaxy server (https://usegalaxy.eu) and supportive self-learning training material is available.

In this talk, we will present this effort and its application for an public education project: BeerDEcoded from the Street Science Community (https://streetscience.community). This voluntary-based non-profit group aims to bring science to the public. It organized workshops to teach pupils, students and citizens the fundamental concepts of molecular biology and genetics, data analysis, and open science by sequencing and analyzing the "DNA of beer" using the Nanopore MinION and Galaxy.

MiCroM: A Comprehensive Pipeline for Gene Amplicon and Metagenomic Data Analysis

Richa Bharti, TUM Campus Straubing for Biotechnology and Sustainability, Weihenstephan-Triesdorf University of Applied Sciences, Straubing, Germany;
Dominik G. Grimm, TUM Campus Straubing for Biotechnology and Sustainability, Weihenstephan-Triesdorf University of Applied Sciences, Straubing, Germany

Background

In diverse habitats, microorganisms act as an essential component in maintaining the earth's ecosystem. These microbes play a fundamental role in many processes (including recycling, breakdown, modulating immune responses and many more) and are widely studied using culture independent sequencing methods. Advances in sequencing technologies made it possible to study these complex roles of microbes, which led to the initiation of several important microbiome projects. For the analysis of this data a large variety of computational methods, statistical tools and visualizations have been developed. However, the plethora of these tools and the complexity of these analyses complicate the process of conducting such studies tremendously. Although, there exist certain guidelines for metagenomics and 16S rRNA analyses (e.g. Critical Assessment of Metagenome Interpretation (CAMI) and The Microbiome Quality Control project (MBQC)), there is no single tool that unifies all necessary steps in a common pipeline, while following these guidelines.

Methods and Results

We here present MiCroM, the first comprehensive and easy-to-use microbiome analysis pipeline that combines all steps in a single tool. For this purpose, MiCroM unifies a variety of state-of-the-art microbiome tools for amplicon, whole genome shotgun metagenomics and long read sequencing data in a common pipeline, following CAMI and MBQC standards. Further, MiCroM provides an integrated visualization module to automatically generate interactive visualizations and statistical analyses to help scientists with the interpretation of the results.

Discussion

Eventually, MiCroM will facilitate the analysis of microbial data and will help to increase the reproducibility of scientific results.

Gene Regulatory Networks Inference using Fuzzy Cognitive Maps by Improved Compressed Sensing Algorithm

Alireza Khanteymoori, Postdoc Researcher, Bioinformatics Lab, Albert-Ludwigs-Universität, Freiburg im Breisgau, Germany

Elpiniki Papageorgiou, Associate Professor, Faculty of Technology, University of Thessaly, Larisa, Greece

Jamshid Pirgazi, Ph.D., Computer Engineering Department, University of Zanjan, Zanjan, Iran

Nikolas Papandrianos, Assistant Professor, Nursing Department,University of Thessaly, Larisa, Greece

Seyyed Javad Mousavi, M.Sc., Computer Engineering Department, University of Zanjan, Zanjan, Iran

Abstract

Recently with the advancement of high-throughput sequencing, gene regulatory network inference became an interesting subject in bioinformatics and system biology. But there are many challenges in the field such as noisy data, uncertainty, time series data with numerous gene number and low data, time complexity and so on. In recent years, many researchers have been conducted to tackle those challenges which resulted in different methods in gene regulatory networks inference. A number of models have been used in modeling of the gene regulatory networks including Boolean networks, Bayesian networks, Markov model, relational networks, state space model, differential equations model, artificial neural networks and so on. In this paper, the fuzzy cognitive maps were used to model gene regulatory networks due to their dynamic and learning capabilities of handling non-linearity and inherent uncertainty.

Fuzzy cognitive maps belong to the family of recurrent networks and are well suited for gene regulatory networks. In this research study, the Kalman filtered compressed sensing is used to infer the fuzzy cognitive map for the gene regulatory networks. This approach using the advantages of compressed sensing and Kalman filters allows noise robustness and learning of sparse gene regulatory networks from data with high gene number and low samples. In proposed method, stream data and previous knowledge can be used in the inference process. Furthermore, compressed sensing finds likely edges and Kalman filter estimates the weights of them.

The proposed approach uses a novel method to decrease the noise of data. The proposed method was compared to CSFCM, LASSOFCM, KFRegular, ABC, RCGA, ICLA, and CMI2NI. The results show that the proposed approach is superior to the other approaches in fuzzy cognitive maps learning. This behavior is related to the stability against noise and creates a proper balance between data error and network structure.

Keywords: Gene Regulatory Network; Fuzzy Cognitive Maps; Compressed Sensing; Kalman Filter; Time Series Data

Predicting Gene Expression from Histone Modifications Signals using Machine Learning Techniques in Galaxy

Alireza Khanteymoori, Bioinformatics Lab, Albert-Ludwigs-Universität Freiburg im Breisgau, Germany; Björn Grüning, Bioinformatics Lab, Albert-Ludwigs-Universität Freiburg im Breisgau, Germany; Anup Kumar, Bioinformatics Lab, Albert-Ludwigs-Universität Freiburg im Breisgau, Germany;

Abstract

Epigenetic variations alter gene expression among various diseases and histone modifications play an important role in controlling gene expression. Predicting gene expression from histone modification signals can be useful for designing epigenetic drugs to combat diseases. Multiple computational models have been proposed to use histone modifications to predict gene expression. Machine learning based methods have shown success in modeling and understanding interactions among chromatin and they are highly desirable for modeling histone modification effects in gene regulation.

Galaxy is a scientific platform that helps scientists to analyze data without any programming skills. To analyse data, workflows can be created which make the analyses reproducible and shareable. We demonstrated machine learning approaches by reproducing and improving multiple publications by implementing Galaxy workflows to classify gene expression patterns using multiple datasets. The data includes five core histone modifications (H3K4me3, H3K4me1, H3K36me3, H3K9me3, H3K27me3). Hyperparameters in machine learning are important because they control the behavior of a machine learning algorithm and have a significant impact on its predictive performance. We have used two common strategies, grid search and random search in Galaxy and demonstrate which impact the correct hyperparameters have on the accuracy.

The training and test data have been obtained from the Roadmap Epigenomics Project (REMC) for different cell types. Using this data, we have applied a machine learning algorithm available in Galaxy to predict gene expression from histone modification signals and it is comparable with the other computational models.

ClinVAP: A Reporting strategy from variants to therapeutic options

Bilge Sürün[1,5], Charlotta P. I. Schärfe[1,5], Mathew R. Divine[1,5], Julian Heinrich[1,5], Nora C. Toussaint[2], Lukas Zimmermann[3,5], Janina Beha[4], and Oliver Kohlbacher[1,3,4,5,6,]

[1]Institute for Bioinformatics and Medical Informatics, Tübingen, Germany, [2]NEXUS Personalized Health Technologies, ETH Zurich & Swiss Institute of Bioinformatics, Zurich, Switzerland, [3]Translational Bioinformatics, University Hospital Tübingen, Tübingen, Germany, [4]Center for Personalized Medicine, University Hospital Tübingen, Tübingen, Germany; [5]Applied Bioinformatics, University of Tübingen, Germany; [6]Max Planck Institute for Developmental Biology, Tübingen, Germany

Background

The ever-decreasing cost of Next Generation Sequencing (NGS) has enabled clinical application of genomics in a personalized setting where optimal treatment plans are developed based on a patient's genetic profile. Especially in cancer, inferences from somatic variants from whole-exome sequencing data have become a key to understanding a patient's cancer profile for targeted treatment identification regardless of cancer type. The increasing volumes of data necessitate robust and automated pipelines to extract structured information from the detected variants. Although tools and databases exist that partially serve this purpose, they often have disadvantages such as complex command line interfaces, lack of integration as well as standardized input and structured output formats, or data privacy issues when web applications are used to process patient data.

Methods and Results

We introduce Clinical Variant Annotation Pipeline (ClinVAP) that processes, filters, and prioritizes variants in an automated manner. Based on the mutational landscape of the patient, it reveals the mechanisms driving the cancer and lists therapeutics shown to target the genes carrying somatic mutations. By integrating various publicly available databases, we created a comprehensive knowledge base that is used to conduct clinical annotation. ClinVAP is a multistep pipeline consisting of variant annotation, clinical annotation, and report generation steps. The pipeline takes somatic variants in Variant Call Format (VCF) file as input and generates clinical reports in both JSON and DOCX formats (the latter based on user-defined templates) and thus provides the results in both machine and human readable formats. It also

ensures reproducibility as it is fully containerized for Docker and Singularity platforms.

Discussion

We provide ClinVAP to create concise and structured clinical reports revealing the important factors in a patient's tumor profile, which is offered as evidence in targeted therapy selection. As a next step, we plan to conduct a utility assessment together with clinicians to prove ClinVAP's efficiency and contribution in the clinical decision-making process.

ClinVAP is available at https://github.com/PersonalizedOncology/ClinVAP

AMEN: a reproducible and reliable trimming, assembly and binning workflow.

Christian Knauth, Department of Biological Oceanography, Leibniz Institute of Baltic Sea Research, Rostock, DE; Johannes Werner, Department of Biological Oceanography, Leibniz Institute of Baltic Sea Research, Rostock, DE

Background

The analysis of microbial communities with the aid of next-generation sequencing allows a cultivation-independent approach to investigate microbial keyplayers in their community. To facilitate the analysis of uncultivable microbes on a genomic level, metagenome-assembled genomes (MAGs) can be inferred from metagenome data . In order to allow an automatic reconstruction of MAGs from metagenomics raw data, automatic workflows have been developed, such as MetaWRAP or SqueezeMeta. Both of them consist of a set of scripts and wrappers for established tools, which are run sequentially, however none of them is built to be error-tolerant.

Methods and results

With the ever-rising relevance of big data tasks in the entire economy, solutions for managing complex workflows and computational resources, such as Snakemake, Nextflow and Luigi have been developed and became a fundamental part of daily data-analysis business. These tools have a large user base, are stable, well-documented and last but not least scalable. Employing the pipeline tool Luigi, we present AMEN (Awesome Metagenomics), a workflow aiming to be user-friendly, scalable, reproducible, offering dependency isolation via conda, and integration with the rich Python scientific computing ecosystem. AMEN offers project-specfic configuration files, live status updates and per-step reports. Our tool is re-entrant and can be called directly from python code, making it very easy to use in conjunction with widespread platforms, such as Jupyter notebooks for instance. It also comes with a suite of tests, which lay a solid foundation for AMEN to be future-proof. Aside some fairly standard steps for trimming, assembly, binning and re-assembly that are also performed by other automatic workflows, AMEN also provides some enhanced mechanisms for more reliable results: We implemented an improved quality-filtering step, seeking out overrepresented sequences after quality trimming. A novel

abundance estimation for the binning step that takes multiple read mappings into account allows a more precise and thorough generation of MAGs.

Discussion

AMEN facilitates a highly reliable generation of MAGs from metagenomics raw data from any environment which allows reproducible and stable results due to its clear design and defined interfaces. With our design choices, we hope to attract contributors in order to establish a solid, yet versatile platform, ready for the future.

Comparing Alignment-Free Distances

Fabian Klötzl, MPI for Evolutionary Biology, Plön; Bernhard Haubold, MPI for Evolutionary Biology, Plön

Background

With the ever-increasing size of genomic datasets alignment-free comparison methods have become popular in recent years. Instead of producing sequence alignment, they give just a matrix of distances. These can be the result of k-mer counting or pairwise sequence comparison. Despite ongoing efforts to categorize the available tools, it remains unclear how to objectively chose the best tool. A common method to judge distances is to first reconstruct a phylogeny and then compare that to a reference using the Robinson-Foulds distance. Here we make a case for directly comparing the distance matrices.

Methods & Results

Fitch and Margoliash (1967) developed a criterion to pick the optimal tree for a given distance matrix. We propose to adapt this approach into a comparison of two distance matrices, *D* and *d*:

$$\Delta = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\left(D_{ij} - d_{ij}\right)^2}{\left(\left(D_{ij} + d_{ij}\right)/2\right)^2}$$

This approach allows the direct comparison of distance matrices without the intermediate step of a tree construction. Thereby it takes the implied branch lengths into account without being susceptible to artifacts from the tree construction.

Discussion

With the Robinson-Foulds distance informing about tree topology, the *Δ* statistic quantifies the difference between distance matrices.

Phybema: An extensible software pipeline for the systematic and reproducible evaluation of distance based phylogenetic reconstruction methods

Birgitta Päuker and Stefan Kurtz, Universität Hamburg, MIN-Fakultät, Center for Bioinformatics, Hamburg, Germany

Background

To estimate pairwise distances for the reconstruction of phylogenetic trees from sequence data one often uses alignment-free methods, as these are much faster than the traditional alignment-based methods. There are many software tools implementing such alignment-free methods, but reproducible benchmark studies evaluating such tools on real and simulated data are scarce.

Methods and Results

To allow for systematic and reproducible evaluations of phylogenetic reconstruction methods based on sequence data we have developed the software pipeline *Phybema* (**Phy**logenomic **Be**nch**Ma**rking). *Phybema* provides a simple interface to integrate distance estimators, i.e. software tools estimating pairwise distances from sequence data. The integrated distance estimators can be applied to diverse sequence datasets, provided with the pipeline. For each dataset the chosen tools will be called and from the resulting distance estimations the neighbor joining tree is constructed. These trees are visualized and compared against each other or against the reference tree provided with each sequence data set. The comparison delivers four different metrics to quantify the topological congruence of the resulting trees.

An exemplary run of *Phybema* with four distance estimators revealed that the best estimator on simulated data (with mutations and lateral gene transfer as evolutionary events) is not necessarily the best on real data and vice versa.

Discussion

Phybema provides a software infrastructure including data sets for benchmarking distance estimators in the context of reconstructing phylogenetic trees. The software has only a few dependencies, which are satisfied by a standard anaconda3 installation of Python3. The extensible and simple interface eases the integration of new estimators and their comparison to existing results.

Application of stepwise Elastic Net regression (sENR) with pairwise interactions to improve interpretation of drug response prediction with multiple data types

Nan Li, German Cancer Research Center and Heidelberg University, Heidelberg, Germany; Roman Schefzik, German Cancer Research Center, Heidelberg, Germany; Angela Goncalves, German Cancer Research Center (DKFZ), Heidelberg, Germany

With the fast development of sequencing technology, it has become possible to carry out large-scale genomic screens for cell lines, resulting in large datasets with associated analytical challenges.
Regarding this situation, Elastic Net regression (ENR), which is a linear method with both L1 and L2 penalty, has been widely used by scientists for predictive or interpretive analysis, due to its ability to do shrinkage and smoothing when the number of features (e.g. genes) is much bigger than the number of samples.

One of the most prominent applications of ENR prediction from multiple omics data is in cancer drug screen projects, with CCLE and GDSC being the most famous ones, where it indicates that most of contribution to drug response comes from gene expression, regardless of other omics data.
In order to gain insight into the other data types available, and to improve biological interpretability, we have developed a stepwise ENR approach (sENR) with pairwise interactions, , which can measure the contribution of genomic upstream data (e.g. type of tissue, mutation) and genomic downstream data (e.g. expression) in a more balanced way than classic ENR, and can check the importance of interactions between upstream and downstream data and even among different upstream data types.

sENR, together with other shrinkage methods, were tested in CCLE data set. Results reveal that, for some drugs, all linear regression methods fail. Meanwhile, for the set of drugs for which there is some predictive power, sENR tends to find more genomic upstream data than other existing methods, which could aid in biological interpretation.

Surveying the Local Sequence Complexity in the Axolotl Genome

Bernhard Haubold, Max-Planck-Institute for Evolutionary Biology, Plön/Germany;

Vertebrate genomes are full of transposons. These are not distributed uniformly across the genome; instead, clusters of transcriptional regulators, like the *Hox* clusters, contain comparatively frew transposons. The reason for this is presumably that most transposon instertions in such regions are highly deleterious. This suggests that vertebrate genomes can be read as the result of a transposon mutagenesis experiment carried out over evolutionary time. Highly conserved regions would then lack close homologues elsewhere in the genome.

To quickly quantify homology, Pirogov et al. (2018) developed and implemented the *match complexity*, which ranges from zero in regions repeated exactly elsewhere, to 1 in regions without close homoloues in the rest of the genome. As hypothesized, they found that in the human and mouse genomes regions with maximal match complexity were highly enriched for developmental genes.

Nowoshilow et al. (2018) sequenced the axolotl genome, which is ten times larger than the human genome. The authors noted that the axolotl lacked the developmental gene *Pax3*. In this poster I revisit this result by investigating the gene complement contained in the high-complexity moiety of the axolotl genome.

Omics Profiling and integration: Deeper insight into Pancreatic Ductal Adenocarcinoma (PDAC) survival heterogeneity using integrative analyses of individual and global transcriptome based networks

Archana Bhardwaj[a], Claire Josse[b,f], Daniel Van Daele[c], Marcela Chavez[d], Ingrid Struman[e], Kristel Van Steen[a]

[a]BIO3, GIGA Research, University of Liège, Belgium

[b]Laboratory of Human Genetics, GIGA Research, University Hospital (CHU), Liège, Belgium

[c]Department of Gastro-enterology, University Hospital (CHU), Liège, Belgium

[d]Department of Medicine, Division of Hematology, University Hospital (CHU), Liège, Belgium

[e]GIGA-R Centre, Laboratory of Molecular Angiogenesis, Liège, Belgium

[f]Medical Oncology Department, CHU Liège, Liège Belgium

Background: Pancreatic ductal adenocarcinoma (PDAC) is categorized as the seventh leading cause of cancer mortality in the world. PDAC survival rate is 5%, but a very small subset of patients survives longer. Its highly significant and challenging task to identify the factors that determine the long-term survivorship. To clarify the discrete roles of heterogeneity between individual profile of PDAC survival patients, ideally individual (patient level) and group level (comparative level) data could be integrated. Hence, we aimed to propose and implement a combinatorial genomics approach for the omics profiling of ST (short term) and LT (long term) PDAC cohort that exhibit variation at molecular profiling representative of one of the clinical stage of this disease.

