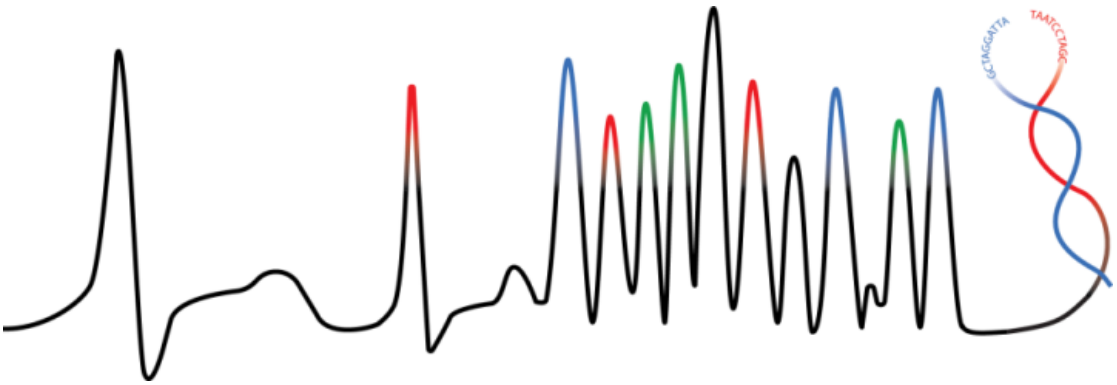
Statistical Challenges in Medical Data Science

Ascona, 16-21 June 2019



Organised by Niko Beerenwinkel, Peter Bühlman, Wolfgang Huber

Contents

Preface

About the workshop

Rapid biotechnological advances have turned the biomedical sciences into a data science. Today, large-scale high-dimensional data is generated routinely by new imaging modalities, DNA sequencing technologies, and many other molecular profiling techniques. These profiles promise to reveal the molecular basis of diseases and to guide the design of novel therapeutic interventions. In addition to molecular and clinical data, mobile health data obtained from internet-based pervasive monitoring can also provide useful information. However, integrating and analyzing complex clinical, molecular, and mobile health data is extremely challenging, and new statistical models and computational inference methods are needed. Our workshop will (i) explore recent advances and open problems in statistical modelling, inference, and integration of molecular profiling, electronic health record, and mobile health data; (ii) identify opportunities and challenges for translation of data science approaches to health and disease, such as the construction of data-driven medical decision support systems; and (iii) facilitate meaningful interactions between engineering, biomedical, and quantitative researchers.

Sponsors

Congressi
Stefano Franscini
Swiss Federal Institute
of Technology Zurich

Swiss Institute of
Bioinformatics

Strategic Focus Area
Personalized Health and Related Technologies

Venue

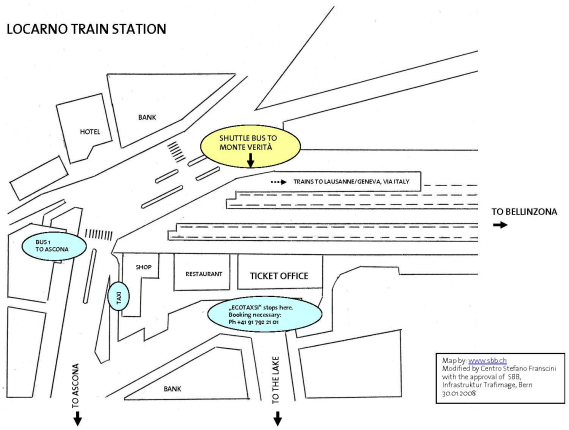
Monte Verità
Via Collina 84
CH-6612 Ascona
tel. +41 91 785 40 40

About the Congressi Stefano Franscini (CSF)

The Congressi Stefano Franscini (CSF) is the international conference centre of the Swiss Federal Institute of Technology (ETH) in Zurich, situated in the south of Switzerland (Canton Ticino) at Monte Verità. It has been named after the Federal Councillor Stefano Franscini, a native of Ticino who, in 1854, played an important part in establishing the first Federal Institute of Technology in Switzerland, ETH Zurich. Every year, the centre hosts 20 - 25 conferences organised by professors working at Swiss universities and concerning all disciplines (sciences and humanities) taught at academic level. The centre is also open to the local population with a regular program of public events (lectures, concerts, films, etc.) organised in the context of its international conferences and/or Monte Verità's cultural programme.

Shuttle service from Locarno Station

A free 13-seater shuttle bus to Monte Verità leaves from Locarno railway station Sunday June 16 at the following times: **14.05**; **14.45**; **15.35**; **16.15**; **17.05**; **17.45**; **18.35**. The meeting point is on the right side of the train platforms in Locarno (see image).



Keynote lectures

Towards A Self-Learning EHR System

Tianxi Cai

Harvard University

The wide adoption of electronic health records (EHR) systems has led to the availability of large clinical datasets available for discovery research. EHR data, linked with bio-repository, is a valuable new source for deriving real-word, data-driven prediction models of disease risk and progression. Yet, they also bring analytical difficulties. Precise information on clinical outcomes is not readily available and requires labor intensive manual chart review. Synthesizing information across healthcare systems is also challenging due to heterogeneity and privacy. In this talk, I'll discuss analytical approaches for mining EHR data with a focus on scalability, reproducibility and automated knowledge extraction. These methods will be illustrated using EHR data from Partner's Healthcare and Veteran Health Administration.

Embracing the complexity of complex traits

Marylyn Ritchie

University of Pennsylvania

Genome science has advanced at a tremendous pace during recent years with dramatic innovations in molecular data generation technology, data collection, and a paradigm shift from single lab science to large, collaborative network science. Still, we predominantly explore one gene and one phenotype at a time. Comprehensive collections of phenotypic data can be used in more integrated ways to better subset or stratify patients based on the totality of his or her health information. With the complexity of the networks of biological systems, the likelihood that every patient with a given disease has exactly the same underlying genetic architecture is unlikely. Through applying machine learning and expert knowledge to the rich phenotypic data of the electronic health records along with the wealth of omics data available, these data can be mined to identify new and interesting patterns of disease expression and relationships. We have been exploring machine learning technologies for embracing the complexity as we explore both the phenomic and genomic landscape to improve our understanding of complex traits. These techniques show great promise for the future of precision medicine.

Delineating the rates and routes of metastasis

Christina Curtis

Stanford University

Metastasis is the most lethal and insidious aspect of cancer. Despite significant therapeutic advances, metastatic disease is generally incurable. To date, the molecular and microenvironmental determinants of metastasis are largely unknown, as is the timing of systemic spread, hindering effective treatment and prevention efforts. In this talk, I will describe two distinct quantitative frameworks to delineate the dynamics of distant metastasis and their application to different solid tumors and types of cohorts. First, I will outline a suite of computational tools we have developed to infer the evolutionary dynamics of tumor progression from patient genomic data by coupling population genetic theory, spatial computational modeling and approximate Bayesian computation. Building on these efforts, I will describe a new method, termed SCIMET (Spatial Computational Inference of MEtastatic Timing) to infer the timing of metastatic spread based on patterns of genomic divergence between paired primary tumors and distant metastases. I will show how application of this approach to colorectal cancer enables quantification of the rates and routes of metastasis in a patient-specific fashion and yields fundamental insights into the drivers of this lethal process with attendant clinical implications. Lastly, I will describe a statistical approach to model the dynamics of breast cancer relapse and its application to a cohort of nearly 2000 breast cancers with detailed genomic information and long-term clinical follow-up (Rueda et al. Nature 2019). Throughout, I will discuss context dependencies that underlie disease progression and how this may inform strategies for patient stratification and therapeutic targeting.

Deciphering geographical complexity of the tumour ecosystem

Yinyin Yuan

IRC London

Tumours are complex, evolving ecosystems with dynamic crosstalk among cancer and normal cells. Measuring spatial heterogeneity in the tumour microenvironment is critical for understanding the spatial context in which cancer evolves. I will discuss the quantitative spatial measures of immune response we developed by combining automated histology image analysis and spatial statistics. Such measures led to new appreciation of the clinical relevance of immune response in breast cancers and high-grade serous ovarian cancer. I will also share our recent progress on studying the geospatial complexity of tumour microenvironment in the TRACERx study of lung cancer evolution.

Challenges in Bayesian analysis with complex models arising in modern health applications

Chris Holmes

Oxford University

Modern studies in the biomedical sciences increasingly involve integrating data at scale across multiple measurement modalities. On the one hand, the joint probabilistic nature of Bayesian statistics is attractive as it provides a formal framework for the inclusion of prior knowledge alongside hierarchical, adaptive, model structures with full uncertainty quantification. On the other hand, computational issues in the model fitting and model evaluation, as well as robustness to modelling assumptions, are a major challenge. We will discuss some recent work in robust Bayesian modelling that seeks to address some of these issues in the context of large-scale genetic association studies.

Brief excurse into the problem of drug response prediction

Anna Goldenberg

University of Toronto

Deciding on the best treatment for the patient, often via predicting drug response, is one of the holy grails of precision medicine. There is much data collected to try to address this problem and yet many computational challenges remain. In this talk I will touch on several issues in predicting drug response and our solutions to these problems including using variational autoencoders for predicting drug response in cell lines, quantifying drug response in patient derived xenografts (PDX) and explainable models in patients with some observations regarding difficulties with domain adaptation when translating cell-line based models to predicting response in patients.

Big data and health

Michael Snyder

Dept. of Genetics,Stanford University School of Medicine

Recent technological advances as well as longitudinal monitoring not only have the potential to improve the treatment of disease (Precision Medicine) but also empower people to stay healthy (Precision Health). We profiled 107 participants using multiomics (genomics, immunomics, transcriptomics, proteomics metabolomics and microbiomics etc) and wearables technologies for up to eight years and made 67 major health discoveries. We have also invented ways to measure the airborne exposome at high resolution and found the people are routinely exposed to 1000s of biological and chemicals. Altogether, we conclude that deep longitudinal profiling using advanced technologies can lead to actionable health discoveries and provide important information relevant for precision health.