Method and Results: Our study is the first to perform extensive integrative individual and group based transcriptome profiling in PDAC patients contrasting LT (> 36 months) and ST (< 12 months) survival. Using a discovery cohort of 19 PDAC patients from CHU-Liège (Belgium), we first identified differentially expressed genes (DEGs) between LT/ST. Second, we performed unsupervised system biology approaches to obtain

meaningful gene modules. In particular, important modules obtained via weighted gene co-expression network analysis (WGCNA) showed significant correlation with clinical features, including tumor size and interval between surgery and chemotherapy. Next, we created individual-level perturbation profiles from normalized gene expression data. Detailed inspection of individual specific omics changes across LT survival individuals revealed biological signatures associated to focal adhesion and Extracellular matrix (ECM) receptors evident to PDAC survival. Finally, we prioritized cancer genes by integrating group based and individual-specific based DEGs on a directed functional interaction network. In parallel, analyzed exome sequencing data of same PDAC cohort to identify the survival associated SNPs and implementing various machine learning approaches for the integration of identified omics profiles of transcriptome and mutational (SNPs) profiling of PDAC patients (work in progress).

Discussion: We demonstrated first time in PDAC, integrative approach where individual based profiling could help to rank cancer specific genes together with the group based gene expression analysis. Coupling of multiple omics profiling might be always beneficial. We were able to find out the factors linked to basal cause of survival status in two different groups of survival groups, which could lead to improvement in gene prioritization and thus effect personalized target selection.

Phase-aware neoantigen prediction from Whole Exome and RNA Sequencing Data

Jan Forster, German Cancer Consortium, Partner Site Essen/Düsseldorf, Essen/Germany

Johannes Köster, Algorithms for reproducible bioinformatics, Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, University Hospital Essen, Essen/Germany

Background

Cancer immunotherapy is a current focus topic in oncology. In order to train the immune system of a patient to detect and destroy cancer cells, it is important to discover novel cancer-specific peptides which function as antigens towards the tumor. The effectivity of this approach depends on a variety of different factors.

Neoantigen candidates need to derive from proteins which are actually expressed in the tumor, form a stable complex with the patient's MHC alleles and be recognizable by T-cells . Since it is not yet possible to confidently predict the immunogenicity of a peptide, several other predictors such as MHC binding affinity or similarity to non-somatic peptides are used in the process. Typically, when testing in-silico predicted neoantigen in-vitro, only a smaller percentage of peptides will be tested positive in a T-cell assay. To get optimal results during neoantigen prediction, it is therefore important to correctly define the search space (neopeptidome) of a patient and then narrow down the candidates as accurately as possible.

Methods and Results

The presented pipeline tries to achieve maximal accuracy in neoantigen prediction, starting directly on raw sequencing data. To correctly assess the entirety of cancer-specific neopeptides, we use small-scale phasing in order to determine all clonal and subclonal haplotypes of adjacent variants. To our knowledge, this step is not satisfactorily incorporated in most established pipelines, at least not on DNA sequencing data. Phasing is performed using our recently developed tool Microphaser (https://github.com/koesterlab/microphaser). Given a GTF annotation file, Microphaser uses a sliding window approach over all annotated transcripts and returns phased neopeptides of a predefined length as well as their corresponding non-somatic normal peptides. Currently, we are able to handle SNVs as well as small insertions and deletions (inframe and frameshift). The resulting neoantigen

candidates are translated into amino acid sequences and compared against the reference proteome of the patient to filter self-similar neopeptides. In downstream analysis, the MHC alleles and their binding affinity towards the neopeptides are predicted using well-established tools such as NetMHCpan, Optitype and HLA-LA. Transcript expression is assessed from RNA-seq data using the quasi-mapping approach of kallisto. The workflow itself is written as a Snakemake pipeline (https://github.com/jafors/Neoantigen_Prediction).

The pipeline has been used on various datasets, including published metastatic melanoma samples in which we were able to replicate all sucessfully validated neoantigens and other impact variables such as LOH in several MHC alleles. Validation of other predicted neoantigen candidates is currently in progress. A concordance analysis performed on for different sequencing protocols of the same sample also showed that the pipeline is robust to technical variance.

Discussion

We present a fully comprehensive workflow for neoantigen prediction from NGS data, which in most cases yields similar results as other established workflows. The benefits of the approach lie in the comprehensive analysis which creates results directly from FASTQ data as well as the improved resolving of adjacent variants in neopeptides due to phasing. We argue that this more accurate representation of the neopeptidome might improve the outcome of neoantigen prediction especially in tumors with a high mutational load.

DeePaC: Predicting pathogenic potential of novel DNA with reverse-complement neural networks

Jakub M Bartoszewicz, Robert Koch Institute, Berlin Germany and Free University of Berlin, Berlin, Germany; Anja Seidel, Robert Koch Institute, Berlin Germany and Free University of Berlin, Berlin, Germany; Robert Rentzsch, Robert Koch Institute, Berlin Germany; Bernhard Y Renard, Robert Koch Institute, Berlin Germany

We expect novel pathogens to arise due to their fast-paced evolution, and new species to be discovered thanks to advances in DNA sequencing and metagenomics. Moreover, recent developments in synthetic biology raise concerns that some strains of bacteria could be modified for malicious purposes. Traditional approaches to open-view pathogen detection depend on databases of known organisms, which limits their performance on unknown, unrecognized, and unmapped sequences. In contrast, machine learning methods can infer pathogenic phenotypes from single NGS reads, even though the biological context is unavailable.

We present DeePaC, a Deep Learning Approach to Pathogenicity Classification. It includes a flexible framework allowing easy evaluation of neural architectures with reverse-complement parameter sharing. We show that CNNs and LSTMs outperform the state-of-the-art based on both sequence homology and machine learning, cutting the error rate almost in half when predictions for both mates in a read pair are integrated. To provide a degree of interpretability of a learned model, we identify and visualize sequence motifs contributing to the final classification decision. Investigating the relation between particular subsequences and class membership leads to identification of prospective markers of pathogenicity in a human host.

A Novel Tool for Quantitative Measures of Genome-wide Cell-free DNA Fragmentation in Cancer Patients

Pitithat Puranachot, Applied Bioinformatics, German Cancer Research Center, Heidelberg, Germany; Steffen Dietz, Cancer Genome Research, German Cancer Research Center, Heidelberg, Germany; Holger Sültmann, Cancer Genome Research, German Cancer Research Center, Heidelberg, Germany; Benedikt Brors, Applied Bioinformatics, German Cancer Research Center, Heidelberg, Germany

Contact information: p.puranachot@dkfz-heidelberg.de, s.dietz@dkfz-heidelberg.de, h.sueltmann@dkfz-heidelberg.de, b.brors@dkfz-heidelberg.de

Background: As an alternative approach for cancer patient management to monitor therapeutic response, tumor-derived cell-free DNA (cfDNA) or circulating tumor DNA (ctDNA) can be detected in the plasma of patients. However, larger quantities of cfDNA from noncancerous cells dilute the overall concentration of ctDNA. Recent tools are mostly capable of detecting DNA mutations from tumor tissue while inferior to detection of low tumor DNA contribution. Developing a tool that consider short DNA fragmentation which differentiate ctDNA from non-tumor cfDNA will improve detection sensitivity.

Methods and Results: We are developing a novel method to evaluate genome-wide short DNA fragmentation from low-coverage whole-genome sequencing data of cfDNA. A fraction of short DNA fragment, having a length of 150 bases or less, were calculated for each 100kb genomic non-overlapping window. A z-score statistic was used to determine if the short DNA fragment in a window would be significantly higher or lower compared to a group of healthy individuals. As a result, our method reports genomic regions with deviant fragmentation pattern which could infer copy-number alteration in tumor cells. Quantitative measures of fragmentation deviation, referred to as genome-wide z-scores and median absolute deviation (MAD scores), were significantly higher in patients with chronic lymphocytic leukemia and non-small cell lung cancer as compared to healthy individuals.

Discussion: cfDNA from cancer patients with high fragmentation deviation (e.g. genome-wide z-scores) reflects higher contribution of ctDNA. Our approach genome-widely measures in cfDNA the aberrant short DNA contribution which infer copy-number alteration in tumor cells. This measurement would also reflects accordingly the therapeutic outcome when serial plasma samples were taken throughout the treatment course.

The Statistical Stability of Consensus Independent Component Analysis for Patient Classification and Prediction of Survival

Maryna Chepeleva[1,2], *Thomas Eveno*[1], *Mikalai M Yatskou*[2], *Petr V Nazarov*[1]

[1] *Luxembourg Institute of Health, Strassen, Luxembourg;*

[2] *Belarusian State University, Minsk, Belarus*

Background

Independent component analysis (ICA) allows decomposing heterogeneous transcriptomics data and extracting relevant transcriptional signals that correspond either to relevant biological processes or to technical biases. Using independent components as features for downstream analysis requires high reproducibility of decomposition. Here we investigated the stability of ICA and tested reproducibility of its results for single and multiple runs, and in the case of reduced number of samples.

Method

We applied the developed consensus ICA algorithm (https://gitlab.com/biomodlih/consica) to TCGA RNA-seq gene expression data on patients with skin cutaneous melanoma (SKCM) and non-small cell lung cancers: squamous cell carcinoma (LUSC) and adenocarcinoma (LUAD). The stability of the deconvolution was investigated in respect to the number of parallel consensus ICA runs and the size of the patient cohort. Two predicting models were used to classify the patients based on ICA results: random forest and xgboost from corresponding R packages. We also carried out survival prediction with Cox regression (R package `survival`) on the weights of independent components and examined the dependence between prediction quality and the number of components. Jaccard indexes, coefficients of determination and cosine similarities between identified metagenes were used to assess the stability of deconvolution.

Results

We validated the optimal number of parallel consensus ICA runs that provided reproducible deconvolution for each cancer type. The dependency on real datasets was contrasted with such on permutated datasets. Also, we estimated the effect of parallel runs on the quality of lung cancer type classification (LUSC/LUAD). Using ICA as a feature engineering method before cancer type classification gave a higher accuracy compared to classifiers trained on the raw gene expression data. We identified biologically relevant signals that were stably detected for skin and lung tumours, including cell cycle, activity of infiltrating immune cells and keratinization. Finally, we estimated the boundary values for

the number of components that allow detecting these biologically relevant signals in smaller patient cohorts.

Finding sequence motifs in ChIP-Seq data without peaks

Michael Menzel, Technische Hochschule Mittelhessen, Gießen, Germany; Andreas Gogol-Döring, Technische Hochschule Mittelhessen, Gießen, Germany

Background

The default way to identify sequence motifs from chromatin immunoprecipitation sequencing (ChIP-Seq) data utilizes peak calling with tools like MACS2 [1] and motif discovery around those peaks with Homer [2] or similar tools. However, this identification relies on sufficient data quality to identify peaks. If none or only few peaks can be identified the motif discovery is unusable. [3]

Method and results

Here we propose a novel method to analyze ChIP-Seq reads. Using enrichment profiles of K-mers around each read we are able to identify binding motifs without peak calling and analyzing data-sets that cannot be utilized in traditional tools.

Reads from ChIP-Seq data on binding sites were removed around peaks in several steps until no peaks could be identify. For each step we analyzed the remaining reads with our method, the remaining peaks with Homer and scored the motifs against a reference. The results show that our method is only slightly affected by the decreasing read count while the quality of the results by Homer decrease with each step until there are no peaks left and therefore peak calling is unable to search for sequence motifs.

Discussion

Based on our results we assume that our method can be used to analyze data-sets that are not suitable for common motif discovery pipelines. Additionally, enrichment profiles can be used to identify binding site characteristics like dimer binding.

[1] Zhang, Yong, et al. "Model-based analysis of ChIP-Seq (MACS)." *Genome biology* 9.9 (2008): R137. [2] Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589. [3] Nakato, Ryuichiro, and Katsuhiko Shirahige. "Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation." *Briefings in bioinformatics* 18.2 (2016): 279-290.

Determinants of transcription factor activity - resolving the riddle of absent expression correlation of transcription factors and their targets

Adam Zaborowski; Dirk Walther, Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

Transcription factors are one of the major players in the regulation of gene expression. However, contrary to expectations, the average expression correlation coefficient between transcription factors (TFs) and their targets is hardly different from that of transcription factors and non-target genes. Nonetheless, the overall distribution may hide the presence of a subset of genes, which follows a simple regulatory model where expression of a gene can be predicted by the expression of the cognate transcription factor. For *Arabidopsis thaliana*, we created a set of regulatory pairs based on available DAP-seq experimental data and 150 RNA-seq expression data. For each pair, a multitude of genomic features, along with data from high throughput experiments were collected. With machine learning algorithms we identified those features, which are associated with highly correlated pairs. Chosen features concern physical properties of TFs such as length of the protein and 5' UTR, expression variation of gene and TFs, along with other characteristics. Using this information, one can predict, which pairs could be correlated in terms of expression based on genomic features and use this information to more reliably reconstruct gene regulatory networks.

Identification of novel intronic cis-regulatory motifs in *Arabidopsis thaliana*

Georg Back; Dirk Walther, Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

Introns have been known to influence gene expression and being able to promote transcription, even in the absence of promotors, referred to as intron mediated enhancement (IME), the mechanistic details of which are not yet fully understood. We hypothesize that the mechanisms at play will involve the specific recognition of regulatory sequence motifs inside introns by DNA-binding proteins or other molecules (e.g. RNA). Assuming that regulatory motifs are subject to evolutionary selection, we exploit the available deep genome sequence information for the plant species *Arabidopsis thaliana* to identify conserved intronic sequence motifs and associate them with the observed gene expression patterns. In addition, and as the first intron is believed to be the mediator of IME, we contrasted presence/ absence occurrences of motifs in first versus the remaining downstream introns. As a result, candidate motifs have been identified, whose potential to act cis-regulatorily has been confirmed based on co-expression analysis. Our findings suggest the existence of specific intronic motifs involved in gene-expression regulation.

Signature refitting with decompTumor2Sig: identification of active mutational processes in individual tumor samples

Sandra Krüger, Institute of Computer Science, Freie Universität Berlin, Germany;
Rosario M. Piro, Insitute of Computer Science, Freie Universität Berlin, Germany, Institute of Medical Genetics and Human Genetics, Charité-Universitätsmedizin, Berlin, Germany, German Cancer Consortium (DKTK) partner site Berlin and German Cancer Research Center (DKFZ), Heidelberg, Germany

The somatic mutations of a tumor often stem from multiple mutational processes such as cigarette smoke, UV light, or age-related spontaneous deamination of 5-methylcytosine. The identification of the processes which have contributed to a tumor genome is a current research question in cancer genomics and may be of clinical relevance, e.g., when it indicates a DNA repair deficiency which may impact the response to cytotoxic treatments.

Two different mathematical models of mutational processes have been developed recently, one developed by Alexandrov et al[1] and one by Shiraishi et al[2], describing them as "mutational signatures" in terms of mutation frequencies of base changes within their immediate sequence context. These models differ with respect to the number of required parameters and their accuracy of describing mutational events.

The de novo inference of mutational signatures requires a large number of tumor genomes[1,2] to decompose them into (i) a set of signatures that represent the processes driving the somatic mutations, and (ii) a set of "exposures" or contributions of the signatures to the individual tumor genomes. In a clinical setting the necessity for large set of tumors is unsatisfactory and often impractical. However, once accurate signatures have been defined, they can be used to estimate their contributions to the overall mutational load of an individual tumor sample, a process called “signature refitting”.

We present the Bioconductor R package "decompTumor2Sig" that we explicitly developed for dissecting tumor genomes into a given set of mutational signatures of both the Alexandrov- and the Shiraishi-type model. So far such tools existed only for Alexandrov signatures. We show the application of our tool to examples of both signature types.

References:

[1] Alexandrov et al: Signatures of mutational processes in human cancer. Nature 500, 415-421 (2013).

[2] Shiraishi et al: A simple model-based approach to inferring and visualizing cancer mutation signatures. PLoS Genet 11(12), 1005657 (2015).

Drug response prediction in primary blood tumor cells using models trained on cell line data

Roman Kurilov, German Cancer Research Center (DKFZ), Heidelberg, Germany;
Benedikt Brors, German Cancer Research Center (DKFZ), Heidelberg, Germany

Data from a number of large high-throughput cell line screens have became available to scientific community over the last decade. By combining molecular and drug response data from such screens it's possible to identify biomarkers of response and build predictive response models. However, accuracy of such prediction still remains limited, especially in the case when we would like to make predictions for patients using models trained on cell line data. In this work we investigated various aspects of model training in order to identify those aspects that affect the accuracy of drug response prediction.

Particularly we tested our ability to predict drug response in primary blood tumor cells (data from the study Dietrich et al.[1]) using models built on cell line data (GDSC dataset[2]). We built models for the drug panel consisted of 21 drugs covering the wide range of molecular targets. As a base model for each drug we used a model with only expression features and we examined how addition of other data types and sample subsetting influence the accuracy. We found that addition of mutation and methylation information, and selecting only blood cancer cell lines for training have a strong impact on a model's accuracy. Interestingly depending on drug this impact can be either positive or negative.

References:

1. Dietrich S. et al. "Drug-perturbation-based stratification of blood cancer." *The Journal of clinical investigation* 128.1 (2018): 427-445.

2. Iorio F. et al. "A landscape of pharmacogenomic interactions in cancer." *Cell* 166.3 (2016): 740-754.

Methylomes and transcriptomes of the adult neural stem cell lineage suggest a role of DNA-methylation in cell type-specific gene expression

Lukas P. M. Kremer, DKFZ & ZMBH, Heidelberg; Sascha Dehler, DKFZ, Heidelberg; Santiago Cerrizuela, DKFZ, Heidelberg; Dieter Weichenhan, DKFZ, Heidelberg; Ana Martin-Villalba, DKFZ, Heidelberg; Simon Anders, ZMBH, Heidelberg

The subventricular zone (SVZ) is one of the two major neurogenic niches of the adult mammalian brain. It harbors a lineage consisting of adult neural stem cells (NSCs) and their progeny, including neuroblasts and glial cells. DNA-methylation is essential for embryonic NSC function, but due to the scarcity of adult NSCs, the role of DNA-methylation in adult brains has remained elusive. Here, we employ T-WGBS, a low-input bisulfite-sequencing protocol, and RNA-seq to assess the methylome and transcriptome of the adult NSC lineage within the natural environment of the brain.