Machine Learning for Protein Engineering

Jennifer Listgarten

UC Berkeley

With the advent of more and more high-throughput technologies to measure protein properties of interest such as binding, expression, fluorescence, the time for machine learning to act synergistically with protein design is here. I will touch on two stories in this space. The first will be about how machine learning can be leveraged to improve CRISPR gene editing. The second will touch on how one can accelerate design/optimization of proteins with machine learning approaches – a sort of in silico approach to the method of Directed Evolution, which won the 2018 Nobel prize in Chemistry.

A simple new approach to variable selection in regression, with application to genetic fine-mapping

Matthew Stephens

University of Chicago

We introduce a simple new approach to variable selection in linear regression, and to quantifying uncertainty in selected variables. The approach is based on a new model – the "Sum of Single Effects" (SuSiE) model – which comes from writing the sparse vector of regression coefficients as a sum of "single-effect" vectors, each with one non-zero element. We also introduce a corresponding new fitting procedure – Iterative Bayesian Stepwise Selection (IBSS) – which is a Bayesian analogue of stepwise selection methods. IBSS shares the computational simplicity and speed of traditional stepwise methods, but instead of selecting a single variable at each step, IBSS computes a *distribution* on variables that captures uncertainty in which variable to select. The method leads to a convenient, novel, way to summarize uncertainty in variable selection, and provides a Credible Set for each selected variable. Our methods are particularly well suited to settings where variables are highly correlated and true effects are sparse, both of which are characteristics of genetic fine-mapping applications. We demonstrate through numerical experiments that our methods outperform existing methods for this task.

Machine Learning for Biomarker Discovery: Combinatorial Association Mapping

Karsten Borgwardt

ETH Zürich

The enormous data wealth in Medicine offers huge opportunities for Machine Learning to detect biomarkers in patient data that can be exploited for improved diagnosis, prognosis and therapy decisions. In this talk, I will describe our work on developing new combinatorial approaches to Biomarker Discovery that try to identify disease-related combinations of patient properties.

Contributed talks

Representation Learning of Patient Health States

Gunnar Rätsch

ETH Zurich

I will discuss approaches for learning latent representations of health states from temporal data. We use unsupervised and supervised methods to find low-dimensional representations characterizing the health state of the patient that is predictive of future outcomes. I will discuss applications of the method in the analysis of data from intensive care units.

Stable prediction with radiomics data

Tim van de Brug

Amsterdam University Medical Centers / Location VUmc

Tim van de Brug and Carel F.W. Peeters
Radiomics refers to the high-throughput mining of quantitative features from medical images. These features characterize a volume of interest (such as a tumor) and enrich the standard radiological lexicon. The promise of radiomic data is that it may be a basis, through the information contained in standard of care images, for non-invasive medical decision support at low additional cost. The problem with these data, however, is that they are usually high-dimensional as well as highly collinear. For many prediction methods this is a precarious combination. We present a novel method, based on factor analysis, that copes with this challenge. It projects the original high-dimensional feature space onto a lower-dimensional orthogonal latent-feature space. This projection produces a compact set of stable features that can be directly used in any classifier

or predictor. We apply the method to a multicenter database of PET-CT images of diffuse large B-cell lymphoma patients. The proposed method outperforms other machine learning and feature selection techniques in terms of the stability of time-to-event predictions. Joint work with Mark van de Wiel.

Evaluation strategies to examine impact of a risk-based allocation algorithm for transplantation.

Jean Yee Hwa Yang

University of Sydney

Kidney transplantation is the optimal treatment for many patients with end-stage kidney disease as it improves survival and quality of life for the hundreds of thousands of Australians on dialysis. Rapidly emerging tools in omics technologies have generated complex data and have more recently been incorporated to further fundamentally understanding risk factors in organ transplantation. Using statistical learning, we first examine the impact of higher resolution immunological signature on transplant allocation. In addition, we examine the impact of a new risk-based, deceased donor kidney allocation formula in comparison to the current Australian algorithm.

We use over 20 years transplant data with more than 7000 patients to build a new evaluation framework and examine the impact of risk-based matching algorithms. We measure its impact based on waiting time, quality of life and graft years. Our results show that risk-based matching engendered a moderate, a moderate, overall increase in graft and patient survivals, accrued through benefits for recipients 45 years or younger but part of the advantage is offset by recipients older than 60 years.

Using network analysis to illuminate neurological dysfunction

Kasper Hansen

Johns Hopkins University

Coding variants in epigenetic regulators are emerging as causes of neurological dysfunction and cancer. Here, we study the co-expression patterns of these genes across

tissues. We remove unwanted variation from the expression data and we assess the impact of doing so on co-expression analysis using both positive and negative controls. We find a single group of 74 epigenetic regulators which are highly co-expressed across all adult tissues. We show that co-expression is associated with intolerance to loss-of-function coding variation; more than 90% of these highly co-expressed genes are intolerant. Furthermore, we find that these highly co-expressed genes are associated with neurological dysfunction compared to non-co-expressed epigenetic regulators (odds-ratio: 5.3) but not with cancer. Finally, brain-specific regulatory regions of these genes are enriched for explained heritability for common neurological traits such as bipolar disorder and neuroticism. This finding establishes that co-expression plays a functional role in normal neurological homeostasis.

Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq

Simona Cristea

Dana-Farber Cancer Institute

Triple-negative breast cancer (TNBC) is an aggressive subtype characterized by extensive intratumoral heterogeneity. To investigate the underlying biology, we conducted single-cell RNA-sequencing (scRNA-seq) of >1500 cells from six primary TNBC. Here, we show that intercellular heterogeneity of gene expression programs within each tumor is variable and largely correlates with clonality of inferred genomic copy number changes, suggesting that genotype drives the gene expression phenotype of individual subpopulations. Clustering of gene expression profiles identified distinct subgroups of malignant cells shared by multiple tumors, including a single subpopulation associated with multiple signatures of treatment resistance and metastasis, and characterized functionally by activation of glycosphingolipid metabolism and associated innate immunity pathways. A novel signature defining this subpopulation predicts long-term outcomes for TNBC patients in a large cohort. Collectively, this analysis reveals the functional heterogeneity and its association with genomic evolution in TNBC, and uncovers unanticipated biological principles dictating poor outcomes in this disease.

Estimation of Mutation, Drift and Selection in Single-Driver Hematologic Malignancy

Marek Kimmel

Rice University

Severe congenital neutropenia (SCN) manifests itself through an inability to produce enough granulocytes to prevent infections. SCN commonly results from a germline ELANE mutation. Large doses of the blood growth factor granulocyte colony-stimulating factor (G-CSF) rescues granulocyte production. However, SCN frequently transforms to a myeloid malignancy, commonly associated with a somatic mutation in CSF3R, the gene encoding the G-CSF Receptor. We built a mathematical model of evolution for CSF3R mutation starting with bone marrow expansion at the fetal development stage and continuing with postnatal competition between normal and malignant bone marrow cells. We employ tools of probability theory such as multitype branching process and Moran models modified to account for expansion of hematopoiesis during human development. We estimate coefficients, to obtain agreement with the age range at which malignancy arises in patients. Our model predicts the existence of a pool of cells with mutated CSF3R before G-CSF treatment begins. Estimated CSF3R mutation rates appear to be within or not far from the range typical for human somatic cells.

Genome-wide cell-free DNA fragmentation as a biomarker for early detection of cancer

Stephen Cristiano

Johns Hopkins University

The high morbidity and mortality of cancer results from late diagnosis where therapeutic intervention is less effective, yet clinically proven biomarkers to broadly diagnose patients are not widely available. Analyses of cell-free DNA (cfDNA) in blood provide a noninvasive diagnostic avenue for patients with cancer. However, cfDNA analyses have largely focused on targeted sequencing of specific genes. Genome-wide analyses of cfDNA features may increase the resolution of changes in circulating tumor DNA compared to healthy cfDNA and promote more sensitive cancer detection. We developed an approach to analyze fragmentation profiles and cfDNA features across the genome

and applied this method to analyze cfDNA from 236 patients with breast, colorectal, lung, ovarian, pancreatic, gastric, or bile duct cancers and 245 healthy individuals. Machine learning incorporating these features resulted in sensitivities of detection from 57% to >99% among seven cancer types at 98% specificity, as well as narrowed the tissue of origin to a limited number of sites. The results of these analyses highlight important properties of cfDNA and provide a facile approach for early detection of human cancer.

Pairtree: constructing mutation trees from multi-sample sequencing data using pairwise interactions between mutations

Quaid Morris

University of Toronto

Tumours are not homogeneous masses, but instead contain multiple subpopulations of cancerous cells. Each subpopulation has unique genetic mutations that can affect treatment response. Using short-read sequencing, we can detect these mutations, then use them to build an evolutionary tree for each cancer describing its evolutionary trajectory. These trajectories can provide insights into carcinogenesis and inform optimal treatment.

Building cancer evolutionary trees is challenging because the noisy, low-resolution nature of sequencing data leads to ambiguity in evolutionary relationships between mutations. To reduce ambiguity, we can use multiple tissue samples from a cancer, each of which is composed of a different mixture of subpopulations. These tissue samples can be taken from different points in space (e.g., primary tumour and metastasis) or in time (e.g., diagnosis and relapse).

Existing algorithms do not efficiently exploit the information provided by multiple samples from a cancer. Here we present Pairtree, a novel algorithm for evolutionary tree reconstruction that overcomes these limitations. Pairtree infers probabilistic pairwise constraints between mutations to constrain a Markov Chain Monte Carlo tree search. Then it uses a fast projection algorithm to score trees according to an approximation of their marginal likelihood. Combining these two advances allows Pairtree to quickly and accurately reconstruct evolutionary trees with dozens of subpopulations across dozens of samples, a task impossible with existing methods.

Large-scale variational inference for hierarchical multiple-response regression modelling of high-dimensional genetic data

Hélène Ruffieux

EPFL

Joint work with: Leonardo Bottolo, Sylvia Richardson, Anthony C. Davison and Armand Valsesia.
We present an efficient Bayesian approach for joint analysis of molecular quantitative trait locus (QTL) data on a genome-wide scale. We consider a series of parallel sparse regressions combined in a hierarchical manner to flexibly accommodate high-dimensional responses (molecular levels) and predictors (genetic variants).
Our novel framework allows information-sharing across outcomes and variants, thereby enhancing the detection of weak effects, and directly controls the propensity of variants to be hotspots, i.e., to remotely control the levels of many gene products, via a dedicated top-level representation. It implements annealed variational inference procedures that enhance exploration of multimodal spaces and allow simultaneous analysis of data comprising hundreds of thousands of predictors, and thousands of responses and samples.
Our methods are publicly available as packages implemented in R and C++. We illustrate their advantages in simulations and in a large-scale proteomic QTL study on two clinical cohorts that highlights novel candidate biomarkers for metabolic disorders.

Parsing latent factors in high dimensional classification

Johann Gagnon-Bartsch

University of Michigan

High throughput biological data often contains signals from multiple unobserved latent factors in addition to the signal of primary interest. In a classification analysis, some of these latent factors may be partially correlated with the phenotype of interest and therefore helpful, some may be uncorrelated and thus merely contribute additional noise, while more perniciously, some may be spuriously correlated with the phenotype in the training set but not in the target population, leading to poor generalized predictive performance. Moreover, whether potentially helpful or not, these latent factors may

obscure weaker direct effects that capture the signal of primary interest. It is therefore desirable to separate out these latent variables. This talk has two parts. The first outlines a classification algorithm that first isolates the signal of primary interest from other latent factors, but then exploits both to improve prediction, leading to sometimes substantial gains. The second discusses how to remove uncorrelated, non-stationary, or otherwise harmful latent factors.

Know when you don't know: Quantifying the uncertainty of deep learning based image classifications

Beate Sick

ZHAW and University of Zurich

In many biological and medical applications a highly reliable classification of image data is essential since critical treatment decisions or expensive further experiments are to follow. Besides the accuracy, a measure for the certainty of an individual classification is of great interest.
Convolutional neural networks (CNN) have revolutionized classification of image data. However, these approaches are currently mainly used to provide point estimates without uncertainty measures and without taking into account dependencies between different images. To address this shortcomings, we utilize a recently developed Bayesian approach for deep neural networks called Monte-Carlo (MC) dropout to determine different uncertainty measures for each classification. Here, we show how these uncertainty measures can be used to in real world medical and biological applications, such as high content screening, MRI analysis, and digital pathology to identify uncertain classifications on image and patient level and/or indicate potential novel classes.
Authors: B. Sick1,2*, L. Herzog1,2, E. Murina1, O. Dürr1,3

Inference of single-cell copy number profiles from parallel DNA and RNA sequencing

Francesco Marass

ETH Zurich

Single-cell sequencing has been instrumental for studying heterogeneity in the composition of cellular populations. Due to the limited starting material, however, it has been difficult to concurrently analyse such compositions across different data modalities. While this constraint has been partially lifted by the introduction of multi-omics approaches for single-cells, this has come at the cost of lower data quality. In cancer, single-cell genome and transcriptome sequencing (G&T-seq) provides matched mutational and expression profiles, with both data types offering information on copy-number alterations. As DNA and RNA are processed differently, the resulting data show different errors, and integration of the two modalities can lead to more accurate copy-number profiles. We introduce a Bayesian model to infer copy numbers jointly from DNA and RNA data of single cells. The method is benchmarked on data generated by the model, and compared against current techniques for single-cell CN analysis. In particular, we show that robust profiles can be inferred despite the presence of high levels of noise with a dataset of 337 cells that were subjected to MDA amplification and exome capture.

Latent factor modelling applied to multi-omics data integration and to precision oncology in chronic lymphocytic leukaemia

Junyan Lu

EMBL Heidelberg

Stratifying patients into different risk subcategories and drug sensitivity groups is one of the most important steps in precision medicine. Genomic, transcriptomic and epigenetic data have been explored for patient stratification in different types of cancer but current approaches often treat them as separate. For example, in chronic lymphocytic leukemia (CLL), one of the most common leukemias in the western world, only a few genetic markers, such as IGHV (immunoglobulin heavy-chain variable region gene), TP53 and NOTCH1 mutations, are considered for treatment design. However, since different omics data types can provide complementary information for understanding

complex biological systems, integrating them should provide a more comprehensive view of cancer heterogeneity and improve patient stratification.
I will report on the analytical and conceptual aspect of a large study in which we applied multi-omics factor analysis (MOFA) to the genomic, transcriptomic, and DNA methylomic data from 274 CLL patients to identify clinically relevant latent factors or subgroups. In particular, MOFA identified a "mystery" latent factor that appears to define a novel aggressive CLL subtype with activated energy metabolism and increased mTOR pathway activity. We validated the presence of this aggressive subtype in three external cohorts. Drug sensitivity profiling in our cohort also revealed that this novel CLL subtype exhibited specific vulnerability towards MEK/ERK and mTOR pathway inhibitors. Overall, our study suggests that integrating multi-omics datasets can lead to better and more biologically interpretable patient stratification, and thus contribute to better individualized patient care.

Sleepwalk: a tool to interactive explore dimension-reduced embeddings

Svetlana Ovchinnikova

University of Heidelberg

Dimension reduction techniques such as MDS, t-SNE or UMAP are routinely used to explore high-dimensional data like, for example, single-cell transcriptomes. These tools help to visualzie structure in the data, but the process of dimension reduction unavoidably induces potentially misleading artifacts, and analysts need a means to distinguish these from genuine structure.
We present "sleepwalk", a simple but powerful visual and interactive tool, which displays an embedding, and dynamically colours the data points according to their high-dimensional-space distance to the point under the mouse cursor. By moving the mouse and observing how the colours change, the user obtains an immediate and intuitive grasp of hitherto hidden aspects of the embedding. The tool can also be used to compare between data sets or between analysis or dimension-reduction approaches.
The Sleepwalk R package, and a live demo, can be found at anders-biostat.github.io/sleepwalk

---

Posters

01: Understanding AML using Data Science

Lisa Amrhein

Helmholtz Zentrum Munich

Acute Myeloid Leukemia (AML) is the most common acute leukemia affecting adults. Even after complete remission, leukemic cells likely remain in numbers below detection limit. Without further postremission or consolidation therapy, most AML patients will eventually relapse and die.

AML patients frequently carry mixtures of different cancer cell types, so-called subclones, which evolve over time, so that the mixture at relapse is different from the one at diagnosis. Understanding clonal evolution and identifying rare subclones is still an open challenge. We aim to parameterize transcriptional heterogeneity from RNA-Seq counts taken from small groups of cells (one to 50 cells per sample).

First, we investigate how current analysis methods work with these varying amounts of material per sample and which cell-pool size is the best option. Nest, we use our Stochastic Profiling Method (REF) to infer single-cell regulatory states by mathematically deconvolving n-cell measurements. This averaging-and-deconvolution approach allows us to quantify single-cell regulatory heterogeneities while avoiding the technical measurement noise of single-cell techniques.

02: Using interactive visualization to explore big-data analyses

Simon Anders

University of Heidelberg

In molecular medicine research, we often perform large assays or studies, e.g., probing thousands to millions of perturbations or potential interactions, then analyse this huge amount of raw data by long complex pipelines to reduce the data to a few high-level summarizing plot or short list of "interesting" results. The risk is great that some of the results are caused by artifacts or unexpected behaviour of the pipeline.

We argue that it is hence most valuable to be able to "walk" through any analysis pipeline "backwards": starting from a single result value, we want to see plots of the intermediate values that the result summerise, and for each of these, of the raw data that these are based on. Interactivity is crucial here, to give the analyst the tools to navigate through the analysis.

We present LinkedCharts, a R and JavaScript framework, that makes it very easy for bioinformaticians to set up tools for this purpose. We show how this provides not only a powerful and convenient analysis methodology, but also a new, complementary, approach to ensuring soundness of any big data analysis.

Web link: `https://anders-biostat.github.io/linked-charts/`

03: Data-driven clustering of airway disease phenotypes in west Sweden adult population

Rani Basna

University of Gothenburg

Machine learning solution for biology is an emerging area of research. The high complexity level of the biological system motivates for more advanced and robust models to cope with such complexity. Specifically, the use of such machine learning to airway diseases that integrate personal, clinical, and genetic variables has not been investigated properly. We propose a hybrid deep learning model that identify a set of phenotypes and perform a pathway analysis to predict gene functions and classify individuals within a cohort. Studying the genetic variation of each phenotype and the impact of gene variation on the correlated hierarchical structure of gene expression. Our pipeline suggests a CNN model to implement feature selection on the genetic data to infer the differentially expressed genes. secondly, we perform a clustering analysis to identify the disease phenotypes. Finally, we plan to construct RNN model to capture the sequential structure of the disease formation and lead to infer the gene regulatory network. On a high level, we propose a convolutional embedded LSTM model with attention to achieve this target.

04: Unsupervised method for automated segmentation of lung nodules on a 3D series of Computed Tomography scans.