We compare the methylomes of NSCs, astrocytes, oligodendrocytes and neuroblasts and find that astrocytes and neuroblasts share a surprisingly similar methylome, despite divergent transcriptomes. Overall, we identify ~45,000 differentially methylated regions (DMRs) in the lineage. Motif analysis identifies several transcription factors as candidates affecting the methylome, some of which are known to regulate NSC differentiation. We use k-means to identify three main types of DMRs, including a rare set of regions that is demethylated in neuroblasts only. Intersection with the Ensembl regulatory build and gene ontology enrichment shows that DMRs are enriched for enhancers and promoter flanks and located near cell type-specific genes, suggesting a role in gene regulation. Indeed, integrating our methylome and transcriptome data shows that promoter demethylation in the lineage is associated with increased gene expression and *vice versa*. Surprisingly, this relationship breaks down in neuroblasts, suggesting that another regulatory mechanism may act in neurogenesis specifically.

Finally, we show that knockout of two interferon receptors has drastic effects on the methylome of NSCs, but not on the other cell types. Specifically, we show that the methylation changes induced by interferon receptor knockout fundamentally resemble those previously observed in wild type NSC-to-astrocyte differentiation. Our study contributes to unraveling the control of the methylome over the transcriptome and sheds new light on the role of interferon in regulation of the epigenome.

Characterizing Ambiguity in Cancer Evolutionary Histories

Linda K. Sundermann, The Donnelly Centre, University of Toronto, The Vector Institute, Toronto, Canada; Jeff Wintersinger, The Donnelly Centre, University of Toronto, The Vector Institute, Toronto, Canada; Gunnar Rätsch, ETH Zurich, Zurich, Switzerland; Jens Stoye, Bielefeld University, Bielefeld, Germany; Quaid Morris, The Donnelly Centre, University of Toronto, The Vector Institute, Toronto, Canada

Cancer within an individual is not a homogeneous disease. Instead, a patient's cancer consists of multiple distinct evolutionary lineages, each of which harbors unique genomic mutations. Using genomic sequencing data, we can reconstruct the evolutionary history of these lineages, granting insights into how cancer develops and potentially improving treatment.

Inferring the evolutionary history of a patient’s cancer is complicated by ambiguity in the data, meaning that multiple solutions may be equally plausible under a certain model. Understanding this uncertainty is critical when seeking biological and clinical insights.

Methods to characterize this ambiguity are an active area of research. Existing approaches, however, can only summarize the set of solutions provided by the user. Possible solutions not represented in this set will remain undiscovered, meaning that the user will not gain a full understanding of possible evolutionary histories.

Here, we present the first algorithm that represents the complete set of possible evolutionary histories encoded in a single given solution, subject to some constraints. Our algorithm produces a data structure that efficiently identifies which lineage relationships in that solution are certain and which are ambiguous. Using this information, we can enumerate the full set of consistent solutions.

As our algorithm can fully characterize lineage relationship ambiguity in evolutionary history reconstructions, we can now use simulated data to systematically explore the extent to which richer datasets reduce this ambiguity. Initial results demonstrated that adding more tissue samples from the same cancer substantially reduce uncertainty.

Structure-Informed Variant Prioritization in Personalized Oncology

Siao-Han Wong, Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany; , Benedikt Brors, Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg

In personalized oncology, treatment strategies take individualized genomic aberrations into consideration. The interpretation of their pathogenicity relies on our understanding of regional significance within genes. There had been initiatives identifying mutation hotspots based on protein sequences or protein structures, where residues distant in sequence can be brought together through protein folding. Nonetheless, these information are not well utilized in the clinical setting.

In this study, we compiled prior knowledge as a hotspot list to molecularly stratify patients, and further consider the structural neighborhood to gain insights into underlying mechanisms. Specifically, 1,572 residue hotspots are predicted from three sequence approaches and an additional 174 structural hotspots from three structural algorithms. With setting up a mechanism to map genomic variants to protein structures, we were able to annotate the structural neighborhood of variants and investigate the interplay between them. Focusing on protein-protein interaction interfaces, we have identified interfaces preferentially disrupted and also the exact contributing mutations.

In summary, this approach enables us to quickly prioritize variants in patient genomes by their prevalence, and further annotate structural significance to inform potential mechanisms behind driving events. This would help facilitate clinical evaluations to meet the growing number of patients recruited to personalized oncology trials.

Target-driven gene expression signatures and pre-clinical drug sensitivity models for predicting patient response to chemotherapy

Tony Kaoma, Quantitative Biology Unit, LIH, Strassen, Luxembourg; Jean-Emmanuel Sarry, CRCT, Toulouse, France; Francisco Azuaje, Quantitative Biology Unit, LIH, Strassen, Luxembourg

Background

Although 60-80% of acute myeloid leukemia (AML) patients respond to the standard chemotherapy (Cytarabine - araC plus an anthracycline), 65% relapse within five years. We previously identified a plasma-membrane enzyme whose gene expression is strongly correlated with resistance to araC *in vitro* and *in vivo*. Based on its potential role as a therapeutic target, we hypothesized that this gene may provide the basis for building relevant models to predict AML patient drug sensitivity (DS) to araC.

Methods and results

First, we generated 8 lists of genes by analyzing the transcriptomic changes resulting from the knock-out of our gene candidate in an AML cancer cell line (CCL) resistant to araC. We then applied elastic net regression to test the ability of each gene list to predict sensitivity to araC using their gene expression profiles (GEP) in 846 different CCLs and their *in vitro* DS to araC. As a reference, we also considered all genes. The CCL's GEP and DS were obtained from the GDSC database. Lastly, we validated the clinical relevance of our models with data from a patient-derived-xenograft preclinical (PDX) model of resistance to araC, primary cells from AML patients and the TCGA.

The majority of our models achieved good performance on CCL data with one model (M3) ranked at the top (concordance index = 0.7). On PDX data, 7 models out of 9 correctly separated sensitive samples from resistant; M3 gave the largest boundary between the two groups. On primary cells, only one model (M8) reached an AUC = 0.7. Focusing on M3 and M8, we correlated the predicted DS with overall survival using TCGA data. For both models, most sensitive patients tend to live longer. Predicted DS was positively correlated with over-expression of gene sets involved in resistance to araC.

Conclusion

By restricting our DS modeling to genes linked to a candidate therapeutic target, we generated and validated clinically-relevant models for accurately predicting AML patient response to araC, which offers new patient-oriented research opportunities.

Improving Characterisation of Cancer Patients by Consensus Independent Component Analysis of Tumour Transcriptomes

Petr V Nazarov, Thomas Eveno, Maryna Chepeleva, Tony Kaoma, Arnaud Muller, Francisco Azuaje, Luxembourg Institute of Health, Strassen, Luxembourg

Background

The majority of tumour samples collected from patients and studied by high-throughput transcriptomics are heterogeneous at three levels. First, bulk samples contain a mixture of several cell types, with proportions that are difficult to control. Second, cancers naturally develop inter and intra-tumour heterogeneity of malignant cells. Third, the evolving technology may introduce technical biases and limit comparison of data originated from new patients to large publicly available datasets, such as TCGA.

Method

We recently proposed to use a data-driven deconvolution method – consensus independent component analysis (ICA) to decompose heterogeneous transcriptomics data and extract features suitable for patient diagnostics and prognostics. The method separates biologically relevant transcriptional signals from technical effects and provides information about cellular composition and biological processes. We applied this method to RNA-seq data originated from several studies on patients with skin cutaneous melanoma (SKCM), brain tumours: glioblastoma (GBM) and low-grade gliomas (LGG), and non-small cell lung cancers: squamous cell carcinoma (LUSC) and adenocarcinoma (LUAD). We investigated the stability of the deconvolution method using TCGA datasets and validated the approach on independent public and in-house datasets from patients with corresponding tumours.

Results

The proposed method efficiently cleans the data from technical biases and allows making diagnostic and prognostic conclusions about the new patients using TCGA datasets as references. In case of brain and lung tumours, we validated classification on independent cohorts of patients obtaining high accuracy of classification between tumour subtypes. ICA showed accuracy at least comparable to other, task-specific feature selection methods. We also were able to build a survival predictor score that showed a strong significance on independent melanoma (pv=1.3e-3, 44 new samples) and glioma (pv=3.4e-18, 352 new samples) datasets. In addition, the method provided information about biological

processes activated in the new patient tumours. It was used to stratify and estimate tumour and stroma-specific signals.

Automated cell type gating on Chip cytometry-data

Svenja Gaedcke[1], Adan Chari Jirmo PhD[1,2], David S. DeLuca PhD[1], Christine Happle MD,PhD[1,2], Anna-Maria Dittrich MD[1,2], Gesine Hansen MD[1,2]

[1]Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), [2]Clinic for Pediatric Pneumology, Allergology and Neonatology, Hannover Medical School, Hannover

BACKGROUND: Cell type identification in both Flow cytometry and Mass cytometry such as in the case of Chip cytometry is usually based on manual gatings, which requires the investigator to define a strategy based on previous knowledge about the cell types leading to judgement calls which inevitably introduce subjectivity. We developed an automated and unsupervised gating method to reduce the uncertainty introduced by 'manual' gating steps and to enable the potential discovery of new leukocyte sub populations.

METHODOLOGY: For each chip a set of cell type-identifying biomarkers was typically used for surface characterisation and subsequent identification through manual gating of various leucocyte subpopulations within peripheral blood. Those cell type-identifying markers were used analogously in this approach for the development of an automated gating method. The automatic clustering was achieved by applying a combination of multiple unsupervised cluster steps using *Gaussian mixture model based clustering* and dimensionality reduction using *t-SNE*.

DISCUSSION: Our automatic gating approach of Chip cytometry generated data based on various surface biomarkers showed that it is possible to identify clusters of cells with similar expression patterns as defined using a manual gating strategy and thus demonstrates the possibility for applying our automatic gating approach on Chip cytometry generated data. Importantly, the automated supercluster outcomes in our model are consistent with the manual defined immunological cell types. We hereby conclude that this approach enhances the objectivity in analysing Chip cytometry generated data, reduces the time required for data processing and increasing the possibility of discovering previous undefined novel leukocyte subpopulation.

A deep learning approach to short antimicrobial peptide prediction[+]

Pratiti Bhadra[1,2] , and Shirley Siu[1]

[1]Department of Computer and Information Science, University of Macau, Macau, China and [2]Center for Bioinformatics, Saarland University, Saarbrücken, Germany

[+]work done at Department of Computer and Information Science, University of Macau, Macau, China

Antimicrobial peptides (AMPs) are promising candidates in the fight against multi drug-resistant pathogens owing to AMPs' broad range of activities and low toxicity. Nonetheless, identification of AMPs through wet-lab experiments is still expensive and time consuming. In particular, short-length (≤ 30 AA) AMPs are promising drug agents which have enhanced antimicrobial activities, higher stability, lower toxicity to human cells, and low production cost. However, existing AMP prediction methods achieved only 60 to 70% accuracy for short length AMPs, even for our previously developed method AmPEP[1]; the problem is due to the largely mixed sequences of different lengths in the training dataset. To provide a solution to short AMP prediction, we have developed Deep-AmPEP30, which is a deep learning method based on an optimal feature set of reduced amino acids composition and convolutional neural network[2]. Deep-AmPEP30 yields an improved performance of 83% in accuracy, 92% in AUC-ROC, and 94% in AUC-PR over existing machine learning-based methods. To show usage of Deep-AmPEP30 in the field of drug discovery, we have screened all open reading frames from the genome sequence of *Candida glabrata* for potential AMPs. *Candida glabrata* is a gut commensal fungus expected to interact with and/or inhibit other microbes in the gut. Selected high-scoring peptides were subjected to experimental validation for antimicrobial activities and a 20-AA peptide P3 (FWELWKFLKSLWSIFPRRRP) that showed strong anti-bacteria activity against *Bacillus subtilis* and *Vibrio parahaemolyticus* were identified. This peptide has a potency comparable to that of ampicillin in the bacterial inhibition assay. Our proposed prediction methods, AmPEP and Deep-AmPEP30 are freely available at http://cbbio.cis.umac.mo/AxPEP for both individual sequence prediction and genome screening of AMPs.

[1] Bhadra, Pratiti, Jielu Yan, Jinyan Li, Simon Fong, and Shirley WI Siu. "AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest." *Scientific reports* 8, no. 1 (2018): 1697.

[2] Jielu Yan, Bhadra, Pratiti, Hio Kuan Tai , Ang Li, Pooja Sethiya, Longguang Qin, Koon Ho Wong and Shirley WI Siu. "Deep-AmPEP30: Improve short antimicrobial peptides prediction with deep learning " (Submitted)

Methylation analysis of combined drug application in leukemia cell lines

Yvonne Saara Gladbach[1,2,3], *Anna Richter*[4], *and Mohamed Hamed*[1]

[1] Rostock University Medical Center, Institute for Biostatistics and Informatics in Medicine and Ageing Research (IBIMA), Rostock, Germany

[2] Faculty of Biosciences, Heidelberg University, 69120, Heidelberg, Germany

[3] Division of Applied Bioinformatics, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT) Heidelberg, 69120, Heidelberg, Germany

[4] Department of Internal Medicine, Medical Clinic III - Hematology, Oncology, Palliative Medicine; Rostock University Medical Center, University of Rostock, 18057 Rostock, Germany

Background

Methylations play essential roles in diverse biological processes, and cancer aberrant DNA methylation patterns can drive e.g. malignant transformation. Further, distinct DNA methylation patterns might be critical in chemotherapy resistance and need; therefore, further investigation. The PI3K/AKT pathway is involved in acute B-lymphoblastic leukemia (B-ALL) pathogenesis. Casein Kinase II (CK2) is upregulated in a variety of neoplasms and enhances PI3K signaling via phosphorylation of tumor suppressor PTEN.

Methods and results

Incubation with CX-4945 resulted in PI3K pathway downregulation and induced anti-proliferative effects on B-ALL cell lines. Combination with DEC synergistically decreased the metabolic activity.

Finally, *in vivo* assessment of CX-4945 anti-tumor potential in B-ALL xenografts resulted in decreased leukemic blast frequency. We further showed experimentally that CX-4945 in combination with DEC further decelerated B-ALL growth significantly.

Discussion

Our analysis showed that treatment of patient-derived xenografts with CX-4945 alone did not change tumor cell proliferation or infiltration while addition of DEC reduced blast frequency in one sample and evaluates the effect of combined CK2 inhibition and DEC-mediated epigenetic modification. We herein demonstrate significant anti-

tumorigenic potential *in vitro* while CX-4945's contribution of the observed anti-proliferative effect *in vivo* remains to be elucidated.

Chemokine signaling in different cancer subtypes

Atanas Aleksandrov, Faculty of Biosciences, Heidelberg University, Heidelberg, Germany; Thorben Hennig, Faculty of Biosciences, Heidelberg University, Heidelberg, Germany; Benedikt Brors, German Cancer Research Center (DKFZ), Division of Applied Bioinformatics, Heidelberg, Germany, German Cancer Consortium (DKTK), Heidelberg, Germany; Maria Dinkelacker, Faculty of Biosciences, Heidelberg University, Heidelberg, Germany, German Cancer Research Center (DKFZ), Division of Applied Bioinformatics, Heidelberg, Germany

Chemokine signalling in cancer plays an important role in the recruitment of different immune cells into the primary site of cancer. The microenvironment of cancer has a high impact on the expression of these signalling molecules. Different subtypes of immune cells are attracted through the chemical gradient of chemokine ligands through their receptors to the tumor and lead to a progression or inhibition of the cancer growth, depending on the subset of immune cells, which are attracted to the tumor. The subtypes of immune cells attracted to the primary tumor sites might decide upon tumor prognosis and patients‘ survival.

Several chemokine ligands and receptors have been shown to be involved in the past, but a systematical analysis of all chemokine ligands and receptors has not been done. In this work we systematically studied the gene expression of all chemokine ligands and receptors in breast cancer, lung cancer, prostate cancer and melanoma. For this we have analyzed both gene expression datasets as well as single cell data, including many immune cell subsets.

In this context especially the chemokine ligands CCL2, CCL5, CXCL9 and CXCL10 have shown to be important. CCL2 as well as CCL5 mainly attract macrophages, which then express CXCL9 as well as CXCL10, which lead to the recruitment of CD8 cytotoxic T cells, which inhibit tumor growth and lead to good overall prognosis.

Analysis of methylation profiles in hepatic cancer

Eugen Rempel[1], Benjamin Goeppert[1], Peter Schirmacher[1], Albrecht Stenzinger[1], Stephanie Roessler[1], and Jan Budczies[1]

[1] Institute of Pathology, University Hospital Heidelberg, Im Neuenheimer Feld 224, 69120 Heidelberg, Germany

Background

Epigenetic processes have been shown to play an important role in hepatic cancer. However, the analysis of methylation patterns in clinical tissues is challenging since the samples represent a complex mixture of different cell types. Here, we present the application of various bioinformatic methods to infer the composition of tissue samples these cell types and their contribution to differential methylation.

Methods and results

The study comprised of EPIC BeadChip data from 60 patients with liver cancer. Samples of normal, premalignant and malignant tissues were analyzed. After exploratory data analysis and batch correction, we applied various methods to decompose the heterogeneous methylation profiles and to elucidate the cell types responsible for differential methylation. We used both reference-based (CellDMC) and reference-free (MeDeCom) approaches. The methods were applied in parallel fashion and their results w.r.t. differential methylation are compared.

Discussion

We observed a strong influence of the cell type composition on the methylation patterns in the hepatic cancer cohort.

Therefore, decomposition of methylation patters into different cell types is crucial for biological interpretation and biomarker development.

Image processing on the brain: 3D printing MRI images for further observation

Lautaro Franco Soler, Madison High School, Madison NJ, United States; Lautaro Soler

Abstract: The purpose of this project was to take partially processed MRI images of the brain and further process them in order to create a 3D model for further, physical analysis once printed. This approach attempted to mirror the processes medical officials use on a daily basis for reconstructing 3D models of organs to determine if any abnormalities are present, for use prior to surgery or any other medical intervention. In this case, the goal of this project was to use publicly available MRI data to recreate a 3D model of the brain in order to be able to 3D print and then inspect it for evident of any irregularities. In order to do this, the MRI image slices needed to be cleaned up, normalized and filtered. Histograms were used to determine brightness cutoff ranges and set a specific brightness threshold. Blurring techniques and edge detection were used to determine the most likely edges of the images of the brain. In the end, edge detection allowed for a 3D model to be generated and exported for 3D printing.