Franciszek Binczyk

Politechnika Śląska

Automated lung nodule detection is a crucial step in the diagnosis of lung cancer. In this work, we present an algorithm for automated nodule detection on a 3D series of CT scans. The proposed algorithm is composed of three steps. The first step is a lung detection, based on data-driven a cascade of morphological operations combined with Gaussian mixture modelling. The second step is the modelling of the bronchial tree based on the original extension of the ellipses modelling routine (DOI: 10.1109/TIP.2015.2492828). The idea of the algorithm was extended in a way that it is capable of reconstructing continuous ellipsoidal shapes (bronchial tree) in a 3D series of images. The last step of the proposed algorithm is anomaly detection performed with the modification of MiMSeg algorithm, by Binczyk (doi:10.1016/j.ins.2016.07.052) combined with the previously detected bronchial tree. The algorithm was validated on a publicly available set of images (doi.org/10.7937/K9/TCIA.2016.6fC8z46u). The obtained quality indexes: Sensitivity: 84.97% (95% CI: 83.13–86.81), Specificity: 99.79% (99.78–99.80) and Dice Index: 86.63% (85.40–87.86), proves the effectiveness of proposed algorithm.

05: Function and structure of HDL in diabetes and coronary heart disease

Mathias Cardner

ETH Zurich

High-density lipoproteins (HDL) are complex particles carrying more than 80 proteins and hundreds of lipid species. There is a known inverse correlation between plasma levels of HDL cholesterol and risk of type 2 diabetes and coronary heart disease (CHD), but a better understanding of the functional and structural properties of HDL particles is required for prevention or treatment based on HDL. To this end, we isolated HDL in plasma samples taken from 51 healthy volunteers and 98 patients with diabetes or

CHD. We integrated data from functional bioassays with proteomics and lipidomics measured through mass spectrometry, as well as lipoprotein particle profiling quantified through NMR spectroscopy. For disease association, we performed elastic net logistic regression with stability selection in order to find candidate biomarkers. We applied a high-dimensional Gaussian copula graphical model to infer partial correlations between functional and structural features of HDL. Clinically interesting associations were assessed in validation experiments, which established a novel causal link between a species of sphingomyelin and the ability of HDL to inhibit apoptosis in endothelial cells.

06: Single-cell RNA-Seq data batch correction tool benchmarking

Ruben Chazarra - Gil

Sanger Institute

Single-cell trancriptomics has become an immensely powerful technology in biological research with the capacity of interrogating the expression profile of thousands to millions of individual cells. World-wide distributed efforts aiming to reveal the transcriptional identity of all cells through entire organisms are being established. This is leading to the discovery of new cell types and other biological insights that are potentially masked when only studying individual datasets. However, data integration has to be assessed carefully as variability arising from different technologies, different laboratories, or different donors may compromise the interpretation of the final results. In this context, there are many tools that correct for this experimental intrinsic variability, commonly known as batch effect. In our study, we benchmark the most popular tools for batch correction. Our test cases reflect the sources of variability that may be present when integrating and joint analysing scRNA-Seq data of various origins.

07: Utilization of Cancer Clinical Trial Data

Dung-Tsa Chen

Moffitt Cancer Center

Cancer clinical trial data has been underutilized due to challenges of dealing with many types of sophisticated data. Each data type is often stored in a multi-dimensional and longitudinal layout with potential hierarchical structure. Such complexities often scare away researchers and lead to the use of naïve descriptive statistics for analysis. As a result, it loses valuable information and becomes a major hurdle to advance cancer precision medicine. Here we share our thought to tackle the problem by analyzing the data from various domains to better understand the unique features for each data type. In the meantime, through this process, numerous new variables will be generated for new hypotheses (i.e., from analytic role to experimental design role). Modern data integration strategies will be utilized to link the new variables within and between the data types. Contemporary statistical methodology will be explored to unravel the complicated clinical data into useful information. The success will pave a smooth road to optimize treatment and prevention strategies to achieve precision medicine in cancer.

08: DEEP-WEB BASED SELF ORGANIZING MESH NETWORKS FOR SAFE AND SECURE DATA SHARING FOR CLINICAL RESEARCH

Luca Clivio

E.O.C. Bellinzona

Controlled Clinical Trials are at the heart of Clinical Research, and are about comparing different treatments in several patients. Since with the advent of High Throughput -Omics data generators and Imaging the number of variables per patient to be compared grew incredibly, the number of involved patients must grow as well till not having mono-centric trials possible anymore for having enough statistical power. This forces huge efforts to organize safe and secure data sharing mechanisms that often are in contrast with the local policies of the hospitals. This often become the main obstacle to the organization of the Clinical Trials themselves. Self Organising virtual Networks as philosophy represent the hope for the Investigators wanting to be involved in a multi-centric Trials. Here we present the use of TOR as transport layer for obtaining a secured data sharing for a Peer to Peer database compliant with the "FDA cfr 21 part 11" guidelines for handling Electronic Health Records. The resulting software is open-source, multi platform and able to perform a network auto topology discovery for establish an ad-hoc peer to peer self-organizing Mesh Network for Virtual Organizations.

09: Cancer Pathway Disruption Score Reveals Underlying Tumor State from Expression Data

Natalie Davidson

ETH Zürich

Molecular pathways dictate a cell's every action; in cancer, they are disrupted to promote proliferation. Therefore, it's key to understand the extent to which pathways are altered when considering therapies. To address this, we designed a cancer pathway disruption score (CPDS) that quantifies the divergence of a patient's pathway activity from normal tissue. Our method assumes that tumors have different pathway expression patterns than normal tissue. Using this belief, we used discriminant component analysis on TCGA RNA-Seq data to identify a pathway-specific gene weighting that separates tumors from normal samples. To validate our CPDS we calculated a somatic mutation pathway burden and associated it with our expression-based CPDS. We found a significant association in 20/44 pathways, including those typically affected by somatic mutations like mTORC1 and KRAS. In comparison, GSVA found an association in 7 pathways, (not mTORC1 or KRAS). We also looked at the hypoxia pathway and validated our score on external datasets and found significant associations in TCGA with genetic markers of hypoxia. In conclusion, CPDS reflects both genetic and micro-environment changes in the tumor.

10: Extracting Meaningful Patterns from Big Binary Data using E-BiBit Algorithm

Ewoud De Troyer

Janssen Pharmaceutica (J&J)

Biclustering is a data analysis method that can be used to simultaneously cluster the rows and columns in a (big) data matrix in order to identify local patterns in a big data matrix. For binary data matrices, these local patterns consists of rectangles of 1's. Most binary biclustering methods (Bimax (Prelić et al., 2006), BiBit (Rodriguez-Baena et al., 2011) focus on the discovery of perfect biclusters (i.e., zeros are not included in the bicluster). We present an extension for the BiBit algorithm (E-BiBit) that allows for noisy biclusters. While this method works very fast, its downside is that it often produces a large number of biclusters (typically >10000) which makes it very

difficult to recover any meaningful patterns and to interpret the results. We propose a data analysis workflow to extract meaningful noisy biclusters from binary data using an extended version of BiBit and combine it with traditional clustering methods. We illustrate this workflow using heart disease patient data and tourism data (e.g. homogeneous subsets of patients who share the same disease symptom profiles). The algorithm is available in the BiBitR R package, as well as in the BiclustGUI R package.

11: Accounting for physical activity to personalize insulin treatment in children with type 1 diabetes mellitus

Julia Deichmann

ETH Zurich

Treatment of type 1 diabetic patients relies entirely on exogenous insulin, and timing and dosing require constant adjustment. Physical activity (PA) increases the risk of hypoglycemia for several hours, thus posing unique challenges for accurate glycemic control. We aim at developing a model of glucose-insulin regulation and a corresponding model-predictive control algorithm to propose personalized insulin treatment adjustments for minimizing this risk. The model needs to address patient variability by using characteristics such as age, weight and disease severity. Application to free-living conditions requires including unscheduled meal ingestion. Modeling of several glucose metabolic processes on different time-scales is crucial for predicting blood glucose during and after exercise. We train the model on data from 50 children, with continuous glucose measurements and accelerometer counts as well as sparse and imprecise diary information about insulin injections, meals and exercise recorded over the course of six days. We use current and previous blood glucose and accelerometer data or duration and type of PA for proposing treatment adjustment to patient and physician.

12: Predicting disease progression in neurodegenerative diseases with high phenotypic variability

Frank Dondelinger

Lancaster University

Identifying factors that influence the clinical progression of neurodegenerative diseases is of critical importance to both experimentalists trying to understand the disease mechanisms, and clinical researchers trying to develop improved therapies. While much effort has gone into the detection of risk factors for a given disease, most of these approaches ignore the inherent variability in the clinical phenotypes. We have developed a high-dimensional mixture model approach for jointly solving the problem of data-driven estimation of clinical phenotypes and prediction of disease progression. Longitudinal dynamics are captured via a mixed model approach, and we take into account both the distribution of the response and the distribution of the covariates for estimating the disease phenotypes. We demonstrate the performance of our method by applying it to data from the PROACT database on amytrophic lateral sclerosis as well as data from the Alzheimer's Disease Neuroimaging Initiative (ADNI, Mueller et al., 2005). We show that in both cases joint inference of the subtypes and predictors improves the prediction performance, and hence the clinical usefulness of our results.

13: Functional Data Visualization in Medical Data Science

Marc G. Genton

KAUST

Visualization of modern complex data arising in medical Data Science has become more important than ever. Many such data take the form of functional data, i.e., data that are recorded continuously through time or images/surfaces. Moreover, the data are often multivariate as several variables are of interest. We propose a two-stage functional boxplot for the visualization and exploratory data analysis of multivariate curves. Specifically, the original functional boxplot is combined with an outlier-detection procedure based on the functional directional outlyingness, which accounts for both the magnitude and shape outlyingness of functional data. This combination is robust to various types of outliers and, hence, captures the data structures more accurately than the functional boxplot alone. It also allows for both marginal and joint analysis of the multivariate curves. We apply the proposed tool to various functional datasets. In particular, we analyze log periodograms of EEG time series data in the spectral domain and we explore the variation of the spectral power for the alpha (8–12 Hz) and beta (16–32 Hz) frequency bands across the brain cortical surface.

14: A systems medicine approach to understand crosstalk between chronic lymphocytic leukaemia and the tumour microenvironment reveals key mechanisms of drug sensitivity and resistance

Holly Giles

EMBL Heidelberg

CLL is the most common adult leukaemia in the western world. While treatment regimes are often successful in managing the disease, significant rates of relapse remain an issue. We set out from the working hypothesis that cross-talk between the malignant B cells and accessory cells in the bone marrow and lymph nodes underlies drug resistance and disease persistence. We aimed to gain a comprehensive understanding of the signalling pathways involved in this cross-talk, and how they integrate with the underlying tumour genetics, with the ultimate goal to improve patient prognoses. We modelled interactions between CLL cells and the microenvironment, with a drug-cytokine combinatorial screen on 196 primary patient samples. We integrated exome, RNAseq and DNA methylation data for the patients, to reveal how the microenvironment and the underlying genetics integrate to affect B cell survival and drug response. Analysis and modelling of the dataset revealed key microenvironmental signalling pathways that contribute to CLL persistence and interact with drug action. Such targetable pathways point to strategies for improving patient drug responses in vivo. We highlight the IL4 pathway as a resistance mechanism during BCR inhibition and apply a systems medicine approach to shed light on the underlying mechanism of this effect. We propose IBET-762, a transcriptional inhibitor, as a drug capable of inhibiting IL4 signalling and re-sensitising CLL cells to BCR inhibition in the context of IL4. We anticipate that understanding the effects of microenvironmental signalling on drug action in such a way may point to strategies for combinatorial treatments to improve responses to current available drugs.

15: Bayesian statistical approaches to identification of shared genetic signals: colocalisation and fine-mapping

Hui Guo

University of Manchester

A number of genes have been found associated with certain clinical outcomes of interest from multiple studies. Identification of shared causal genes of these outcomes is crucial to understand the aetiology of certain diseases and the underlying causal pathways.

To date, several statistical methods have been developed for such purpose in relevant research field. However, there is lack of comprehensive review of these approaches in the literature to guide researchers in health data science. In particular, Bayesian approaches such as colocalisation and fine-mapping have attracted much attention of the research community. Despite different underlying assumptions required between the two methods, we see similarities of them. We will be investigating colocalisation and fine-mapping by using summary statistics from large-scale association studies, e.g. UK Biobank and via a number of simulations.

16: Regularized pathway regression for improved disease stratifications

Kim Jablonski

ETH Zurich

The goal of my research is to tackle the problem of cancer heterogeneity by creating a pathway enrichment tool that computes significantly enriched pathways for a given sample and then uses this information for additional clustering steps.

In order to do this, a novel tool is developed that computes the statistical enrichment of a given pathway from affected gene lists, while incorporating the similarity of all pathways in the used database (e.g. obtained from PPI colocalization) as a regularization term.

Subsequently, this measure is applied to pan-cancer stratification and patient-survival analyses with the goal of improving accuracy and sensitivity when compared with existing methods. At this point, it is crucial to combine the obtained disease clusters with existing healthcare data in order to facilitate their interpretation. This is advantageous in the field of personalized medicine, where drugs have to be delivered on a patient-specific basis.

17: Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data

Katharina Jahn

ETH Zurich

Understanding the clonal architecture and evolutionary history of a tumour poses one of the key challenges to overcome treatment failure due to resistant cell populations. Previously, studies on subclonal tumour evolution have been primarily based on bulk sequencing and in some recent cases on single-cell sequencing data. Either data type alone has shortcomings with regard to this task, but methods integrating both data types have been lacking. Here, we present B-SCITE, the first computational approach that infers tumour phylogenies from combined single-cell and bulk sequencing data. Using a comprehensive set of simulated data, we show that B-SCITE systematically outperforms existing methods with respect to tree reconstruction accuracy and subclone identification. B-SCITE provides high-fidelity reconstructions even with a modest number of single cells and in cases where bulk allele frequencies are affected by copy number changes. On real tumour data, B-SCITE generated mutation histories show high concordance with expert generated trees.

18: Single-cell analysis of epigenetic and transcriptional states during cardiac aging

Bogna Kliem

DZHK@Max Delbrück Center for Molecular Medicine

Aging is accompanied by a progressive decline of physiological integrity. To gain insights in the molecular trajectories of cardiac aging, we used nucleic acid sequence data derived from simultaneous single-nucleus (sn) detection of the transcriptome and the DNA methylation states of cross-sections of snap-frozen hearts of 12 weeks and 20 months old healthy mice. In general single-cell/nucleus data sets are incomplete, in particular in the case of bisulfite-treated DNA, making analysis of DNA methylation difficult. To deal with low detection sensitivity we defined single-nucleus DNA methylation states through pairwise comparisons of single CpG methylation measurements. Sn-epigenetic states could be described by stochastically sampled fractions of CpGs (even <1%!). Using relationships found this way, we defined near complete

methylomes for epigenetically distinct subpopulations of various cell types in the heart. Furthermore, we found statistically significant epigenetic changes at genes that were relevant to cell-specific function. Our approach allowed for contributing a first integrated epigenomic and transcriptional data atlas of a mammalian heart with single-cell resolution.

19: Differentially regulated translation events in T cell differentiation and activation

Jana Koch

University of Zurich

Background: The presence and relative abundance of specific CD4+ T cell (TC) populations are highly relevant for allergic reactions. The differentiation of naïve CD4+ TC into different TC subsets is a precisely regulated process. Accumulating evidence shows that short ORFs localized on mRNAs and non-coding RNAs can be translated into small peptides, yet their functionalities are so far largely unknown. The aim of this project is to identify small peptides that are involved in TC activation and differentiation to possibly gain a better understanding of the mechanisms of allergic responses and immune tolerance. Methods: Ribosome Profiling and RNA Sequencing of in vitro differentiated human CD4+ TC subsets are employed to identify previously unknown translation events. The differential expression of short ORFs between TC subsets is assessed and identified peptides are further analysed for potential immunomodulatory activities. Conclusion: This approach provides a great opportunity to decipher the regulatory network of TC activation and differentiation beyond transcription factors, in order to identify new mechanisms and potential future targets for the treatment of allergic diseases.

20: Modeling drug efficacy in cancer

Krzysztof Koras

University of Warsaw

The ability to predict response of a specific cancer type to a therapy is one of the main goals in precision medicine. Various machine learning methods concentrate on predicting drug response based on genomic profiles of cell lines and discovering molecular features relevant to compound sensitivity. In this work, we developed predictive models which make use of compound properties in order to narrow down the initial biological feature set. We compared their performance with baseline model which utilizes genome-wide data using GDSC pharmacogenomics database. We show that for the majority of drugs, models exhibit insufficient prediction performance, due to the low variance of drug response across cell lines. We demonstrate that conventional evaluation metrics can be misleading in this context and provide the means to overcome this. Two main groups of compounds were identified: those for which models showed better overall ability to predict drug efficacy, as well as those for which modeling performance was improved when using significantly less initial features in comparison to conventional, genome-wide model.

21: The Use of Genome Scale Data for the Prediction of Pathway Activity and Drug Efficacy

Julian Kreis

University of Heidelberg

Biological pathways are maps of interacting proteins and entities which are responsible for tasks like the production of substances or the regulation of molecules. Their activity is an important biomarker, representing possible causes of malfunctioning cells. In research, however, the identification of treatment options is most often derived from genomic data, precursor states of proteins. Mass spectrometry (MS) and reverse phase protein assays (RPPA) may be used as a proxy for pathway activation. Nevertheless, both approaches have technical limitations. Although RPPA is sensitive to low levels of protein expression and requires less sample material, there are only a few hundred antibodies available. MS, as opposed to this, is able to quantify thousands of proteins but is less sensitive, yet more complex in its application. As an alternative, several statistical approaches model pathway activity using molecular data. Most of these approaches, however, use only gene expression data and only a few integrate more than two data types. The aim of my project is the integration of multiple molecular and non-molecular data types to predict the activity of proteins and pathways.

22: Copy number phylogenetics for single cells

Jack Kuipers

ETH Zurich

A comprehensive picture of the genomic aberrations that occur during tumour progression and the resulting intra-tumour heterogeneity, is essential for personalised and precise cancer thera- pies. Single-cell sequencing offers resolution down to the level of individual cells and is playing an increasingly important role in this field. High-throughput methods now allow the profiling of thousands of cells, in particular for copy number alterations through shallow whole-genome sequencing. For such data, we present CNTree: a statistical model and an MCMC algorithm to infer the evolutionary history of copy number alterations of a tumour. In evaluations on simulated data, we show the accuracy of this approach in inferring the copy number profiles of each individual cell, and we demonstrate its applicability to real sequencing data.

23: Network-constrained biclustering of patients and multi-omics data

Olga Lazareva

Technical University of Munich

Unsupervised learning approaches are frequently employed to identify disease-associated genes. In particular, biclustering is a very powerful technique that can be used when the objective is to cluster genes along with patients. It can be used to stratify patients while directly revealing biomarker genes that support the stratification. However, the genes selected in this approach may not necessarily share a joint biological function or participate in the same pathway, making it difficult to interpret these finding.
We propose a novel network-constrained biclustering approach with the aim to obtain interpretable subsets of genes for stratifying patients. We developed a metaheuristic algorithm BiGAnts which uses Ant Colony Optimisation to cluster genes and patients simultaneously with the constraint that genes inside a bi-cluster are also functionally connected in a gene interaction network. We analysed the performance of BiGAnts algorithm in various terms and found that BiGAnts clustered patients in agreement with known subtypes while at the same time discovering phenotype-specific genes that are connected and can thus be interpreted with regards to their biological function.