A new approach for the automatic detection of follicular regions in Actin-stained whole slide images of the human lymph node using *shock filter*

Patrick Wurzel[1], *Hendrik Schäfer*[2], *Jörg Ackermann*[1], *Martin-Leo Hansmann*[2],*Ina Koch*[1]

[1]*Department of Molecular Bioinformatics, Institute of Computer Science, Johann Wolfgang Goethe-University Frankfurt am Main, Robert-Mayer-Str. 11-15, 60325 Frankfurt am Main*

[2]*Dr. Senckenbergisches Institut für Pathologie, Universitätsklinikum Frankfurt,Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany*

Background

In the course of antigen-driven processes, lymphoid follicles go through different morphological stages. The analysis of follicular regions with regard to morphological characteristics, distribution and structure can provide conclusions about the clinical picture and the course of disease [1]. For that purpose, we developed a technique for the automatic detection of follicular regions.

Methods and results

The new method is based on histological *whole slide images* (WSI) of Actin-stained tissue sections. It is divided in three phases. In an initial step, we preprocess the image to reduce image noise. Furthermore, we apply a *shock filter*-based detection method to identify possible transition points [2]. In a third conclusive step, follicular structures gets remodeled and segmented on the basis of the identified transition points. A subsequent validation of the final segmentation revealed on average a reasonable *Zijdenbos similarity index* (ZSI) of 70.8%.

Discussion

Applying state of the art segmentation methods like watershed segmentation and otsu thresholding failed to produce satisfactory results with averaged ZSIs of 47.9% and 48.9%. This is caused by different challenges. At first, lymphoid follicle surrounding sinus structures and vessels exhibit similar characteristics to those of follicular regions. Furthermore, we had to handle properly obstacles related to clustered lymphoid follicles. Finally, we had to deal with the occurrence of intensity differences throughout single tissue sections and multiple WSI's.

[1] Belkacem-Boussaid et al., Computerized Medical Imaging and Graphics, 2011

[2] Belean et al., BMC Bioinformatics, 2015

Single-Cell RNA Sequencing Uncovers a Large-Oncosome-Signature in Medulloblastoma

Marta Interlandi, Institute of Medical Informatics – University Hospital, Münster, Germany; Thomas K. Albert, Department of Pediatric Hematology and Oncology - University Children's Hospital, Münster, Germany; Kornelius Kerl, Department of Pediatric Hematology and Oncology - University Children's Hospital, Münster, Germany; Martin Dugas, Institute of Medical Informatics – University Hospital, Münster, Germany

Background: Medulloblastoma (MB) constitutes the most frequent pediatric malignancy of the central nervous system and shows a wide spectrum of clinicopathological features. Despite multiple efforts have been made in defining molecular subgroups of MB, they do not provide an explanation as to why very high-risk tumors consistently emerge from discrete molecular subpopulations that share little overlap in their defining (epi)genotypes.

Methods and results: We performed single-cell RNA sequencing on four samples of a MB mouse model specific of a certain molecular subgroup, namely sonic hedgehog. The analysis conducted on this data, as well as on a published single-cell transcriptome atlas of the developing mouse brain, uncovered a gene expression signature that results from a mutual communication between tumor cells and infiltrating non-tumor cells through extracellular vesicles, named Large-Oncosome-Signature (LOS). Importantly, this signature is also detected in bulk RNA-seq data and RNA microarray profiles of medulloblastoma patients. Through the implementation of a score, we show that LOS provides an independent prognostic marker in two large MB human cohorts.

Discussion: LOS informs about biologically and clinically relevant subtypes and might help to refine the current classification of medulloblastoma.

Single-cell RNA Analysis in Galaxy

Mehmet Tekman, Bioinformatics Lab, Albert-Ludwigs-Universität Freiburg im Breisgau, Germany; Bérénice Batut, Bioinformatics Lab, Albert-Ludwigs-Universität Freiburg im Breisgau, Germany; Björn Grüning, Bioinformatics Lab, Albert-Ludwigs-Universität Freiburg im Breisgau, Germany

Background

Single-cell RNA (scRNA) analysis provides an important means to understanding the development and function of cell identity within heterogeneous tissue through transcriptome profiling and differential analysis. Several independent packages and pipelines have been developed to address the non-trivial demultiplexing and the statistical complexity of the downstream analysis, each with varying degrees of compatibility. Here, we outline https://singlecell.usegalaxy.eu, a webserver to process, analyse and visualize single-cell data, containing several Galaxy workflow components to streamline this entire process. In addition we developed hands-on training material, so that everyone can start learning scRNA data analysis.

Method and Results

The workflows consist of two main parts: the pre-processing, that demultiplexes and deduplicates scRNA data to produce a cell/transcript count matrix; and the downstream analysis, to perform cell clustering and lineage inference. The pre-processing tutorials are driven through the discussion of cell barcodes and Unique Molecular Identifiers to provide a comprehensive understanding of how different wet labs generate their FASTQ data. The downstream analysis tutorial makes use of the RaceID3 and the ScanPy suite to educate users in robust clustering, and the common pitfalls associated with scRNA analysis.

Discussion

Our Galaxy workflows are flexible to different library setups and can be applied to other use-cases beyond transcriptomics. Additional modules, like STARsolo, provide one-click solutions to 10X data as it becomes more prevalent in the field. The pre-processing and downstream workflows together in Galaxy facilitate the analysis of scRNA data in a transparent, useable, and reproducible manner.

Comparative single cell transcriptomic analysis of the hematopoietic system between human and mouse

Shouguo Gao, Zhijie Wu, Xingmin Feng, Vivian Dong, Sachiko Kajigaya, Neal S. Young, Hematology Branch, NHLBI, NIH, Bethesda, MD, USA

The cross-species comparison of hematopoietic hierarchy is still not thoroughly studied. We constructed a single-cell resolution transcriptomic atlas of HSPCs in human and mouse, from a total number of 32,805 single cells. By known marker genes, we grouped human cells as hematopoietic stem cell (HSC), multilymphoid progenitor, granulocyte-monocyte progenitor (GMP), ProB cell, earliest thymic progenitor, and megakaryocytic-erythroid progenitor (MEP); and mouse cells as long-term HSC, lymphoid multipotent progenitors, and multipotent progenitor (MPP), GMP, MEP and common myeloid progenitors. Using orthologous genes of human and mouse in InParanoid (http://inparanoid.sbc.su.se), after alignment with CCA in Seurat 2.0 (https://satijalab.org/seurat/), the cells of mouse and human were well-mixed and separated by same cell type categories. The cells were grouped into 17 subpopulations and cluster-specific genes were species -conserved and shared same functional themes.

After calculating an average of expression of cells in each population of human and mouse, the clustering dendrogram indicated that cell types were highly conserved between human and mouse. For example, MEP and GMP of mouse and human shared a very similar transcriptome pattern. Human HSC was firstly clustered with mouse LTHSC and then with mouse MPP. We used Monocle to examine the differentiation trajectory of hematopoiesis in human and mouse. Graphically, an intuitive representation of HSPC differentiation emerged. In both mouse and human, three branches (Erythroid/megakaryocytic, Myeloid, and Lymphoid) derived directly from HSC and LTHSC. To understand the species conservation of hematopoitic subpoluations, single-cell transcriptomes of human cells were compared with those of mouse cells using scmap. Most human MEP cells (85%) mapped to mouse MEP cell types based on transcriptional similarity, suggesting species conservation of functional organization. Further, 45% and 24% human HSC cells were mapped to mouse LTHSC and MPP cell types, respectively. Mapping of other cell types was consistent to the mechanism of hematopoiesis. Our analysis confirms evolutionary conservation in the hematopoietic systems between mouse and man.

Exploring gene expression landscape of DMBA-TPA induced skin carcinoma at the single cell level

Luca Penso-Dolfin, Deutsches Krebsforschungszentrum, Heidelberg, Germany
Mikaela Behm, Deutsches Krebsforschungszentrum, Heidelberg, Germany
Duncan Odom, Deutsches Krebsforschungszentrum, Heidelberg, Germany
Michaela Frye, Deutsches Krebsforschungszentrum, Heidelberg, Germany
Angela Goncalves, Deutsches Krebsforschungszentrum, Heidelberg, Germany

ABSTRACT

DMBA-TPA-induced carcinogenesis is the most commonly used mouse model of skin cancer. In this mouse model, carcinogenesis progresses through the development of benign papillomas into squamous cell carcinomas. Here we assay the single-cell genome wide expression patterns of normal versus treated mouse skin samples in order to characterize this model at an unprecedented level of detail. We explore changes to tissue composition, cell-type specific expression patterns and cell-to-cell communication, and we further use somatic mutations to track clonal expansions in the tissue. The identification of the key events during mouse skin carcinogenesis will be useful for rationalizing future functional mouse studies and determining which genes present conserved functions between mouse and human cancers. Future work will focus on the comparison of gene expression between the mouse (M. musculus) and the Naked Molerat (H. glaber), in order to further elucidate the molecular basis of the resistance to cancer observed in the latter.

Biomarker Discovery with Robust Bagging Feature Selection

Giampaolo Pileggi, Brandon Malone

NEC Laboratories GmbH, Heidelberg, Germany

Background

Biomarker discovery lies at the heart of many advances in diagnostic and prognostic medical testing; however, noisy, high-throughput assays, such as RNA sequencing, performed on samples from different conditions are often the first step in identifying potential biomarkers. It is important to identify a small, robust set of biomarkers in this setting for further validation for the development of widely-applicable, inexpensive tests.

Methods and results

In this work, we propose a robust feature selection approach based on bagging (RoBag FS) to address the biomarker discovery problem. Briefly, we split the training data into multiple overlapping subsamples, called bags. Multiple base feature ranking approaches are applied to each bag, and we select the ranking which is most consistent with the rankings from all bags and ranking strategies.We validate RoBag on a publicly-avaliable dataset which measures microRNA expression from primary tumor tissue for more than 8000 patients from 29 different tumor classes. We demonstrate that using only 10 microRNAs selected by RoBag gives an accuracy of 0.89 ± 0.01, while improving to 0.94 ± 0.01 when using 50 or 100 features. RoBag yields comparable performance to a recently-proposed approach on the same dataset, which also obtained a mean accuracy of 0.94 ± 0.01 using a subset of 100 features. (They did not report results for fewer features.)

Discussion

We have shown that RoBag can select a small, robust set of features which give predictive performance comparable to all features. This robustness is very important in the context of biomarker discovery since different cohorts of patients may introduce subtle bias in the training data. Our approach is robust against both overfitting due to small sample size and the choice of ranking strategy.

Standard bagging approaches, such as random forests, learn models which use different features from each bag. In contrast, we select a small set of features which are consistently important across all bags. Consequently, our approach is appropriate for biomarker discovery, while standard bagging approaches result in ensembles which are not suitable for identifying a small set of important features.

Our results show that our selected features are effective across a broad cohort of patients. This biomarker selection is an important first step in the development of inexpensive, yet effective, diagnostic and prognostic tests.

How to “appreci8” Variant Calling in NGS Data

Sarah Sandmann, Institute of Medical Informatics, University of Münster, Münster, Germany; Martin Dugas, Institute of Medical Informatics, University of Münster, Münster, Germany

Background: Calling SNVs and indels in next-generation sequencing (NGS) data is a long-discussed topic. Numerous variant calling approaches are currently available. Despite considerable advances in recent years, detecting low-frequency variants with equally high sensitivity and positive predictive value (PPV) is still challenging. However, for the application of NGS in clinical routine it is essential to rely on valid variant calling results – for high- as well as low-frequency variants.

Methods and results: In close collaboration with medical and biological experts, we developed a variant calling approach called appreci8 - A Pipeline for PREcise variant Calling Integrating 8 tools. Our pipeline combines and filters the output of 8 previously evaluated open-source variant calling tools, using a novel artifact- and polymorphism score. The algorithm aims at reproducing a biologist's manual work when investigating a raw list of calls. Detailed analysis of 7 well-characterized targeted sequencing data sets (678 samples) shows that appreci8 succeeds in calling variants with allelic frequencies at 1% with high sensitivity (0.93 to 1.00) and high PPV (0.65 to 1.00). Appreci8 is able to outperform every individual tool as well as alternative approaches for combining the 8 tools (best alternative: GATK with sensitivity=0.82-0.95 and PPV=0.31-0.94).

Extending the idea of appreci8, we also developed the appreci8R, which is an R-version of our algorithm. The appreci8R provides an additional shiny GUI and checkpoints for re-starting the analysis from intermediate results. Furthermore, all analysis parameters, including the artifact- and polymorphism score, are user-definable. Even the combination and number of input tools can be defined without restriction. Comparing appreci8 and the appreci8R we observe highly comparable performance of both approaches.

Appreci8 is available via Docker (https://hub.docker.com/r/wwuimi/appreci8/), the appreci8R is available via Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/appreci8R.html).

Discussion: Appreci8 and the appreci8R provide a novel approach for accurate calling of high- as well as low-frequency variants in NGS data.

Evaluation of Pipelines for Meta-Analysis of ChIP-seq Data

Ihsan Muchsin[1]; Moritz Kohls[1]; Klaus Jung[1], [1]Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Hannover, Germany

Background

Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq) is a high-throughput method to analyse DNA-protein interactions. The level of evidence for individual experiment is usually limited and error-prone. Since the data of many experiments and studies are deposited in public repositories such as ENCODE (The ENCODE Project Consortium 2012, Science, 306, 636-40), the option for meta-analyses to increase the level of scientific evidence exists. A pipeline for meta-analysis of ChIP-seq data by merging normalized NGS reads has been implemented (Chen et. al. 2011, Genome Biol., 12, R11). Additionally, quantile normalization followed by p-values combination has been studied (Chen 2015, Dissertation, Univ. Pittsburgh). In this study, we evaluated comparatively and refined these existing pipelines.

Method and results

We used ChIP-seq data from computer simulations and public repositories. We compared two distinct approaches, i.e., reads merging and peaks merging. In the former case, we directly merged the reads from all studies and then performed the reads mapping and peaks calling consecutively. In the latter case, reads mapping and peaks calling were done for each individual study. Then, we determined the consensus peaks and their corresponding p-values. In this case, the consensus peak can be reflected by either the union of the overlapping peaks or the intersection of the overlapping peaks. A p-value reflecting the binding activity was assigned to the consensus peak using p-value combination methods, e.g., Fisher's method and Stouffer's Z-score. Peaks merging showed higher precision and sensitivity with the simulated data. Peaks merging also showed better motif enrichment value with the experimental data. In peaks merging approach, union and intersection methods showed equal performance.

Discussion

The results show that peaks merging outperformed reads merging, but in our current simulation studies we are still evaluating different p-value combination methods. In

addition, we are also evaluating the effect of different normalization and imputation strategies in p-value combination.

ExtraCytoplasmic Function (ECF) sigma factor classification: from phylogenesis to protein regulation

D. Casas-Pastor, LOEWE Center for Synthetic Microbiology (SYNMIKRO), Marburg; S. Chandrashekar Iyer, Max Planck Institute for Terrestrial Microbiology, Marburg; R. Müller, Justus-Liebig-Universität, Gießen; T. Mascher, Technische Universität Dresden, Dresden; A. Goesmann, Justus-Liebig-Universität, Gießen; S. Ringgaard, Max Planck Institute for Terrestrial Microbiology, Marburg; G. Fritz, LOEWE Center for Synthetic Microbiology (SYNMIKRO), Marburg

ExtraCytoplasmic Function sigma factors (ECFs) are the most abundant and diverse subfamily of sigma factors in bacteria. Sigma factors drive the RNA polymerase (RNAP) to specific promoter regions, from which they initiate transcription. In ECFs this function is dependent upon a signal, transmitted to the ECF via its regulators, including anti-sigma factors or two-component systems. ECF phylogenetic groups are associated to a common regulator encoded in their genetic context. Nevertheless, previous classification efforts focused only on ~500 genomes (*1-3*). In this work we expand the ECF classification using the full array of genomes in the NCBI database. Using a combination of clustering tools, we classified an extended the ECF library into two levels, defining subgroups of closely-related ECFs that are further hierarchically clustered into groups with a common genetic neighborhood. Then, we generated hypothesis about the regulation of the ECF groups from the conserved elements encoded in their genomic context and confirmed it for members of ECF43. The genetic context of group ECF43 contains a conserved transmembrane eukaryotic-like serine/threonine protein kinase (STK). Interestingly, members of ECF43 diverge from canonical ECF sigma factors in the main RNAP binding region, where members of ECF43 contain a conserved threonine instead of a negatively charged amino acid. We show for the first time that members of ECF43 are activated upon phosphorylation of this residue, that this phosphorylation is dependent on the STK and is required for resistance to polymyxin antibiotics in *Vibrio parahaemolyticus*. Taken together, our new ECF classification provides predictions for the regulation of for more than 150 phylogenetic groups, covering the full bacterial spectrum, which the community can use as a source for testable hypothesis.

References

1. A. Staroń *et al.*, The third pillar of bacterial signal transduction: classification of the extracytoplasmic function (ECF) σ factor protein family. *Molecular Microbiology*. **74**, 557–581 (2009).
2. X. Huang, D. Pinto, G. Fritz, T. Mascher, Environmental sensing in Actinobacteria: a comprehensive survey on the signaling capacity of this phylum. *J. Bacteriol.* **197**, 2517–2535 (2015).
3. C. Jogler *et al.*, Identification of proteins likely to be involved in morphogenesis, cell division, and signal transduction in Planctomycetes by comparative genomics. *J. Bacteriol.* **194**, 6419–6430 (2012).

TportHMM: Classifying Transport Proteins Using Profile Hidden Markov Model and Specificity Determining Sites

Qing Ye, Concordia University, Montreal, Canada; Greg Butler, Concordia University, Montreal, Canada

Background

Transporters make up a large proportion of proteins in a cell, and play important roles in metabolism, regulation, and signal transduction by mediating movement of compounds across membranes. There is a need for tools that predict the substrate that is transported at the level of substrate class and the level of specific substrate.

Methods

Our work investigates the effect of a profile Hidden Markov Model (HMM) classifier for predicting the substrate class of a transport protein when combined with our EPRCS methodology. In this case, the EPRCS methodology uses multiple sequence alignment (MSA) algorithms to utilise evolutionary information, specificity-determining site (SDS) algorithms to highlight positional information, and HMM to utilise sequence information in our predictor.