24: Assessing the effect of germline and somatic mutation on gene expression changes in 1,188 human tumours

Kjong Lehmann

ETH Zurich

We analyze matched whole-genome sequencing and RNA-seq data from the TCGA and ICGC repository to assess the regulatory effects of germline and somatic variants. We map expression quantitative trait loci (eQTL) of germline variants identifying 3,532 genes with a germline effect (FDR<5%), including prominent cancer genes. A systematic comparison for replication of this eQTL map to regulatory variants in matched normal tissues from GTEx, resulted in 422 cancer-specific eQTLs. Using this germline map, we quantify the amount of the regulatory effect of germline mutations in different genomic regions and contrast it to the effect of somatic mutational burden. We also assessed the regulatory effect of somatic mutations using allele-specific expression (ASE) analyses and QTL mapping. In particular, we map fine-grained spatial categories of somatic mutational burden on ASE. We identify 649 somatic eQTL (FDR<5%) among which several genes have known roles in the pathogenesis of specific cancers. We also considered associations between global mutational signatures and gene expression changes. This study enabled us to assess the transcriptional landscape of 1,188 cancer patients.

25: ShrinkIso: a mixed-effects model for differential exon usage suggests differential splicing between cancer types

Renee Menezes

Amsterdam UC, location VUmc

Alternative splicing increases the variety of transcripts produced by a single gene. In particular, different transcripts may be involved in different cancer types, yielding clues about the biological processes in the background. Data to study differential splicing often consists of exon-level counts obtained from RNA-seq data. Methods to find differential splicing ideally take counts for all exons belonging to a single gene into

account.
We propose such a method, ShrinkIso, based upon a mixed-effects model. It includes a exon-group interaction representing the random effect of interest. This allows for selection of effects with a minimum variance, whereupon multiple testing correction is computed. The model is robust to both differential total gene expression and exon-specific effects, such as from poly-A tails. It allows for zero-inflated data and inter-exon correlation. Fitted by empirical-Bayes, it can consider exon-bridging read counts to help improve estimates.
Analysis of breast, kidney and colon cancer data shows that results are reproducible even for small datasets. Differential exon usage between these cancer types suggest the different ways these cancer types develop.

26: Feature selection based on dose profiles for radiation response biomarker research

Anna Papiez

Silesian University of Technology

Machine learning techniques have the potential for unraveling a great amount of knowledge regarding the molecular mechanisms underlying response to ionising radiation. One of the most common and important applications of ionising radiation is radiotherapy in cancer treatment. The focus of this research is to combine statistical and machine learning tools in application to a high-throughput biological data set on ionizing radiation response. A customized approach based on statistical tools and current knowledge of biological mechanisms was tested on a radiation study on breast cancer patients. The group consisted of radiosensitive and radioresistant patients whose blood samples were exposed to high doses of ionising radiation. Gene profiles were applied as a filtering factor to adjust data using linear interpolation, allowing for efficient classification in a multiple random validation setting. The implementation of integrative techniques combined with custom data interpolation between doses led to successful determination of potential biomarkers of radiation response, which have been confirmed with an independent computational approach and literature study.

27: Using smartphone app data to infer personal eating patterns

Nick Phillips

EPFL

The global epidemic of obesity and the related metabolic syndrome presents a major risk to the global population. Personalised medicine has the opportunity to tackle obesity and metabolic syndrome by harnessing new advances in diagnostic technology and cutting-edge machine learning techniques. The core aim of this project is to collect clinical and molecular data from human subjects and use Bayesian statistics to investigate the links between the internal state of an individual's circadian clock (also termed chronotype), eating habits, control of body weight and possible reversal of metabolic syndrome. As part of an ongoing chrononutrition study we are collecting clinical data on the health/disease state of a cohort of volunteers, which includes a smartphone app recording subjects' eating patterns. Here we show how the probability of eating through the day can be described on an individual basis using Gaussian processes. We will subsequently add transcriptome analysis (RNA-seq) of blood samples to uncover the relationship between eating patterns and individual chronotype.

28: *to be announced*

Natalia Pietrosemoli

Institut Paris

Lassa fever (LF) is a major public health issue in Western Africa recently classified as an epidemic threat requiring urgent preparedness by the WHO. There is neither treatment with demonstrated efficiency nor a licensed vaccine to fight against this deadly emerging virus, thus there is an urgent need of developing a vaccine against Lassa virus (LASV). Here, we analyze transcriptomic and proteomic profiles of peripheral blood mononuclear cells in LASV-infected non-human primates immunized with a single shot of different LF candidate vaccines to compare the different vaccines' efficacy. The most efficient candidate will enter phase I clinical trials in the coming months. We apply different omic integration strategies, from result integration (i.e. analyzing transcriptomic and proteomic datasets by separate and comparing their results) to joint analysis approaches, including multi-omic pathway enrichment analysis. Then, we systematically compare these results and evaluate their efficacy in providing a systems-level understanding of the molecular basis of the vaccine's response as com-

pared to the classical independent analysis of transcriptomic and proteomic datasets.

29: Estimating perturbation profiles from patients mutation and gene expression data with a causal network approach.

Martin Pirkl

ETH Zurich

A vast collection of data has been assembled in The Cancer Genome Atlas (TCGA). Different data types are available across patients. The integrative analysis of those types is still an unsolved problem. We introduce a novel approach to jointly analyse gene expression and mutation data.
A mutation can be viewed as a perturbation of a specific gene. Driver mutations are identified, which are assumed responsible for the emergence of cancer. However, many patients in the same group of cancer patients have only a subset or none of the most prominent genes mutated. Our assumption is, that these genes are still perturbed in those patients, but not directly via a mutation.
We use a causal network approach to impute this missing perturbation information. We achieve this by combining the labeled mutation data and the corresponding gene expression to learn a network of the perturbed genes. We use the learned network to impute the missing data and revise known labels. We show in simulations that our approach performs better than state of the art methods. Additionally we apply our method to cancer types available in TCGA.

30: Deciphering single-cell heterogeneity in melanoma for clinical decision support

Michael Prummer

ETH Zurich

With the advent of large-scale profiling of cancer samples, it is now clear that no two tumors are the same and that they can evolve very differently on the molecular and cellular level. Fortunately, scRNAseq today allows for the profiling of molecular

phenotypes of individual cells from culture, biopsies, and whole organisms. We have established a clinical diagnostics pipeline for personalized treatment recommendations for melanoma patients based on single cell RNAseq. The successive steps are as follows: (i) a robust sample preparation protocol to process single-cell suspensions; (ii) 10X Genomics droplet-based sequencing library preparation followed by Illumina NGS; (iii) Cellranger preprocessing and normalization; (iv) cell-cycle correction; (v) classification of melanoma and immune cell types and cell type-specific gene lists; (vi) unsupervised clustering; (vii) differential expression and gene set analysis; (viii) treatment prediction using known gene-drug associations. Our integrated single-cell platform for the analysis of tumor biopsies in a validated and rapid processing pipeline demonstrates the feasibility of the approach for use in real-world molecular tumor boards.

31: Hierarchical inference for genome-wide association studies

Claude Renaux

ETH Zürich

The goal is to perform high-dimensional statistical inference for genome- wide association studies (GWAS). We develop meta analysis for multiple studies and novel software in terms of an R-package hierinf. Inference and assessment of significance is based on very high-dimensional multivariate (generalized) linear models: in contrast to often used marginal approaches, this provides a step towards more causal-oriented inference.

32: *to be announced*

Heba Sailem

University of Zurich

Characterisation of gene functions in a context-dependent manner is crucial for characterisation of gene roles in cancer progression. High Throughput Imaging (HTI) allows the determination of gene function by monitoring phenotypic changes following genetic perturbations. However, a framework for a comprehensive analysis of HTI

datasets is still lacking due to challenges in data analysis and the complexity of gene functions. To address these challenges, we developed a Knowledge-Driven Machine-Learning framework (KDML) to automatically link gene quantitative phenotypes to various biological functions in a context-dependent manner. We apply KDML to a genome-wide siRNA screen in a colorectal cancer cell line where we extensively profile cell morphology and context. We show that KDML can identify genes associated with many biologically relevant phenotypes such as cell protrusion and adhesion which we validate using orthogonal biological databases. Through an integrative analysis we link KDMA predictions to colon cancer patient outcome. In summary, KDML is a flexible and systematic framework for analysing HTI datasets and identifying context and tissue-dependent gene functions.

33: Joint Mean and Dispersion of Real-time Continuous Glucose Monitoring (CGM) Data

Agus Salim

La Trobe University

Real-time Continuous Glucose Monitoring (CGM) devices have been shown to be beneficial in diabetes management. Using inserted sensors, the device returns reading of blood glucose regularly every few minutes. CGM data has been used to estimate measure of glucose variability such as global standard deviation. However, there is growing recognition the timing and as well as the magnitude of glucose variability is important. To estimate time-specific glucose variability, we develop a joint mean-dispersion model where effect of factors such as medication and exercise on the mean glucose levels is modelled while simultaneously modelling the residual variance as a function of 24-hour time using B-splines and residual covariance is modelled using AR (1) structure. Using UVA/PADOVA Type 1 Diabetes Simulator, we show that our model has acceptable mean absolute relative deviance (MARD). Subsequently, using real dataset of 88 patients, our model shows that patients with similar mean and overall glucose variability can have different short-time and long-term outcomes when the timing of their main glucose variability differs.

34: Omics-based tool kit for the diagnosis of metabolic diseases

Sascha Sauer

Max Delbrück Center for Molecular Medicine

The incidences of metabolic diseases such as obesity and type 2 diabetes are rapidly increasing worldwide. We explored the potential of "omics" technologies to detect even slight physiological deregulation by integrating transcriptome, proteome, and metabolome data sets derived from metabolic target tissues. Using network and pathway analyses, we identified tissue-specific hub proteins and deregulation of major metabolic pathways in disease- and treatment-states. Moreover, application of our web-tool PHOXTRACK, which applies non-parametric statistics to calculate whether defined kinase-specific sets of phosphosite sequences indicate statistically significant concordant differences between physiological conditions, enabled us to identify key active regulatory proteins. Importantly, conceptually similar protein set analyses offered the potential to assess the efficacy and but also the cardiovascular side effects of the diabetes drug rosiglitazone, which was not possible with any other currently popular (pre-) clinical assays. The here shown robust statistical analyses of integrated "omics data" shall allow for supporting medical decision and for more efficiently developing drugs.

35: Modelling cancer progression using Mutual Hazard Networks

Rudolf Schill

University of Regensburg

Motivation: Cancer progresses by accumulating genomic events, such as mutations and copy number alterations, whose chronological order is key to understanding the disease but difficult to observe. Instead, cancer progression models use co-occurence patterns in cross-sectional data to infer epistatic interactions between events and thereby uncover their most likely order of occurence. State-of-the-art progression models, however, are limited by mathematical tractability and only allow events to interact in directed acyclic graphs, to promote but not inhibit subsequent events, or to be mutually exclusive in distinct groups that cannot overlap. Results: Here we propose Mutual Hazard Networks (MHN), a new Machine Learning algorithm to infer cyclic progression models from cross-sectional data. MHN model events by their spontaneous rate of fixation and by multiplicative effects they exert on the rates of

successive events. MHN compared favourably to acyclic models in cross-validated model fit on four datasets tested. Availability: Implementation and data are available at `https://github.com/RudiSchill/MHN`.

36: Predicting amyotrophic lateral sclerosis from genotype using deep, regulatory principle guided neural network architectures and protocols

Alexander Schönhuth

Centrum Wiskunde & Informatica Amsterdam

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease, for which a major part of heritability has remained unexplained. ALS is believed to have a complex genetic architecture where non-additive combinations of variants constitute disease. Unlike linear models, deep learning is highly promising for identifying such complex relations. Guided by recent insight about disease-associated variants, we have designed a deep neural network architecture that leverages convolution and mobile vision techniques for implementing genetics principles.
We evaluated our approach on the Dutch cohort ($>$ 11,000 individuals) of the Project MinE data. Our approach reveals novel, hitherto overlooked combinations of variants, beyond confirming known ALS-associated variants. In general, our approach clearly outperforms all alternative approaches in predicting ALS status from individual genotype. By raising a principled, reproducible protocol, we establish a first, promising step towards integration of deep learning methodology in GWAS, of particular interest when investigating genetically complex diseases such as ALS.
See also doi.org/10.1101/533679 and github.com/byin-cwi/ALS-Deeplearning

---

37: Predicting cancer genes with graph convolutional networks

Roman Schulte-Sasse

Max-Planck Institute for molecular genetics

The availability of experimental data in cancer research has rapidly increased in the last years, while in-silico methods still focus on prediction of oncogenes or tumor sup-

pressor genes (TSGs) from mutation data. This is despite the fact that genes might have different roles, depending on tissue & cancer type and usually without leveraging the power of multi-omics data. We present graph convolutional networks (GCNs) in the context of disease gene prediction and show how they can combine protein-protein interaction (PPI) networks, mutation information, DNA methylation at promoters and gene expression to predict cancer genes. The algorithm uses graph relations together with high-dimensional feature vectors for it's prediction and is trained on a set of known cancer genes. We show that our model is able to confidently recover known cancer genes from multi-omics data. Further, we can employ feature interpretation techniques on a trained model to find important features and neighbors in the PPI network, thereby gaining knowledge about the mechanisms driving cancer in different cancer types and tissues.

38: Detecting regulatory genetic variants with transcription factor binding affinity test

Sunyoung Shin

University of Texas at Dallas

Understanding the regulatory roles of non-coding genetic variants has become a central goal for interpreting results of genome-wide association studies. The regulatory impact of the variants may be interrogated by assessing their influence on transcription factor binding. We propose an efficient and scalable motif-based method for identifying potential regulatory variants, named atSNP (affinity testing for regulatory SNP detection). atSNP implements an importance sampling algorithm coupled with a Markov model for the background sequences to evaluate motif matches to both reference and variant alleles and assess variant-led changes in motif matches. Further, we have developed atSNP Search (`http://atsnp.biostat.wisc.edu`), a comprehensive web resource for evaluating the regulatory significance of human variants and visualizing motif alternation by the variants. Users can test more than 37 billion variant-motif pairs with marginal significance in motif matches or alteration. Computational evidence from atSNP, when combined with experimental validation, may help with the discovery of disease mechanisms. This poster is based on Zuo, Shin, Keles (2015) and Shin et al. (2018).

39: Discovery of mutation-specific synthetic lethals from large-scale pan-cancer perturbation screens

Sumana Srivatsa

ETH Zurich

A pair of genes forms a synthetic lethal (SL) pair if an aberration in either gene is innocuous to cellular viability but aberrations in both are detrimental to the cell. In oncology, synthetic lethal partners of mutated genes pose as good drug targets, and hence are critical in light of treatment options. Despite several large-scale cell-line based perturbation experiments, identifying SL pairs remains a challenge. Here, we present SLIdR (Synthetic Lethal Identification in R), a statistical method implemented in R for predicting mutation-specific SL partners in both pan-cancer and cancer-specific settings from large-scale perturbation screens. We identified 104 pan-cancer SL pairs including some well-established SL pairs. Further, we show that pan-cancer hits exhibit differential sensitivities based on the primary site of the tumor, and therefore we applied SLIdR to 17 different cancers and identified both novel and well-known cancer-specific SL pairs. We performed a comprehensive experimental validation of SL pair between mutated AXIN1 and SL partner URI1, thus presenting a prospective new candidate for hepatocellular carcinoma treatment.

40: The genetic landscape of mitochondrial disease: exploration of over 2000 cases by whole exome sequencing

Sarah Louise Stenton

Technische Universität München

Mitochondrial diseases (MD) pose a diagnostic challenge due to clinical and genetic heterogeneity, propelling unbiased WES into early diagnostics. To date, 326 disease genes implicated in mitochondrial energy metabolism are recognised, and this number continues to grow. Through global collaboration, we assimilate data from 2116 patients investigated under suspicion of MD. Systematic analysis establishes a genetic diagnosis in 952 (45%). Variants in 635 (30%) cases harbour within 167 MD genes and in 317 (15%) cases within 225 non-MD OMIM disease genes, highlighting their vast genetic underpinning and phenotypic mimicry of other genetic diseases. Reaching a diagnosis paves the way for development of efficacious treatment and is fast becom-

ing prerequisite for clinical trial inclusion. Currently, 40 cofactor metabolism defects potentially amendable to specific supplementation are recognised, encompassing 130 of our cases. However effective defect-targeted treatments for the majority are missing. I hope attendance to Ascona to provide invaluable knowledge on integration of molecular and health (HPO) data, to further aid diagnosis and elucidate the heterogeneous genetic landscape of MD.

41: Coherence-based Time Series Clustering for Statistical Inference and Visualization of Brain Connectivity

Ying Sun

KAUST

We develop the hierarchical cluster coherence (HCC) method for brain signals, a procedure for characterizing connectivity in a network by clustering nodes or groups of channels that display a high level of coordination as measured by "cluster-coherence." While the most common approach to measure dependence between clusters is through pairs of single time series, our method proposes cluster coherence which measures dependence between pairs of whole clusters rather than between single elements. Thus it takes into account both the dependence between clusters and within channels in a cluster. The identified clusters contain time series that exhibit high cross-dependence in the spectral domain. Simulation studies demonstrate that the proposed HCC method is competitive with the other feature-based clustering methods. To study clustering in a network of multichannel electroencephalograms (EEG) during an epileptic seizure, we applied the HCC method and identified connectivity in alpha (8,12) Hertz and beta (16,30) Hertz bands at different phases of the recording: before an epileptic seizure, during the early and middle phases of the seizure episode. An R Shiny App is also developed.

42: *to be announced*

Chen Suo

Fudan University

Data analysis framework, both commercial and open source, are limited for supporting the detection of cerebral microbleeds (CMB) on Magnetic resonance imaging (MRI) sequences. In this study, we present a novel algorithm for the automated estimation of CMB. The proposed methodology is a combination of classical decomposition of apparent diffusion coefficient (ADC) distribution into a Gaussian mixture model (GMM) with k-means clustering subsequently performed on the parameters of mixture model components. The maximum conditional probability criterion gives the final threshold estimated. The developed signal analysis pipeline was applied to the problem of CMB segmentation, with a dataset of 318 elderlies with MRI. Additionally, a comparison to the obtained CMB regions identified by a human expert was performed.

43: A two-axis latent variable model for unsupervised integrative clustering

David Swanson

University of Oslo

We propose a Bayesian parametric model for integrative unsupervised clustering across data sources. Its unique statistical formulation introduces two co-occurring sets of latent variables acting along perpendicular axes of the concatenated data matrices. We condition the latent variable space on the event that these two sets are equivalent. Formulating the two-way latent structure model in this way allows samples to be clustered in relation to each specific data source, while encouraging cluster information sharing between sources. A common scaling across data sources is therefore not required, and what is otherwise a complex latent structure can be sampled with closed form posteriors. We improve Gibbs sampling of the posteriors with a warm start strategy and modified and improper density functions to robustify convergence. Posterior interpretation gives insight into the degree of pan-genomic clustering at sample-level granularity. We analyze a multi-omic Norwegian breast cancer cohort and find little pan-genomic clustering, suggesting that models assuming a common clustering across data sources might yield misleading results.