We study the impact of performance when different MSA algorithms (ClustalW, Clustal Omega, MAFFT, MUSCLE, T-Coffee and TM-Coffee) are used, and when different SDS algorithms (Speer Server, GroupSim, Xdet) are used. Note that TM-Coffee specifically tries to align transmembrane segments of the proteins. We compare these approaches with the state-of-the-art, TrSSP and TranCEP.

Results

For the dataset from TrSSP, on the independent test dataset, measured by MCC, the best combination, called TportHMM, was MUSCLE with Xdet. It outperformed TrSSP on each of the seven substrate classes, and it outperformed TranCEP on four classes, tied once, and lost twice. The average MCC across the seven classes was TportHMM: 0.71, TranCEP: 0.69, and TrSSP: 0.41.

References

[TranCEP] Alballa M, Aplop F, Butler G: TranCEP: *Predicting transmembrane transport proteins using composition, evolutionary, and positional information*, bioRxiv, 2018.

[TrSSP] Mishra NK, Chang J, Zhao PX: *Prediction of membrane transport proteins and their substrate specificities using primary sequence information*, PLoS One. 2014;9(6):1-14.

HiCExplorer 3: A toolbox for Hi-C data analysis

Joachim Wolff Bioinformatics Lab, Albert-Ludwigs-Universität Freiburg im Breisgau, Germany; Leily Rabbani Max-Planck-Institut für Immunbiologie und Epigenetik, Freiburg im Breisgau, Germany; Gautier Richard Max-Planck-Institut für Immunbiologie und Epigenetik, Freiburg im Breisgau, Germany; Thomas Manke Max-Planck-Institut für Immunbiologie und Epigenetik, Freiburg im Breisgau, Germany; Asifa Akhtar Max-Planck-Institut für Immunbiologie und Epigenetik, Freiburg im Breisgau, Germany; Rolf Backofen Bioinformatics Lab, Albert-Ludwigs-Universität Freiburg im Breisgau, Germany; Fidel Ramirez Target Discovery Research, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riß, Germany; Björn Grüning Bioinformatics Lab, Albert-Ludwigs-Universität Freiburg im Breisgau, Germany

HiCExplorer is a toolbox to analyse and explore the 3D structure of the DNA based on Hi-C (high-throughput sequencing chromatin conformation capture) data. Tools for all steps of an analysis of Hi-C data like creation of Hi-C interaction matrices, quality assessment, correction of Hi-C interaction matrices and identification of A/B compartments, chromatin loops and topological associated domains (TADs) are provided. Users can create publication ready plots of the two dimensional Hi-C interaction matrix and add A/B compartments and chromatin loops information. Additionally, pyGenomeTracks can be used to plot TADs on a selected genomic locus, along with additional information like gene tracks or ChIP-seq signals. Moreover, HiCExplorer supports the analysis of cHi-C data for promoter-enhancer interactions based on a background model. With the Galaxy HiCExplorer web server we provide computational resources and users with little bioinformatic background can perform every step of the Hi-C analysis in a simple to use web browser user interface.

A novel method for achieving an optimal classification of the proteinogenic amino acids

Andre Then[1,a], *Karel Mácha*[1,a], *Bashar Ibrahim*[1], *Stefan Schuster*[1,*]

[1] *Chair of Bioinformatics, Matthias-Schleiden-Institute, University of Jena, Ernst-Abbe-Platz 2, Jena 07743, Germany*

[a] *These authors contributed equally to this work*

* *Corresponding author*

Abstract

Classification of the proteinogenic amino acids is crucial for understanding their commonalities as well as their differences in order to provide a hint why life settled on usage of precisely those amino acids and is also crucial for predicting electrostatic, hydrophobic, stacking and other interactions, for assessing conservation in multiple alignments and many other applications. While several methods have been proposed to find “the” optimal classification, they have several shortcomings, such as the lack of efficiency and interpretability or an unnecessarily high number of discriminating features. In this study, we propose a novel method with repeated binary separation via a combination of the minimum amount for five features (such as hydrophobicity or gyration radius) expressed by numerical values for amino acid characteristics. The features were extracted from the AAindex database. We successfully find four such combinations by simple separation at the medians. We further extend our analysis to separations other than by the median. Here it is shown that obtaining numerous combinations allowing for optimal classification is not an issue. We therefore scored our combinations based on how natural the separations performed by the features included in the obtained combinations are and identified several high ranking combinations in the process. We examined an experimental study from literature where incorporation of unnatural amino acids into the genetic code was used to enhance antibody binding towards a HIV coat protein. Or method is able to suggest the most diverse amino acids for incorporation upon counting how often they occupy vectors unoccupied by natural amino acids in our high ranking combinations.

FastQuality – A machine learning application to assess the quality of early stage sequencing data (fastq files)

Steffen Albrecht, Johannes Gutenberg-Universität, Mainz, Germany; Miguel A. Andrade-Navarro, Johannes Gutenberg-Universität, Mainz, Germany; Jean-Fred Fontaine, Johannes Gutenberg-Universität, Mainz, Germany

Next-generation sequencing is a powerful technology highly relevant in biomedical research and pharmaceutical industry. Applied in combination with molecular assays it provides detailed insights in cell processes such as gene expression, DNA accessibility, and protein-DNA interactions. However, the clinical or biological relevance of results from these assays is extremely sensitive to the quality of the sequencing data. Existing quality control tools either produce comprehensive reports, to be manually inspected, or they reveal their quality metrics after extensive computations. This is critical considering the increasing amount of sequencing data due to decreasing costs. Consequently, a tool is needed that produces actionable and easy to interpret reports and that is applicable to early stage sequencing data.

In this study, we investigated the possibility to predict the quality of sequencing data from standard fastq files containing the raw sequencing reads. Quality metrics were derived for 3024 fastq files from the ENCODE data portal, already labeled to be of either low or higher quality. Based on these quality metrics, calculated by established methods such as FastQC and Bowtie2, we trained and cross validated classification models to predict the ENCODE quality label. In order to find the best performing model we used a grid search that comprehends 10 state-of-the-art machine learning algorithms and a total of 19417 parameter settings. Algorithms such as multilayer perceptron, random forest, and gradient boosting machines achieved remarkable predictive performance scores, reaching area under ROC-curve values up to 0.97. Interestingly, the performance changed depending on various underlying datasets that are defined by different species-assay combinations and different sets of quality metrics. For instance, a lower accuracy was observed for human-ChIP-seq files compared to mouse-ChIP-seq that could be explained by an over-representation of human cancer cell lines, known for higher rate of mutations.

From the high predictive performance, observable across all datasets, we conclude that there is a strong potential of machine learning algorithms to perform automatic quality assessment of early stage sequencing data based on relevant quality metrics.

Predicting the Presence of Redox-Modified Cysteines in Proteins using Machine Learning and Statistical Methods

Marcus Keßler, Goethe-University, Frankfurt am Main, Germany

Ilka Wittig, Goethe-University, Frankfurt am Main, Germany

Ina Koch, Goethe-University, Frankfurt am Main, Germany

Background

Non-toxic levels of reactive oxygen species (ROS) are known to affect many cellular processes through several post-translational modifications (PTMs) of cysteines [1], but their experimental detection is non-trivial. As common sequence motifs surrounding modifiable cysteines have not yet been found, there is currently a lack of computational methods for their prediction. Thus, we have attempted to use statistical and machine learning methods like the support vector machine algorithm to investigate structural and sequence protein data to find commonalities between redox-modifiable cysteines.

Methods and results

To better be able to predict redox PTMs, we used the bsvm 2.08 implementation of the support vector machine algorithm [2], a supervised learning algorithm used for classification and regression analysis. To train our algorithm, we gathered data from different databases like redoxDB [3] and PDB [4]. This data was then used to produce features, such as structural information, half sphere exposure (HSE) [5] and predicted pK_a [6] values.

We tested our methods on data sets with varying sequence identities and confirmed our results via cross validation. For proteins with up to 80% sequence identity, our method had a sensitivity of 20% and a specificity of 80% for modified cysteines, as well as a positive predictive value (PPV) of 0.37 and a negative predictive value (NPV) of 0.76 for a data set consisting of 25% modified cysteines. With 40% sequence identity, a sensitivity of 10% and a specificity of 94% was obtained, with a PPV of 0.38 and an NPV of 0.76. We were also able to confirm which features seemed particularly valuable for prediction. Amino acid sequence, secondary structure and HSE were the most useful features.

Discussion

While our method shows promise, it cannot currently be used reliably to predict redox PTMs, but may be a fast way to assist researchers at choosing which proteins show

promise for further experimentation. More data as well as more features such as protein-protein interactions or subcellular location may improve results.

[1] L. Bleier, I. Wittig, H. Heide, M. Steger, U. Brandt, S. Dröse. Generator-specific targets of mitochondrial reactive oxygen species. Free Radical Biology and Medicine 78: 1–10, 2015.

[2] C. Hsu and C. Lin. A simple decomposition method for support vector machines. Machine Learning 46: 291-314, 2002.

[3] M. Sun, Y. Wang, H. Cheng, Q. Zhang, W. Ge, D. Guo. RedoxDB - a curated database of experimentally verified protein redox modification. Bioinformatics, 28(19): 2551-2552, 2012.

[4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The Protein Data Bank. Nucleic Acids Research, 28: 235-242, 2000.

[5] J. Song, H. Tan, K. Takemoto, T. Akutsu. HSEpred: predict half-sphere exposure from protein sequences. Bioinformatics, 24 (13): 1489–1497, 2008

[6] D. Bas, D. Rogers, J. Jensen. Very fast prediction and rationalization of pKa values for protein-ligand complexes. Proteins, 73:765–783, 2008.

Pathway Activation Prediction using Signed Directed Random Walk with Restart

Julian Kreis, University Heidelberg, Heidelberg, Germany; Dilafruz Juraeva, Merck Healthcare KGaA, Darmstadt, Germany; Eike Staub, Merck Healthcare KGaA, Darmstadt, Germany; Benedikt Brors, Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany

Background

Biological pathways are the units of cellular functions with examples ranging from the cell division machinery to the RAS signal transduction. Together, they represent a large network, consisting of nucleic acids, proteins, and small molecules. In cancer cells, abnormal regulation of pathways is an important biomarker, because it may cause various oncogenic transformations, resulting in uncontrolled cell proliferation or resistance to anti-cancer treatment. Due to their complexity and costs, mass spectrometry or reverse phase protein arrays, which measure protein activity, are not scalable to large sample cohorts or substantial sets of proteins. Statistical models which predict the status of a pathway activation from genomic and transcriptomic data provide an alternative. However, these methods make strong assumptions regarding the complex properties of pathways, e.g. only considering gene expression data or disregarding the diverse regulatory machinery of pathways.

Methods and Results

With the aim to better capture the information of available molecular data and known biological interactions for the prediction of pathway activity, we developed a signed directed Random Walk with Restart (RWR) approach. We integrate data from cohorts of The Cancer Genome Atlas (TCGA) with pathway annotation of the Reactome database. We quantitatively predict the levels of protein phosphorylation, representing protein activity (e.g. Src kinase) or pathway activity (known integrators such as AKT1).

Discussion

To our knowledge, our method is the first RWR approach for protein activity prediction, that captures the downstream effects of inhibiting and activating signal transduction. We will present our underlying model, examples of results for individual

cancers and pan-cancer. A thorough comparison of our predictions with results from phosphor-proteomics will be shown and discussed.

autobiogramML: towards automated machine learning in protein function prediction

Michał Burdukiewicz[1], Dominik Rafacz[1], Katarzyna Sidorczuk[2], Stefan Rödiger[3], Przemysław Gagat[2]

[1]Warsaw University of Technology, Warsaw, Poland, [2]University of Wrocław, Wrocław, Poland, [3]Brandenburg University of Technology Cottbus-Senftenberg, Senftenberg, Germany

Background: The advancements in various 'omics' fields have resulted in the discovery of many new protein sequences. Their functional annotations, however, come in at a much slower pace because they require laborious and often expensive experimental procedures. The machine learning models fill in this gap by providing estimates of protein functions. The challenges of developing appropriate models for protein data exclude from the field scientists with limited machine learning expertise and resources. Therefore, we propose autobiograML, an R package designed to automatically apply our framework for protein function prediction [1, 2].

Methods: autobiograML models the relationships between provided protein sequences (encoded as amino acid motifs) and annotations. The Bayesian framework optimizes the hyperparameters of the model in nested cross-validation. The outer layer of the cross-validation is later used to select the optimal machine learning algorithm.

Results and discussion: Although autobiograML does not cover all the intricacies of machine learning, it offers a reliable way to create a model for the prediction of protein function. Our tool will be a valuable machine learning assistant for many research groups studying areas that are too new to have already well-established computational methods.

[1] Burdukiewicz et al. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports *7*, 12961.

[2] Burdukiewicz et al. (2018). Prediction of Signal Peptides in Proteins from Malaria Parasites. International Journal of Molecular Sciences *19*, 3709.

The JSBML project: a fully featured Java API for working with systems biology models

Nicolas Rodriguez[1], Thomas M. Hamm[2,3], Roman Schulte[3], Leandro Watanabe[4], Ibrahim Y. Vazirabad[5], Victor Kofia[6], Chris J. Myers[4], Akira Funahashi[7], Michael Hucka[8], and Andreas Dräger[2,3,9]

[1]The Babraham Institute, Cambridge, United Kingdom

[2]Computational Systems Biology of Infection and Antimicrobial-Resistant Pathogens, Institute for Biomedical Informatics (IBMI), University of Tübingen, Tübingen, Germany

[3]Department of Computer Science, University of Tübingen, Tübingen, Germany

[4]Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT USA

[5]Marquette University, Milwaukee, WI, USA

[6]University of Toronto, Toronto, ON, Canada

[7]Department of Biosciences and Informatics, Keio University, Yokohama, Japan

[8]The California Institute of Technology, Pasadena, CA, USA

[9]German Center for Infection Research (DZIF), partner site Tübingen, Germany

Background:

SBML is the most widely used data format to encode and exchange models in systems biology. The open-source JSBML project has been launched in 2009 as an international collaboration with the aim to provide a feature-rich pure Java implementation for reading, manipulating and writing SBML files.

Results:

The JSBML project has matured into a stable, actively developed, and well-documented software project with a large number of contributors around the world. A growing number of applications is now available that uses JSBML as their back-end for data manipulation. These cover diverse areas of use cases, such as model building and graphical display, constraint-based modeling, dynamic simulation, model annotation, and many more. JSBML supports all levels, versions, and releases of SBML and provides numerous utility functions that facilitate

working with this standard. JSBML also integrates well with other Java libraries for community standards, such as for SBGN or the COMBINE Archive format.

Discussion:

The JSBML team actively maintains and updates the project. JSBML is being used in students' education and numerous research projects. Major model databases, such as BioModels or BiGG Models, use JSBML-based tools for their curation pipelines. JSBML is also regularly subject of international students coding events.

Availability:

Source code, binaries and documentation for JSBML can be freely obtained under the terms of the LGPL 2.1 from the website http://sbml.org/Software/JSBML/ and on GitHub https://github.com/sbmlteam/jsbml/. The users' guide at http://sbml.org/Software/JSBML/docs/ provides further information about using JSBML.

References:

Dräger, A., Rodriguez, N., Dumousseau, M., Dörr, A., Wrzodek, C., Le Novère, N., Zell, A., and Hucka, M. (2011). JSBML: a flexible Java library for working with SBML. Bioinformatics, doi:10.1093/bioinformatics/btv341.

Rodriguez, N., Thomas, A., Watanabe, L., Vazirabad, I. Y., Kofia, V., Gómez, H. F., Miag, F., Mahes, J., Rudolph, J. D., Wrzodek, F., Netz, E., Diamantikos, A., Eichner, J., Keller, R., Wrzodek, C., Fröhlich, S., Lewis, N. E., Myers, C. J., Le Novère, N., Palsson, B. Ø., Hucka, M., and Dräger, A. (2015). JSBML 1.0: providing a smorgasbord of options to encode systems biology models. Bioinformatics, doi:10.1093/bioinformatics/btr361.

Comprehensive analysis of allele-specific methylation in amplicons with amplikyzer2

Sven Rahmann[1,2], *Marcel Bargull*[1,2], *Jasmin Beygo*[2] *and Bernhard Horsthemke*[2]
[1]*Genominformatik,* [2]*Institut für Humangenetik,*
Universitätsmedizin Essen, Universität Duisburg-Essen, Essen, Deutschland

The Illumina MiSeq sequencing platform allows locus-specific DNA methylation analysis using deep bisulfite amplicon sequencing. Existing tools do not always meet the analysis requirements for complex assay designs with multiple regions of interest (ROIs) from multiple samples [1]. We have developed the amplikyzer2 software to address these challenges. It is a complete rewrite of the former amplikyzer software that only analyzed SFF flowgrams from the Roche/454 platform [2]. Amplikyzer2 offers a variety of options to analyze complex multiplexed samples with several regions of interest and outputs useful statistics and publication-quality analysis plots. For example, amplikyzer2 determines the mean methylation values of all CpG dinucleotides in a region, or of each CpG dinucleotide across all reads or a selected set of reads. It is also possible to compare methylation values between different samples and different alleles within a sample. Especially the allele discrimination option is a unique feature of amplikyzer2. The software may be used in automated workflows as well as interactively from a convenient graphical user interface.

The amplicon sequence analysis is divided into several phases: decomposition of reads into parts (multiplex ID, tag, primer, region of interest), bisulfite-aware mapping and alignment to genomic reference sequences, sorting according to provided options, methylation level calling from sorted alignments, textual or graphical output.

We will present the underlying analysis algorithms, discuss parameters and options and use cases of amplikyzer2 and present examplary results. The software is available at https://bitbucket.org/svenrahmann/amplikyzer/ under the MIT open source license. It is written in Python, partly just-in-time compiled with numba, and conveniently installable with the conda package manager.

References

[1] Leitão, E; Beygo, J; Zeschnigk, M; Klein-Hitpass, L; Bargull, M; Rahmann, S; Horsthemke, B (2018). Locus-Specific DNA Methylation Analysis by Targeted Deep Bisulfite Sequencing. *Methods Mol. Biol.*, 1767:351-366.

[2] Rahmann S, Beygo J, Kanber D, Martin M, Horsthemke B, Buiting K (2013). Amplikyzer: Automated methylation analysis of amplicons from bisulfite flowgram sequencing. *PeerJ PrePrints* 1:e122v2 https://doi.org/10.7287/peerj.preprints.122v2

Fast homology search with COMER

Mindaugas Margelevičius, Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania

Motivation

Sequence homology search, or sequence comparison in general, lies at the heart of computational genomics. As the amount of sequence data steadily grows, existing and new software requires solutions for fast data processing. COMER is one of the most sensitive and accurate computational tools developed for protein alignment and homology search. And this study presents the first results from the development of a fast version of COMER, which will make it greatly attractive for use.

Results

COMER computing performance is increased by harnessing the power of the Graphics processing unit (GPU). While CPU and GPU prices are comparable, a GPU provides several orders of magnitude greater parallelism than can be achieved on a CPU. Still, GPUs have not been utilized to their full power for sequence analysis problems mainly due to data dependencies (e.g., dynamic programming algorithms) and distinctiveness in computer programming for GPUs. This study presents data structures and algorithms for parallel homology search that allow for the maximization of memory and instruction throughput. A new COMER version running on an NVIDIA GeForce GTX 1080 (Max-Q Design) GPU achieved a speedup of greater than a factor of 4000 compared to the previous COMER version (v1.5) running on Intel Xeon CPU E5-2670 v3. The time of the computation of nearly 1e+12 position-specific pair scores, which is 66% of the total computation time, decreased from 328 hours on the CPU to 200 seconds on the GPU.

Conclusion

The results of this study represent one of the fastest implementations of calculations for sensitive protein homology search. Although the new algorithms are developed to perform operations on vectors, they are readily applicable to protein and nucleotide sequences. Therefore, the study offers prospects for accelerating homology search in general.

Funding: This research was funded by the European Regional Development Fund (Grant No. 01.2.2-LMT-K-718-01-0028).

A comparative study of MiSeq and MinION for the classification of 16S rRNA gene sequences

Raf Winand[1], Bert Bogaerts[1], Qiang Fu[1], Julien Van Braekel[1], Maud Delvoye[1], Stefan Hoffman[1], Loïc Lefèvre[1], Nancy H. Roosens[1], Sigrid C. J. De Keersmaecker[1] and Kevin Vanneste[1]

[1] Transversal & Applied Genomics, Sciensano, Brussels (1050), Belgium

Background

A key task of clinical microbiology labs and public health authorities is the accurate and rapid identification of pathogenic bacteria contained in human samples. We compared the feasibility of 16S rRNA sequencing for the classification of organisms, using a mock community reference standard, using two high-throughput sequencing technologies, different software and databases.

Methods and results

From a mock microbial community with known composition, 16S rRNA gene amplicons were generated. Different variable regions were sequenced on the Illumina MiSeq, and the whole gene was sequenced on the Oxford Nanopore MinION. Generated reads for both platforms were classified with Mothur using the NCBI 16S and SILVA databases. MinION sequences were also classified using the proprietary Oxford Nanopore EPI2ME 16S workflow based on the NCBI 16S database, and using GraphMap, a read mapper specifically designed to deal with error-prone reads. Classification of MiSeq reads was highly dependent on the specific amplified region, and generally showed a large number of misclassifications at the species level. Using a majority vote over all sequenced regions, most false positives could be removed at the genus level but not at the species level. Classification of MinION reads by Mothur and EPI2ME resulted in misclassification of less than 1% of sequences at the family and genus level, but ~40% at the species level. Results obtained with GraphMap varied with the database, yielding comparable performance using the NCBI database but demonstrating increasing misclassifications using the SILVA database.

Discussion

Both short and long read high-throughput technologies can be very useful to classify 16S rRNA sequences down to the genus level, but suffer from false positive classifications at the

species level. For short read technologies, the choice of variable region is very important. For long read technologies, a standardized analysis methodology is still lacking.

NJ-networks, turning greedy into all possibilities

Armin Hoenen, Texttechnology Lab, Goethe University, Frankfurt am Main, Germany

Background

The Neighbor Joining algorithm [1] is well used for the computation of phylogenies. Based on a distance matrix, it proceeds in a stepwise manner. At each step one new minimal value is chosen. With two equally minimal distance values, NJ chooses only one of them. Such an event is fairly rare but could lead to different outputs on the same data set from which various methods can produce single networks [2].

Methods and results

We present an implementation of NJ which at each point of encounter of more than one minimum proceeds following all possible ways of computation assembling all equally likely NJ trees on the same distance matrix. Furthermore, we introduce rounding at the distance matrix update steps as a free parameter in order to force minima to converge, so as to be able to produce NJ-networks from any kind of data set. Through a simulation study randomizing large numbers of input distance matrices, we investigate and report the expectable incidence of multiple minima for NJ and estimate the setting of the free parameter when one aims at producing exactly 2 or more trees from one data set.

Discussion

In some biological cases, (e.g. in bacterial and plant evolution), we deal with certain amounts of lateral gene transfer or hybridization. In these cases sometimes exactly 1, 2, .. n such events or a certain proportion may be expectable. Instead of resorting to other algorithms, NJ-networks and the free parameter allow to use the efficiency and accuracy of NJ and produce 2 or more phylogenies by one algorithm e.g. as input to a minimum hybridization network. The rounding procedure achieving this may theoretically correspond to a degree of uncertainty allowing the evolutionary model some flexibility.

References

[1] Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406-425.

[2] Huson, D. H., & Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology*, *61*(6), 1061-10.

A curated knowledgebase on endocrine disrupting chemicals enabling mechanistic insights into systems-level perturbations upon exposure

Janani Ravichandran[#,], Bagavathy Shanmugam Karthikeyan[#], Karthikeyan Mohanraj[$], R.P. Vivek-Ananth[$], Areejit Samal[*]*

The Institute of Mathematical Sciences (IMSc), Homi Bhabha National Institute (HBNI), Chennai, India

*[#] Joint-First authors; [$] Joint-Second authors; * Corresponding authors*

Presenting author: jananir@imsc.res.in; Corresponding author: asamal@imsc.res.in

Background:

Human well-being is affected by exposure to certain chemicals in the environment. Endocrine Disrupting Chemicals (EDCs) are an emerging concern that can perturb the hormonal homeostasis through various mechanisms leading to adverse health effects. Hence, there is immense interest in unraveling the molecular mechanisms via which EDCs perturb the endocrine system. In this direction, we have built a large-scale resource to facilitate research towards mechanistic understanding of systems-level perturbations upon EDC exposure.

Methods and results:

In this work, we have developed a detailed workflow to identify EDCs with supporting evidence of endocrine disruption in published experiments in humans or rodents. Using the workflow, we manually evaluated more than 16000 published research articles to compile 686 potential EDCs along with their endocrine-specific perturbations and dosage information from supporting published literature. Thereafter, we have classified EDCs based on the type of supporting evidence, their environmental source and their chemical properties. Additional information including their chemical structure, physicochemical properties, predicted ADMET properties and target genes were compiled for potential EDCs. The compiled information on EDCs was used to build an open access resource, **D**atabase of **E**ndocrine **D**isr**u**pting **C**hemicals and their **T**oxicity profiles (DEDuCT), accessible at: https://cb.imsc.res.in/deduct/. Using our resource, we have performed a network-centric analysis of EDCs based on their similarity of chemical structure or sets of target genes. Ensuing analysis revealed a lack of correlation between chemical structures and target genes of EDCs.

Discussion:

Our large-scale compilation of the endpoints along with the systems-level perturbations will enhance systems-level understanding of perturbed biological pathways upon EDC exposure.

Reference:

B. S. Karthikeyan[#], **J. Ravichandran[#,*]**, K. Mohanraj[$], R.P. Vivek-Ananth[$], A. Samal*. Science of the Total Environment (2019). DOI:10.1016/j.scitotenv.2019.07.225

Comparison of pathway and GO-network of Salmonella infected HeLa cells.

Jens Rieser, Molecular Bioinformatics, Goethe-University, Frankfurt am Main, Germany
Jörg Ackermann, Molecular Bioinformatics, Goethe-University, Frankfurt am Main, Germany
Ina Koch, Molecular Bioinformatics, Goethe-University, Frankfurt am Main, Germany

Background

Salmonella Typhimurium provokes gastroenteritis and typhoid fever and causes many thousands death every year. A better understanding of the host-pathogen interaction can lead to a better medical therapy. Ubiquitination and Phosphorylation as post-translational modification of proteins are one of the first answers to an infection. The changes of the expression level of phosphorylated proteins in *Salmonella*-infected and uninfected HeLa cells has been investigated by Rogers *et al* [1] and the changes in ubiquitinated proteins by Fiskin *et al* [2].

Methods

We used the proteins of both datasets to search for protein-interactions in three public databases: IntAct, BioGRID and STRING. We combined the data to build a protein-protein interaction network with 1.704 proteins and 18.978 interactions. The data of the ubiquitination and phosphorylation sites were mapped on the nodes.

Results

We have analyzed the network for the basic parameters and clustered the network with the Girvan-Newman algorithm. For each cluster we have performed a GO-enrichment to determine the biological function. For a better overview a GO-interaction network was created and the proteins were mapped on the nodes with their regulated sites. Furthermore, with the information from the Reactome and KEGG database we built a pathway network to investigate the up- and downstream pathways from proteins with highly regulated modification sites.

[1] Rogers, L. D., *et al. Science Signal.* 4.191 (2011): rs9-rs9.
[2] Fiskin, E., *et al*. *Molecular Cell* 62.6 (2016) 967-981.

Topological description of large protein complexes based on cryo-EM data using graph theory

Jan Niclas Wolf, Marcus Keßler, Jörg Ackermann, Ina Koch

Molecular Bioinformatics, Goethe University, Frankfurt am Main

Background

Cryo-EM is capable of yielding atomic resolution structures of protein complexes of more than 99,999 atoms or 62 chains (large structures). The Protein Data Bank (PDB) stores such large structures as macromolecular Crystallographic Information File (mmCIF), because the legacy file format is unsuitable for these. Due to the essential role of large protein complexes in living organisms and the increasing availability of cryo-EM, more and more large structures have been explored. Currently, PDB lists about 900 entries of large structures. No automatic classification tools are available until now.

Methods

We modeled protein structures as Protein Graphs (PGs) and Complex Graphs (CGs). In PGs and CGs, vertices correspond to secondary structure elements and to protein chains, respectively. Edges indicate spatial contacts between vertices. PGs and CGs represent the structure of proteins and protein complexes, respectively, as undirected, labeled graphs. This enables the application of standard algorithms and analysis techniques of graph theory to protein structures. Our Protein Topology Graph Library (PTGL [1, 2]) is an online database of PGs and CGs, including their visualization. Currently, PTGL stores graph-theoretical representations of 143,049 PDB entries. Moreover, PTGL provides a search for predefined and user-defined structure motifs.

Results

We implemented an efficient parser for the mmCIF format, including a fast algorithm for the computation of spatial contacts. We computed PGs and CGs for 715 large structures, from which the largest consists of 2.4 million atoms and 1,356 chains. Exemplary, we analyzed the CG of the human respiratory complex I (PDB ID 5XTH, [3]). A connectivity-based graph clustering, Community Clustering [4, 5], partitioned the CG into the complex's modules and submodules.

[1] T. Schäfer *et al.*, Bioinformatics, 2016

[2] I. Koch & T. Schäfer, Current Opinion in Structural Biology, 2018

[3] R. Guo *et al.*, Cell, 2017

[4] J.H. Morris *et al.*, BMC Bioinformatics, 2011

[5] M.E.J. Newman & M. Girvan, Physical Review E, 2004

Protein Complex Similarity Based on Weisfeiler-Lehman Labeling

Bianca K. Stöcker[1,2], Till Schäfer[4], Petra Mutzel[4], Johannes Köster[1,2,3], Nils Kriege[4] and Sven Rahmann[1,4]

[1]Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

[2]Algorithms for Reproducible Bioinformatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

[3]Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston MA 02215, USA

[4]Computer Science XI, TU Dortmund University, Dortmund, Germany

Background

Proteins in living cells rarely act alone, but instead perform their functions together with other proteins in so-called protein complexes. Being able to quantify the similarity between two protein complexes is essential for numerous applications, e.g. for database searches of complexes that are similar to a given input complex. While the similarity problem has been extensively studied on single proteins and protein families, there is very little existing work on modeling and computing the similarity between protein complexes. Because protein complexes can be naturally modeled as graphs, in principle general graph similarity measures may be used, but these are often computationally hard to obtain and do not take typical properties of protein complexes into account.

Methods and results

Here we propose a parametric family of similarity measures based on Weisfeiler-Lehman labeling. We use a convex combination of two Jaccard coefficients (Weisfeiler-Lehman label multiset after zero and one iteration) and evaluate it on simulated complexes of the extended human integrin adhesome network. We show that the defined family of similarity measures is in good agreement with edit similarity, a similarity measure derived from graph edit distance, but can be computed more efficiently. It can therefore be used in large-scale studies and serve as a basis for further refinements of modeling protein complex similarity.

Stöcker BK, Schäfer T, Mutzel P, Köster J, Kriege N, Rahmann S. 2018.

Protein complex similarity based on Weisfeiler-Lehman labeling. PeerJ Preprints 6:e26612v1. International Conference on Similarity Search and Applications (SISAP) 2019, accepted for publication.

C-Jun drives melanoma progression in *PTEN* wild type melanoma cells

Melanie Kappelmann-Fenzl[1,2], *Claudia Gebhard*[3,4], *Alexander O. Matthies*[1], *Silke Kuphal*[1], *Michael Rehli*[3,4], *Anja Katrin Bosserhoff*[1,5,]

[1]*Institute of Biochemistry (Emil-Fischer Center), Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany*

[2] *Faculty of Applied Health Care Sciences, University of Applied Science Deggendorf, Germany*

[3]*Department of Internal Medicine III, University Hospital Regensburg, Regensburg, Germany*

[4]*Regensburg Center for Interventional Immunology (RCI), c/o University Hospital Regensburg, Regensburg, Germany.*

[5]*Comprehensive Cancer Center (CCC)-EMN, Erlangen, Germany*

Melanoma is a highly aggressive type of cancer, and its incidence is growing faster than any other cancer entity. In recent decades, many altered pathways regulating the development and progression of melanoma and the high migratory and invasive potential of melanoma cells have been identified, but a detailed molecular understanding of this disease is largely lacking.

One crucial step in melanoma development and progression is the deregulation of cancer-supporting transcription factors, especially AP-1 (activating protein-1) transcription factors, including *c-Jun*. Due to the critical impact of active *c-Jun* in melanoma, it is important to define its target genes and to identify and ultimately inhibit oncogenic signals. In this study we mapped the genome-wide occupancy of the AP-1 family member *c-Jun* in different melanoma cells and correlated AP-1 binding (ChIP-Seq) with transcriptome data (RNA-Seq) to identify genes in melanoma regulated by *c-Jun*. Our analysis shows that *c-Jun* supports the malignant phenotype by deregulating genes in cancer-relevant signaling pathways, such as mitogen-activated protein kinase (MAPK) and phosphatidylinositol-3-kinase (PI3K) pathways. Moreover, we detected a coexpression of *c-Jun* and the tumor suppressor *PTEN*. Our study reveals that activation of *c-Jun* overrules the tumor suppressive effect of *PTEN* in early melanoma development. These findings help to understand the relevance of *c-Jun* within cancer pathways in different melanoma cell types, especially in relation to MAPK and PI3K pathways, which are commonly deregulated in melanomas. Consequently, targeting *c-Jun* in *PTEN*$^{+}$ melanoma cells may

represent a promising therapeutic strategy to inhibit survival of melanoma cells to prevent the development of a metastatic phenotype.

Taken together, our study confirms a crucial role of the transcription factor *c-Jun* in melanoma development and progression in *PTEN*$^{+}$ melanoma cells to overcome apoptosis and promote malignancy not only *in vitro* but also *in vivo*. TCGA data analysis by cBioPortal confirm and support our latest findings. Thus, the detection of *c-Jun* and *PTEN* coexpression in melanoma represents a promising diagnostic marker for highly aggressive melanoma cells with the potential to result in metastasis with the loss of both *c-Jun* and *PTEN*.

Deciphering the mechanism of drug resistance in Smoothened receptor: An *in silico* perspective of protein resistivity and specificity

Noopur Sinha [1,2], Saikat Chowdhury [1,2]; Ram Rup Sarkar[1,2]

1. Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Pune, India, 2. Academy of Scientific & Innovative Research (AcSIR), CSIR-NCL Campus, Pune, India;

Background: The Smoothened (SMO) antagonists Vismodegib effectively inhibit the Hedgehog signaling pathway in the proliferating cancer cells. In early–stage of treatment, Vismodegib exhibited promising outcomes to regress the tumors cells, but relapsed due to the drug resistive primary (G497W) and acquired (D473H/Y) mutations in SMO. This study aims to investigate the uncharted insights into the structural and functional mechanism hindering the Vismodegib binding with the mutant variant SMO (SMO^{Mut}). A previously reported inhibitor ZINC12368305; known for making energetically favourable complex with SMO^{WT}, is chosen as a control to assess important factors governing the formation of stable complex with SMO^{Mut}.

Methods and results: Network propagation using heat diffusion principles was applied to identify the modules of amino acid residues which are influenced by the individual mutations and causes unfavorable binding of Vismodegib with SMO^{Mut}. It also identified the distantly located residues from the binding–pocket, responsible for allosterically modulating the drug resistant in SMO^{Mut}. Apart from these, the basic understandings of the structural properties of protein–inhibitor complex conformations and its impact on binding affinity were examined with the help of Molecular dynamics simulation and MM–PBSA method. Furthermore, the second generation inhibitor ZINC12368305 is also depicted as the potential inhibitor of both the $SMO^{WT/Mut}$, which could be used individually or in combination with Vismodegib for the suppression of SMO^{Mut}.

Discussion: These observations could be useful in deciphering the structural mechanisms and the complex interconnections of the intermediate residues regulating the allosteric modulation of SMO. The proposed novel computational framework could also be useful for assessing the potential allosteric sites of other proteins and could be helpful in developing second generation inhibitor molecules.

Struo: a pipeline for building custom databases for common metagenome profilers

Jacobo de la Cuesta-Zuluaga, Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen, Germany; Ruth E. Ley, Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen, Germany; Nicholas D. Youngblut Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen, Germany

Background: Metagenome profiling is the most efficient method of obtaining comprehensive taxonomic and functional data from metagenomes, yet default databases accompanying metagenome profilers are not updated at a pace that reflects the rapid increase in microbial genomics data. The creation of updated comprehensive, custom databases is cumbersome due to the complexity and high computational requirements of retrieving the genomes, and configuring and executing the software. As a result, many metagenomic analyses fail to include the most up to date microbial data, missing critical insights. We address this with the development of *Struo*, an automatized and modular pipeline that assists in the retrieval of genomes and construction of databases for metagenome profilers.

Methods and results: *Struo* uses Snakemake and Conda to unify the workflow and build databases in a straight-forward, reproducible manner on Unix-based high-performance compute clusters. Currently, *Struo* supports Kraken2, Bracken2 and HUMANn2, and can be extended to include other tools. Publicly available or novel genomes can be used; here, we used *Struo* with 21,276 representative genomes of the GTDB to generate databases that broadly encompass known microbial diversity. This resulted in an increase of 25% more reads mapped from simulated and real metagenomes compared to default profiler databases.

Discussion: A carefully curated and tailored selection of genomes to be included in reference databases for metagenome profiling facilitates the exploration of microbiomes by increasing the fraction of reads mapped to a known reference. *Struo* empowers researchers to incorporate previously unexplored taxa in the study of hidden microbial diversity. *Struo* and the custom databases will be made public as open source resources.

Curatopes Melanoma: A database of T cell epitopes from overly expressed proteins in metastatic cutaneous melanoma

Christopher Lischer, Department of Dermatology, Universitätsklinikum Erlangen and Faculty of Medicine, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Martin Eberhardt, Department of Dermatology, Universitätsklinikum Erlangen and Faculty of Medicine, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Tanushree Jaitly, Department of Dermatology, Universitätsklinikum Erlangen and Faculty of Medicine, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Cornelia Schinzel, Department of Dermatology, Universitätsklinikum Erlangen and Faculty of Medicine, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Niels Schaft, Department of Dermatology, Universitätsklinikum Erlangen and Faculty of Medicine, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Jan Dörrie, Department of Dermatology, Universitätsklinikum Erlangen and Faculty of Medicine, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Gerold Schuler, Department of Dermatology, Universitätsklinikum Erlangen and Faculty of Medicine, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Julio Vera, Department of Dermatology, Universitätsklinikum Erlangen and Faculty of Medicine, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Therapeutic anti-cancer vaccination has been adapted as an immunotherapy in several solid tumors. However, the selection of promising candidates from the total quantity of possible epitopes poses a challenge to clinicians and bioinformaticians alike, and very few epitopes have been tested in experimental or clinical settings to validate their efficacy. Here, we present a comprehensive database of predicted non-mutated peptide epitopes derived from genes which are overly expressed in a group of 32 melanoma biopsies compared to healthy tissues, and which were filtered against expression in a curated list of survival-critical tissues. We hypothesize that these "self-tolerant" epitopes have two desirable properties: they do not depend on mutations, being immediately applicable to a large patient collective, and they

potentially cause fewer auto-immune reactions. To support epitope selection, we provide an aggregated score of expected therapeutic efficiency as a shortlist mechanism. The database has applications in facilitating epitope selection and trial design and is freely accessible at https://www.curatopes.com. Curatopes systematically predicts and scores anti-tumor T cell epitopes with a focus on tolerability and the avoidance of severe auto-immunity, offering a supplementary epitope set for further investigation in immunotherapy.

Natrix: A snakemake-based workflow for the processing, clustering and taxonomic assignment of amplicon sequences

Marius Welzel, Heiderlab, Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany, Anja Lange, Bioinformatics and Computational Biophysics, University of Duisburg-Essen, Essen, Germany, Manfred Jensen, Biodiversity, University of Duisburg-Essen, Essen, Germany, Jens Boenigk, Biodiversity, University of Duisburg-Essen, Essen, Germany, Daniela Beisser, Biodiversity, University of Duisburg-Essen, Essen, Germany

Prokaryotes and microbial eukaryotes constitute a large fraction of the biodiversity on earth, still in many environments their diversity and distribution is yet unkown. Sequencing of marker genes amplified from environmental samples has allowed to resolve some of the hidden diversity and elucidate evolutionary relationships and ecological processes among the microorganisms. The analysis of large sample numbers at high sequencing depths generated with recent Illumina sequencing technology requires efficient, flexible and reproducible bioinformatics pipelines.

Few existing workflows include all necessary analysis steps starting from raw sequencing reads up to taxonomically assigned OTUs and can be run user-friendly, scalable and reproducible on different computing devises using an efficient workflow management system.

Here we present Natrix, an open-source bioinformatics workflow for the preprocessing of raw amplicon sequencing data. The workflow contains all analysis steps from the quality assessment over read assembly, dereplication, chimera detection, split-sample merging, OTU-generation, to the taxonomic assignment of OTUs. The pipeline is written in Snakemake, a workflow management engine for the development of data analysis workflows. Snakemake ensures reproducibility of a workflow by automatically deploying dependencies of workflow steps (rules) and scales seamlessly to different computing environments like servers, computer clusters or cloud services. In addition conda environments are used for version control of the utilised programs. The workflow contains separate rules for each step and each rule that has additional dependencies has a separate conda environment that will be automatically created when starting the workflow for the first time. The encapsulation of rules and their dependencies allows for hassle-free sharing of rules between workflows and easy adaptation and extension of existing workflows.

PIC – a basic and extendable Petri net App for Cytoscape

Marcel Gehrmann[(1)]; Marius Kirchner[(1)]; Jens Rieser[(1)]; Jörg Ackermann[(1)]; Ina Koch[(1)]

(1) Molecular Bioinformatics, Goethe-University Frankfurt am Main, Germany

Background

An important part of systems biology comprises the analysis of both the static and dynamic behavior and properties of biological systems. One way to model these networks is by using Petri nets, which are specialized for representing dynamic networks at different scales. Cytoscape is an established tool for analyzing and visualizing biological networks. It is an open source application, enabling developers to implement their own apps to extend Cytoscape based on their needs. PIC is one such newly developed app for Cytoscape, integrating basic Petri net functionalities.

Methods and Results

PIC allows users to import Petri nets from several formats, among them SBML, pnt, apnn and metatool. Alternatively, users can edit a new Petri net within Cytoscape, using PIC's internal functionality or Cytoscape. PIC also can verify the Petri net. Users can simulate the dynamic behavior according to the defined firing rule with the results being visualized within the application. PIC is also capable of importing and visualizing transition invariants as well as checking them for their realizability for the current marking of the Petri net. For small networks, PIC can also compute the transition invariants by itself.

Discussion

Overall PIC enables analysis, editing and visualization of Petri nets directly in Cytoscape. As it is based on Cytoscape's API, one can easily extend it with additional functionalities and possibilities for analysis and visualization. Future updates could include computation of place and manatee invariants, logical places and further improvements.

[1] I. Koch, W. Reisig, and F. Schreiber. Modeling in systems biology: the Petri net approach, volume 16. Springer Science & Business Media, 2010.

[2] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software environment for integrated models

of biomolecular interaction networks. Genome Research, 13(11):2498-2504, 2003.

Translated ORF prediction in bacteria - A comparison of available machine learning approaches

Christopher Huptas, Zachary Ardern and Siegfried Scherer

Chair of Microbial Ecology, Technical University of Munich, Freising, Germany

Background

Ribosome Profiling (Ribo-Seq) is an experimental strategy enabling the deep sequencing of small ribosome-protected mRNA footprints and has become a powerful tool to measure the genome-wide translational landscape *in vivo*. In recent years, several machine learning approaches were developed to delineate the exact boundaries of translated ORFs, each utilizing different signatures inherent to ribosome profiling information. No comprehensive performance comparison of the available approaches has been performed so far.

Methods and results

Several publicly available N-terminomics and proteogenomics datasets were used to judge the prediction performance of REPARATION, DeepRibo and the general approach introduced by Giess *et al.* (herein named 'TIS-Predictor' for simplicity) on corresponding Ribo-Seq data from *Escherichia coli* K12 MG1655 and *Salmonella* Typhimurium SL1344. In the majority of cases, TIS-Predictor yields the highest precision, while REPARATION is the most sensitive tool. However, considering both precision and recall (F_1 score) together, we show that DeepRibo usually performs best. In the cases where a false positive isoform is predicted by any of the tools, the 'true' translation initiation site (TIS) can most often (between 85.0 % and 97.3 % of cases) be found within the top-3 predictions of the isoform family when switching to multiple-start-site prediction mode.

Discussion

The comparisons performed provide a first snapshot of each tool's prediction performance. Extending the repertoire of Ribo-Seq and proteomics data considered during comparison will provide more detailed insights into the strengths and weaknesses of each tool. Rather than finding a 'super tool' which is the best overall, we advocate a pluralistic approach where different tools will be more or less useful for different scientific questions. Computational biologists in the field can make use of our findings to tune already existing approaches or develop superior ones.

metagenomic read and EVE classification with Deep Learning

Florian Mock, Friedrich Schiller Universitaet, Jena/ Deutschland
Adrian Viehweger, Friedrich Schiller Universitaet. Jena/ Deutschland
Manja Marz, Friedrich Schiller Universitaet. Jena/ Deutschland

Abstract

Identification and separation of viral, bacterial, archaeal and eukaryotic reads in metagenomic datasets is an ongoing challenge. The identification, especially of unknown species without a related reference genome, is problematic today [1]. For many species, the three-dimensional structure of their encoded proteins is conserved, even though their sequence identity is low. However, it is computationally infeasible to predict and compare 3D protein structure for all sequences in a metagenomic dataset [2].

We approached this problem by using transfer learning. Transfer learning is a powerful technique whereby neural nets (NN) are trained on one task for which there is numerous data and are then trained on another problem for which there is little data. This enables the NN to focus on the task of interest in the second step, using a priori knowledge gained on the first. We used previously trained models that can represent an amino acid sequence as a series of numerical vectors. These vectors are called embeddings. Each dimension in the embedding encodes a different property of the amino acid its encodes for. This protein embeddings have are a very realistic, powerful, fast representation of the proteins helping predict several different properties of the amino acid [3].

We trained a recurrent NN to classify the superkingdom of a protein from the protein embedding. The quality of the prediction is relatively high with an accuracy of over 70%, which is nearly three times higher than expected by chance (25%). Furthermore is this approach multiple orders faster than using blast. In the future, it should be possible to identify reads of interest while sequencing with this approach. Therefore we would all six different frames to translate a read into its possible protein. Then predict the superkingdom of the most promising frame. Furthermore, it should be possible to identify potential endogenous viral elements (EVEs) in the human genome by a sliding window approach where all viral classified hits are potential EVEs.

[1] Mande, Sharmila S., et al. "Classification of metagenomic sequences: methods and challenges." *Briefings in bioinformatics* 13.6 (2012): 669-681.

[2] Waterhouse, Andrew, et al. "SWISS-MODEL: homology modelling of protein structures and complexes." *Nucleic acids research* 46.W1 (2018): W296-W303.

[3] Rives, Alexander, et al. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." *bioRxiv* (2019): 622803.

PTKTool: Mass spectrometry-based quality control system by reference peptides

Sang-Yoon Kim, Antoine Lesur, Francisco Azuaje, Petr Nazarov, Gunnar Dittmar

Luxembourg Institue of Health, Strassen, Luxembourg

Background

Maintaining high quality of lab instrumentation is an essential procedure for modern analytical experiments focused on obtaining reliable, reproducible and comparable results overtime. Particularly in a proteomics laboratory, the demand for quality control (QC) has been rising constantly due to an increase of the instruments sensitivity and the number of high-throughput experiments. Here, we present PTKTool, a fully automated mass spectrometry (MS)-based QC system that monitors MS instruments and high-throughput acquisition efficiently by analyzing reference peptides (Pierce Retention Time Calibration Mixture; PTK15, Thermo Fisher Scientific).

Methods and results

The PTKTool extracts precursor (MS1) and fragment ion (MS2) of the reference peptides, and then analyses peptide and peak properties in multiple aspects (i.e. peptide fragmentation patterns, elution time, mass accuracy, peak area, peak asymmetry etc.). Our system consists of two main components that allow reporting for each QC experiment in detail, as well as visualizing results over the time to pinpoint aberrant experiments effectively. We have developed PTKTool using PHP, SQLite and Apache in the back-end, while modern web frameworks such as bootstrap and D3 JavaScript libraries were used in the front-end for visualizing results interactively. This system is fully compatible with any operating systems such Linux and Windows. Python3 and MSFileReader libraries (Thermo Fisher Scientific) are required in order to read and extract information from the .raw files.

Discussion

PTKTool is currently functional with Thermo Scientific mass spectrometry, but is open to any reference peptides. Thus, this system can be easily and efficiently implemented in daily routine peptide quantification pipelines.

A Novel *in Silico* Precision Medicine Modelling Tool to Understand Tumor-associated Pathways and Molecular Treatment Decisions

Meik Kunz, Chair of Medical Informatics, Friedrich-Alexander University of Erlangen-Nürnberg, Erlangen, Germany; presenting author

Maximilian Fuchs, Functional Genomics and Systems Biology Group, Department of Bioinformatics, University of Würzburg, Germany

Ralf Bargou, Department of Internal Medicine II, Chair of Translational Oncology, University Hospital Würzburg, Würzburg, Germany

Thomas Dandekar, Functional Genomics and Systems Biology Group, Department of Bioinformatics, University of Würzburg, Germany

Hans-Ulrich Prokosch, Chair of Medical Informatics, Friedrich-Alexander University of Erlangen-Nürnberg, Erlangen, Germany

ABSTRACT

Targeted therapy opens new options for cancer treatment (1). However, the success rates are very disappointing. Thus, it is of clinical importance to simulate the response of the intended treatment to find the optimal therapy and prevent treatment resistance (1,2).

We developed an *in silico* modelling tool combining sequencing analysis with machine learning and signal network modelling that allows to understand individual treatment responses and predict optimal therapeutic strategies (3,4). The modelling tool reflects the logical connectivity of the involved patient mutations and cancer-associated signaling cascades in a simplified view, in which a machine learning approach predicts risk models with valuable clinical outcomes (3,5). The advantage of the tool is that we can individually adjust it to specific patient mutations and treatments, highlighting its relevance for cancer research and precision medicine (3). Our modeling tool is used in the molecular tumor board of the comprehensive cancer center Mainfranken. We exemplified it for patients with lung, adrenocortical and oral cancer that did not respond to clinical therapies. We simulated the different treatments according to the patient mutational background including relevant

resistance mechanisms. Subsequently, the *in silico* drug target screening calculates the optimal targeted therapy (3,6).

In conclusion, the *in silico* cancer modelling tool helps to understand tumor-associated pathways and can be used to pre-evaluate the individual molecular treatment response. This allows to identify optimal patient-tailored treatment strategies and paves the way to support therapy decisions in the clinic.

References

(1) Ferlay J, et al. (2015). Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11, http://globocan.iarc.fr, Int J Cancer; 136(5):E359-86. doi: 10.1002/ijc.29210.

(2) Sanchez-Vega F, et al. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell.; 173(2):321-337.e10. doi: 10.1016/j.cell.2018.03.035.

(3) Göttlich C*, Kunz M*, et al. (2018). A combined tissue engineered/in silico signature tool for patient stratification in lung cancer. Mol Oncol.; doi: 10.1002/1878-0261.12323.

(4) Kunz M, et al. (2019). A comprehensive method protocol for annotation and integrated functional understanding of lncRNAs. Briefings in Bioinformatics; doi: 10.1093/bib/bbz066

(5) Schweitzer S*, Kunz M*, et al. (2018). Plasma steroid metabolome profiling for the diagnosis of adrenocortical carcinoma. *Eur J Endocrinol.*, pii: EJE-18-0782.R1. doi: 10.1530/EJE-18-0782.

(6) Kunz M, et al. (2016). The drug-minded protein interaction database (DrumPID) for efficient target analysis and drug development. Database (Oxford); pii: baw041, doi: 10.1093/database/baw041.

Data mining and machine learning approaches to investigate the biological roles of established biomarkers and their related cancer drivers

Raheleh Sheibani Tezerji, Ludwig Boltzmann Institute Applied Diagnostics, 1090, Vienna, Austria; Gerda Egger, Department of Pathology, Medical University of Vienna, 1090, Vienna, Austria and Ludwig Boltzmann Institute Applied Diagnostics, 1090, Vienna, Austria

One of the challenges in medical sciences is how to translate scientific findings into better clinical results. Information and data are the main input for such a translational process. In the era of postgenomics, public databases such as the Cancer Genome Atlas (TCGA), have contributed to the development of various Omics data sets from 33 cancer types. Omics analysis of these data have facilitated the discovery of effective cancer biomarkers. Aside from biomarker discovery, access to such a big database provides significant insight into understanding biological changes, modifications and functions of established biomarkers in certain patients. The purpose of this study is to provide an independent analysis of public cancer databases from primary or metastatic tumor samples to investigate the biological roles of known biomarkers or novel driver genes and add perspectives to biomarker discovery and application.

In our study, first, we relied on a top down approach based on a highly relevant prostate cancer biomarker, the prostate specific membrane antigen (PSMA), which has shown a promising performance in nuclear imaging of primary and advanced prostate cancer (PCa) as a prognostic marker and therapeutic target. Comparing PSMA high versus PSMA low primary tumors based on RNA-Seqencing analyses of 497 primary PCa samples from the TCGA dataset, we identified a significant difference in gene expression patterns of the two groups. The differentially expressed genes were associated with key oncogenic and metabolic pathways, highlighting the biological significance of PSMA expression for PCa. Based on these findings, we will use computational and systems biology approaches to assess proteomic and epigenomic signatures of PSMA positive versus PSMA negative tumors and their relevance for tumor biology.

As a next step, we will apply an ensemble classifier (EC) machine learning method,
for the discovery of cancer driver candidates (mutations) in primary and metastatic prostate cancer samples from TCGA and dbGAP. The accuracy and performance of the method will be estimated by statistical validation techniques. Top driver genes and novel targets will be predicted, which could be used as potential candidate biomarkers for PSMA

positive PCa, based on biological classification. These top driver genes can be evaluated for biological and molecular pathways that are statistically enriched in PCa.

We expect to identify key tumor driving pathways and metabolic characteristics in a defined group of PCa, associated with high levels of PSMA expression. Our results will address basic scientific questions in tumor biology, but the concept and expected results head towards translational research. These findings might have high translational impact for personalized medicine to construct predictive models from early diagnosis to monitoring prognosis and predict response to treatments.

Substrate prediction for membrane transporters by machine learning

Andreas Denger, Chair of Computational Biology, Saarbrücken, Germany; Volkhard Helms, Chair of Computational Biology, Saarbrücken, Germany

Background

Membrane transport proteins are needed for numerous tasks in any living organism, including energy production, cell metabolism, signal transduction and immune response. With the increasing amount of sequence data available due to developments in next-generation sequencing, and with experimental methods for transporter substrate annotation being costly and time-consuming, computational methods for transport protein substrate prediction are needed dearly.

Methods and results

A machine learning pipeline based on support vector machines (SVMs) was developed to predict substrates of membrane transporters in three organisms. Four different feature generation algorithms were optimized and combined into a meta-approach, which was evaluated using stratified cross-validation. The program was able to distinguish amino acid-, protein-, sugar-, and electron-transporters with a Matthews correlation coefficient (MCC) of 0.89, 0.83 and 0.95 for *E.coli, S.cerevisiae* and *H.sapiens,* respectively. The average pairwise binary classification scores (MCC) for the substrate classes *amino acid, anion, cation, electron, protein, sugar* and *other* were 0.93 for *E.coli*, 0.94 for *S.cerevisiae* and 0.91 for *H.sapiens.*

Discussion

Machine learning is a suitable method for the distinction of membrane transporters based on their substrates. By generating features from datasets for protein sequences, evolutionary information, gene expression, gene ontology and genomic neighborhood, the program was able to classify transporters into up to seven substrate classes with high accuracy.

A multipurpose web service for protein analysis based on the remote homology search method COMER running on GPU

Justas Dapkūnas and Mindaugas Margelevičius, Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania

Motivation

Community-wide CASP experiments revealed that the most reliable approach to protein structure prediction is based on homology to known proteins. Similarly, protein function annotation also benefits from homology-based inference. However, despite recent advances, homology detection itself needs improvement. Moreover, sequence databases are growing steadily, demanding faster search methods. Here we present the ongoing development of a web service with a new version of the remote homology search method COMER at its core, running on the Graphics processing unit (GPU).

Results

The new web service will have a number of useful features. It aims to facilitate homology search across popular protein databases (PDB, SCOP, PFAM, and others) using highly sensitive and accurate homology search and alignment method COMER. The user is allowed to construct and inspect multiple sequence alignments based on COMER output and to generate homology models of protein structures. The web service will provide a RESTful application programming interface (API) for conducting searches remotely on a large scale. This feature is catalyzed by a newly developed COMER version running on multiple GPUs and a dedicated computer server with powerful GPUs (NVIDIA Tesla V100 architecture), enabling large-scale homology search to be performed within minutes.

Conclusion

The web service presented here has two key features: it represents an interface to sensitive and accurate protein homology analysis and provides means to perform it very fast. With these main features, the new web service will allow researchers to benefit from remotely accessible high-performance homology search and analysis.

Funding: This research was funded by the European Regional Development Fund (Grant No. 01.2.2-LMT-K-718-01-0028).

Analysis of Differential Interactome in Colorectal Adenocarcinoma to Identify Candidate Biomarkers and Therapeutics

Hande BEKLEN, Department of Bioengineering, Marmara University, Istanbul, Turkey; Gizem GULFIDAN, Department of Bioengineering, Marmara University, Istanbul, Turkey; Beste TURANLI, Department of Bioengineering, Istanbul Medeniyet University, Istabul, Turkey; Pemra OZBEK SARICA, Departmant of Bioengineering, Marmara University, Istanbul, Turkey; Kazim Yalcin ARGA, Department of Bioengineering, Marmara University, Istanbul, Turkey

Background: Colorectal cancer which affects the large intestine in the digestive tract is one of the most lethal types of cancers commonly seen in both men and women. The high heterogeneity of colorectal cancer leads to difficulties explaining the biology and behavior of this cancer. The aim of this study is to identify prognostic biomarkers and potential therapeutics for colorectal cancer using the protein interactions differentiated among healthy and tumor groups.

Methods and results: The differential protein-protein interactions (dPPI) were identified by using the differential interactome algorithm published by our research group before. For this purpose, two independent data sets were obtained from The Cancer Genome Atlas (TCGA) containing 644 tumor samples and 51 normal samples and Gene Expression Omnibus (GEO) containing 32 tumor samples and 32 normal samples. As a result of differential interactome analysis, significant dPPIs were determined (2434 in GEO data set, 1619 in TCGA data set) and highly interacting protein modules were identified. Principal component analysis for diagnostic purpose and Kaplan-Meier analysis for prognostic purpose were performed for each module. For the common modules in both data sets, 16 modules were found having diagnostic potential, while 6 modules were found having prognostic potential. In addition, common significant dPPIs in both data sets were observed in point of drug repositioning and 6 dPPIs and 13 drug targets for these interactions were identified. By using molecular dynamics simulations, root-mean-square deviation and root-mean-square fluctuation were taken as performance metrics to perform further investigation in vitro cell culture.

Discussion: This study will shed light on the identification of specific biomarkers and drug targets for early detection, disease progression and accurate treatment of colorectal cancer. ***This study was supported under FEN-C-1206-0199 project.***

Cost-optimal assignment of elements in genome-scale multi-way bucketed Cuckoo hash tables

Jens Zentgraf, Computer Science XI, TU Dortmund, Dortmund, Germany;
Sven Rahmann, Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, Essen, Germany

We present the first practical algorithm to solve the minimum cost assignment problem for multi-way bucketed Cuckoo hashing [1] with $h \geq 2$ hash functions [2] and buckets that store $b \geq 1$ elements each [3]. We minimize the average lookup cost over all stored elements, assuming that an element in the bucket indicated by its j-th hash function incurs a lookup cost of j cache misses.

Our method is based on a combination of the Bellman-Ford and Hopcroft-Karp algorithms [4, Chapter 24.1, Chapter 26.6] for finding minimum cost paths in the Cuckoo assignment graph, using multiple sources in parallel and a limited number of iterations. It consists of an *initialization phase*, which greedily assigns as many elements as possible (without moving any element) to their first or second bucket. It then proceeds in several *iterations*, assigning a (ideally large) subset of the still unassigned elements. It never unassigns already assigned elements, but it may move elements to different buckets during an iteration. Each iteration consists of *two phases*. In the first phase, cost-optimal paths between buckets with free slots and unassigned elements are found in parallel. In the second phase, elements are moved along disjoint paths.

We find a cost-optimal assignment for the 2.38 billion unique DNA 25-mers of the human genome in 4–8 hours of CPU time at a load level of 95% for bucket sizes 3–8, using slightly over 40 GB of RAM. For bucket size $b = 4$, we obtain optimal costs of approximately 1.2 cache misses per stored element and achieve ≤ 1.4 for non-existing elements.

References

[1] Rasmus Pagh and Flemming Friche Rodler. Cuckoo hashing. *J. Algorithms*, 51(2):122–144, 2004.
[2] Dimitris Fotakis, Rasmus Pagh, Peter Sanders, and Paul G. Spirakis. Space efficient hash tables with worst case constant access time. *Theory Comput. Syst.*, 38(2):229–248, 2005.
[3] Martin Dietzfelbinger and Christoph Weidling. Balanced allocation and dictionaries with tightly packed constant size bins. *Theor. Comput. Sci.*, 380(1-2):47–68, 2007.
[4] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. Introduction to Algorithms, *3rd Edition*. MIT Press, 2009.

Of gene expression and cell division time – a mathematical framework for advanced differential gene expression and data analysis

Katharina Baum, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Germany, and Luxembourg Institute of Health, Luxembourg; Johannes Schuchhardt, MicroDiscovery GmbH, Germany; Jana Wolf, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Germany; Dorothea Busse, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Germany

Background. Estimating fold-changes of average mRNA and protein molecule counts per cell is the most common way to perform differential expression analysis. However, these high-throughput population-based gene expression data may be affected by cell division. Due to this often-neglected phenomenon, caution is required when concluding on altered gene expression characteristics from differences in these data.

Methods and results. Based on a mathematical model, we develop a quantitative framework that links population-based mRNA and protein measurements to rates of gene expression in single cells undergoing cell division. The derived equations are easy-to-use and widely robust against biological variability. They integrate multiple 'omics' data into a coherent, quantitative description of single-cell gene expression and improve analysis when comparing systems or states with different cell division times. We explore these ideas in the context of resting vs. activated B cells.

Discussion. Applying our framework by analyzing differences in protein synthesis rates enables to account for differences in cell division times. Our example of B cell activation demonstrates that this improves the resolution and hit-rate of differential gene expression analysis when compared to analyzing population protein abundances alone.

Can exosomal RNA serve as biomarker for renal clear cell carcinoma?

Konrad Grützmann[1,2], Susanne Füssel[3], Alexander Krüger[1,2], Andrea Lohse-Fischer[3], Falk Zakrzewski[1,4], Barbara Klink[1,2,5], Daniela Aust[1,4], Evelin Schröck[1,5]

1 Core Unit for Molecular Tumor Diagnostics (CMTD) at the National Center for Tumor Diseases (NCT) Dresden, Germany; 2 German Cancer Consortium (DKTK), Heidelberg, Germany; German Cancer Research Center (DKFZ), Heidelberg, Germany; 3 Department of Urology, University Hospital Carl Gustav Carus, Dresden University of Technology, Dresden, Germany; 4 Institute for Pathology, University Hospital Carl Gustav Carus Dresden, Germany; 5 Institute for Clinical Genetics, University Hospital Carl Gustav Carus Dresden, Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Germany

correspondence: konrad.gruetzmann@uniklinikum-dresden.de

Background

Clear cell renal cell carcinoma (ccRCC) is the 6th/11th most common cancer in male/female and ranks on 13th place of cancer-related deaths in Germany. Many ccRCC are in an advanced state when diagnosed and one third has already metastasized. There is a clear need for early detection markers for better patient treatment. Liquid biopsies from urine and blood have come into focus as potential non-invasive diagnostic tools for cancer. Besides cell-free DNA, such fluids carry cell-derived vesicles like exosomes, which harbor RNA, DNA and proteins of the host cell. While their biological role is not yet fully understood, they are thought to have intercellular communication function. It was observed that exosomes from tumor tissue have molecular profiles distinct from those originating from healthy cells. Since urinary exosomes were described to carry RNA biomarkers for renal disease, their usefulness for the detection of ccRCC should be evaluated in this study. For this purpose, we applied next generation sequencing of RNA from urine-derived exosomes of ccRCC patients to evaluate their potential as diagnostic markers.

Methods

Exosomes were extracted from urine of 66 ccRCC patients and 36 controls with urolithiasis using miRCURY Exosome Isolation Kit (Exiqon/Qiagen) and validated via

presence of typical exosomal proteins and vesicle size distribution. RNA was extracted with the miRCURY RNA Isolation Kit (Exiqon/Qiagen), prepared with the SMARTer smRNA-Seq Kit (Takara) and single-end sequenced on an Illumina NextSeq 500. After quality and adapter trimming (Trimmomatic), reads were mapped against the human genome (STAR) and unambiguous reads from known coding and non-coding transcripts were counted (Rsubsread/featureCounts).

Results

Due to a small effective read lengths (33nt on average), only 24% of the trimmed reads could be mapped uniquely onto the genome. Despite deeper sequencing (in median $22x10^6$ raw reads per sample), relatively few reads could be found for known transcripts. A large fraction of the reads (measured in RPKM) originated from miRNA (37%) and miscRNA (42%), while reads from protein-coding (5.8%) and rRNA (5.3%) were less frequent. Principle component analysis and hierarchical clustering of expression profiles showed that there was a big variance in expression between the samples, which was in part due to the low read counts. Tumor and control samples could only marginally be separated using global expression profiles. However, we found 405 genes that were significantly less represented in exosomes in ccRCC patients and 40 genes that were significantly more strongly represented. Among them were 409 protein-coding genes, one miRNA (MIR142) and two lincRNAs (EIF3J-AS1, RNU12). The magnitude of regulation was relatively low. Only 32 genes had an absolute log-2 fold-change above one. However, six genes showed a potential for ccRCC diagnostics (GGA2, RP11-661A12.9, CC2D1B, ECH1, RP5-837M10.1, ZHX2) with an area under the curve (AUC) of 0.553 – 0.732.

Conclusion

We showed that it is possible to apply next generation sequencing on RNA extracted from urinary exosomes. Very short sequence reads from this particular biomaterial and, hence, ambiguous mapping pose a challenge for expression profiling from urine-derived exosomes. However, some potential expression markers for diagnosis of ccRCC were found.

Impact of abnormal feeding and rearing conditions on gut microbiome composition in piglets

Vitaly Belik[1] *and Robert Pieper*[2]

[1] *System Modeling Group, Institute for Veterinary Epidemiologie and Biostatistics, Freie Universität Berlin, Germany*

[2] *German Federal Institute for Risk Assessment, Berlin, Germany*

Gut microbiome composition might serve as an important health indicator in animals and humans. In particular, gut microbiome composition at an early stage of life might serve as a primer for the development of immune system. How the microbiome develops in neonatal animals is still poorly understood. We analyze the evolution of the gut microbiome composition on species levels in young piglets (from 1 to 14 days old). We compare the gut microbiome composition of normally fed (suckling) piglets with piglets subject to modifications in their feeding (e.g. fed with formula instead of sow milk) and isolation from sow animals. This results in four group of animals in total. We apply complex network methods as well as non-metric multidimensional scaling to differentiate between different groups based on the microbiome composition and immunological data. Moreover, we utilize a machine learning regression model (random forest) to reveal the differences in the development of the microbiome composition in different groups. Our results contribute to the understanding of the microbiome development at the early stage of life and may improve diagnostics of health conditions based on gut microbiome composition.

Deconvolution of methylation data identifies immunological component in sarcomas

Malte Simon, DKFZ, Heidelberg, Germany; Sadaf S. Mughal, DKFZ, Heidelberg, Germany; Benedikt Brors, DKFZ, Heidelberg, Germany; Stefan Fröhling, NCT Heidelberg, Heidelberg, Germany; Charles D. Imbusch, DKFZ, Heidelberg, Germany

Background

Soft-tissue sarcomas (STS) are a heterogeneous group of mesenchymal tumors that remain poorly characterized on a molecular level. With the rise of immunotherapies, it becomes increasingly important to enable patient stratification into groups, which will most likely benefit from these treatment strategies.

Methods and results

We deconvoluted DNA methylomes from tumor samples of the TCGA sarcoma cohort using MeDeCom, a novel non-negative matrix factorization method intended for the discovery of biologically meaningful methylation patterns in the data. As a result, we identified a component associated with tumor-infiltrating leukocytes, which suggests varying degrees of immune cell infiltration between and within the examined STS subtypes and is associated with clinical outcomes. Specifically, our results show high infiltrates in some cases of dedifferentiated liposarcoma and undifferentiated pleomorphic sarcoma and low infiltrates in synovial sarcoma. We further analyzed data on somatic mutations, fusion events and copy number changes to discover genetic alterations that may influence tumor infiltration by leukocytes.

Discussion

To date, there exist only few predictors for immunotherapy response in sarcoma treatment. Our estimated degree of immune cell infiltration for the different STS subtypes is in accordance with previous results from clinical trials using immune checkpoint inhibitors. The results from this study are a solid foundation for further research on the immune compartment in sarcomas, which will support treatment decisions in the future.

Germline Alteration Detection Using a Parents-child Trio Bioinformatics Analysis Pipeline

Layal Yasin, Heinrich Heine University - Medical faculty, Düsseldorf, Germany, Carolin Walter, Institute of Medical Informatics- University of Münster, Münster, Germany, Stefan Janssen, Heinrich Heine University - Medical faculty, Düsseldorf, Germany, Triantafyllia Brozou, Heinrich Heine University - Medical faculty, Düsseldorf, Germany, Ute Fischer, Heinrich Heine University - Medical faculty, Düsseldorf, Germany, Martin Dugas, Institute of Medical Informatics- University of Münster, Münster, Arndt Borkhardt, Heinrich Heine University - Medical faculty, Düsseldorf, Germany

Background

Pediatric tumors are believed to be linked to germline alterations, which either are inherited from parents or exists as de-novo in the child. Since 2015, we have conducted a trio study, where we sequence parents and children with cancer using Illumina Hiseq 2500 to detect possible germline variations. Recently, we started whole genome optical mapping using Bionano genomics for specific trio samples in which no clear causative germline alterations were detected.

Methods and results

Our bioinformatics pipeline processes whole exome sequencing (WES) and Bionano optical mapping data from trio samples. For the WES data, the processing steps start with alignment and quality control followed by variants calling and annotation, where various prediction tools and annotation databases are used. The pipeline currently detects single nucleotide variants, in addition to indels, de-novo mutations, parental and child mosaicism. Afterwards, important cancer-related pathways are analysed in an attempt to detect cancer predisposing germline mutations, de-novo mutations are phased to their parental origin, and digenic mutations are studied for possible pathogenicity. Our pipeline also runs the Bionano's variant calling pipeline to detect inherited and de novo structural variants. As a final step. The detected variants from both WES and optical mapping are uploaded into our in-house

server were they are extensively annotated, and provided through a user friendly interface to our clinicians to have a closer look at, and filter for important genomic alteration, and finally report back to the families. The pipeline is written using snakemake to ensure reproducibility and scalability. It is run on our HPC cluster.

Outlook

Several features need to be integrated in our pipline, such as using structural bioinformatics to predict the pathogenicity of our variants, specially the variants of unknown significance, by predicting the effect of the variant on the stability of the 3D structure of the produced protein. Pipeline source code is available at: https://github.com/sjanssen2/spike.

ORGANISER

DECHEMA e.V.
Theodor-Heuss-Allee 25
60486 Frankfurt am Main
Germany

Matthias Neumann
Phone: +49 (0)69 7564-254
Fax: +49 (0)69 7564-176
E-mail: matthias.neumann@dechema.de