44: Identification of disease-associated immune cell populations in systemic sclerosis

Mehdi Taslimifar

ETH Zurich

Authors: Mehdi Taslimifar*, Edoardo Galli*, Mike Becker*, Gabriela Kania, Rudnik Michal, Burkhard Becher†, Oliver Distler†, Manfred Claassen† *contributed equally † contributed equally
Systemic sclerosis is an autoimmune disease of connective tissue and internal organs, with involvement of various immune cell subsets. While these subsets are indicative about the disease mechanism and of diagnostic value, it remains to map out these subsets in the affected tissue and in the periphery. We conducted immune cell profiling of peripheral mononuclear cells of a cohort of diseased and healthy patients to identify novel immune cell signatures of systemic sclerosis patients. Specifically, we considered a cohort of 26 diseased and 15 control patient samples and performed mass cytometry. We correlate the clinical information of the patients with our previously developed representation learning approach, CellCnn, to identify the disease-associated cell subsets. We identify two peripheral blood immune cell subsets in the natural killer and myeloid compartment to be disease-specific.

45: Exploring the functional effects of genetic risk scores in pseudotime

Shu Mei Teo

Baker IDI Heart and Diabetes Institute

Many complex human diseases are caused by a combination of genetic and environmental exposures that interact with age of an individual. Time-series molecular data obtained from longitudinal studies is ideal but challenging to conduct at a large scale. Recent advances on the inference of pseudotime trajectories from single time-point data has shown that latent temporal information could be extrapolated from cross-sectional data sets, therefore offering new insight into dynamic disease processes without longitudinal studies. In parallel, increasingly large genome-wide association studies have enabled the construction of genomic risk scores (GRS) that represent an individual's risk of disease at birth. Yet, the molecular pathways by which GRS confers disease risk have not been explored. I use covariate-aware pseudotime approaches to model cross-sectional surveys of the human proteome and its interactions with GRS of multiple complex human traits and diseases. Proteins with significant GRS-pseudotime interac-

tion effects are of interest; These represent proteins whose trajectories are modulated by the genetic susceptibility to specific diseases.

46: A new pipeline for the preprocessing and statistical analysis of CRISPR-cas9 datasets

Renaud Tissier

Amsterdam UMC

Gene knockout screens can help understand and identify biological processes affecting cell growth. Genetic CRISPR-cas9 pooled screens use several guide RNAs (gRNAs) per gene for knock-out, for a large number of genes. The read-out is cell fitness per gRNA at different time points (T0 and one or more later time points), assessed by DNA sequencing. Analysis of such data too often yields results hard to reproduce, however. We propose a new pipeline to pre-process and analyse CRISPR screen data in order to improve reproducibility and yield better false discovery rate control. We normalize the data using assay controls in a robust way. Then, different impact on cell fitness between conditions is inferred by using a mixed-effects model which allows to test for both gene knock-out effect and guide-RNA specific variation, using all time points. Estimates are obtained using an empirical Bayesian approach. We apply this pipeline to two datasets, one publicly available and one from the Cancer Center Amsterdam. We show that commonly used approaches for do not control the FDR consistently across cell lines. Our approach, in contrast, yields adequate FDR control and more reproducible results.

47: V-pipe

Ivan Topolsky

ETH Zürich

Viruses are both important models for evolutionary biology and causes of severe infectious diseases, thus representing major public health and economic concern. Viral genetics based on next-generation sequencing (NGS) of viral genomes is now the method of choice for analyzing the diversity of intra- and inter-host virus populations, including epidemiological studies and individual treatment optimization in clinical virology.
Here we present the latest developments of V-pipe.
V-pipe is a bioinformatics pipeline that integrates various computational tools, that we have developed to support the computational analysis of viral NGS data. It is freely available at `https://cbg-ethz.github.io/V-pipe`. It enables the repro-

ducible analysis of genomic diversity in intra-host virus populations, including quality control, read alignment, and inference of viral genomic diversity on the level of single-nucleotide variants and viral haplotypes.

48: MelArray: an integrated targeted sequencing panel to assess mutations, copy number, structural variants, and immunological features in melanoma for clinical decision support

Patrick Turko

Universitätsspital Zürich

Immune checkpoint inhibitors and targeted therapies have improved survival in patients with metastatic melanoma. The detection of specific mutations and genomic features is necessary for therapeutic stratification. Commercially available gene panels cover frequently mutated genes in melanoma, but additional genes can be useful to determine which patients might qualify for specific clinical trials, off-label or compassionate therapies, and as prognostic markers. We designed a gene panel covering 190 genes relevant for melanoma progression and therapy. Our target capture design includes over 4000 exons and 28 introns as well as intergenic probes of heterozygous SNPs for CNV estimation. We have established a complete protocol consisting of DNA capture, Illumina sequencing, and customized bioinformatics post-processing to detect somatic SNVs, CNVs, and common fusion breakpoints in melanoma. We further estimate total tumor mutation burden (TMB), a recognized marker of immunotherapy response. We discuss the implementation of the MelArray platform at two institutions (Yale University and University of Zurich) and the sequencing of samples from 83 patients

---

49: Prediction of genes producing circular RNAs based on genomic and epigenetic features

Alena van Bömmel

Max Planck Institute for Molecular Genetics

Circular RNAs (circRNAs) are a novel class of RNAs which are associated with many deseases and are believed to function as potential biomarkers. However, circRNA biogenesis remains poorly understood. To investigate their formation, we identified circRNAs by RNA-seq and further mapped the chromatin occupancy of modified forms of RNA Polymerase II (RNAPII S2p,S5p and S7p) by ChIP-seq in mouse embryonic stem cells (mESC). We detected approximately 100 genes producing circRNAs (circ+) in mESCs. Furthermore, RNAPII is altered at the promoter of genes producing circRNAs (circ+), when compared with genes not producing circRNAs (circ-) in mESCs. To understand the causes underlying altered RNAPII patterns at circ+ genes, we are exploring the epigenetic landscape of circ+ promoters. Moreover, we build a classifier with elastic net for circ+ genes based on the epigenetic marks and on predicted transcription factor binding on promoters. To identify the underlying mechanisms of circRNA formation, we investigate the most important features in the model. We found modified RNAPII S7p, cyclin dependent kinases and HCFC1 as the strongest predictors for circ+ genes.

50: Improving prediction for high-dimensional data: an Empirical Bayes approach to co-data learning

Mirrelijn van Nee

Amsterdam UMC

We consider generalised ridge regression in clinical prediction settings, in particular binary and survival, for high-dimensional data. We use complementary data (co-data, e.g. related studies, genomic annotation or cell line data) to define possibly hierarchical covariate groups (e.g. gene sets, known signatures, Gene Ontology trees) that may differ considerably in terms of predictive strength. If so, penalising these groups by different ridge penalties would improve prediction.
We present a CV-free Empirical Bayes approach to find estimates for the group penalties. In order to obtain stable group parameter estimates, we provide an extra level of shrinkage. Any type of shrinkage, e.g. ridge or hierarchical lasso, can be used at this level, rendering a new, flexible framework to improve predictions. Moreover, the framework allows for integration and weighting of multiple co-data sets, plus posterior variable selection. We show that the multi-group penalties stabilise variable selection, and improve the performance of parsimonious prognostic models.
We demonstrate the method on an application to cancer genomics and compare predictive performance with other commonly used methods.

51: Performance and stability analysis of sparse network-based survival models

Susana Vinga

INESC-ID

Extending sparse optimization to incorporate biological information has become a valuable framework to build interpretable models from high-dimensional patients data. Classical regularization models, such as elastic net, can be further constrained to include spatial, temporal or relational information of the covariates (gLasso, fused Lasso, Net-Cox, DegreeCox). We developed glmSparseNet, a Bioconductor R package that allows to include network information as a regularization function in generalized linear and survival models. We explored several centrality measures in protein-protein interaction networks to improve TCGA RNA-seq data Cox proportional hazards models. We evaluated their accuracy and stability by fitting six oncological datasets using 5000 cross-validation replicates, and by analysing the obtained log-rank tests and concordance indexes. The results show that network-based models are more stable and consistent in the variables selected, have comparable performance with elastic net, and may add interpretability by identifying interesting highly connected nodes (hubs) and gene sub-networks.

52: Distributed Estimation of Central Subspace

Kelin Xu

Fudan University

We consider the problem of estimating central subspace when data are stored in a distributed fashion across different places. Large data sets are often distributed across different places, uniformly analyzing these data involve data transformation, storage, and privacy problems: electronic health records are stored in different health agencies. A distributed algorithm of sufficient dimension reduction methods is investigated in this paper, wherein sliced inverse regression and cumulative slicing estimation are explicitly discussed. Sufficient dimension reduction is an efficient dimension reduction method

which captures all the information in the explanatory variables related to the response variable via the estimation of central subspace. We show that the estimation error rates of our proposed distributed estimation methods are the same with that using whole data, as long as the sample size on each server is not too small. Besides, we extend our findings to the heterogenous case. Our theoretical results are illustrated through extensive simulations.

53: Quantifying microbial dark matter and its impact on metagenome analyses

Elizabeth Yuu

Robert Koch Institut

Alterations in the microbiome are known to cause severe health problems such as autoimmune diseases and various cancers. For example, closely related bacterial species can present large phenotypic differences, even if they have very similar genomes. It is therefore crucial to focus on how similar species differ and how these deviations affect the microbiome. In metagenomics, mapping reads to reference genomes allows for insight into taxonomic compositions and variations between different microbial communities. We previously introduced DiTASiC (Differential Taxa Abundance including Similarity Correction) for shared read ambiguity resolution based on a regularized, generalized linear model (GLM) framework. This, and other similar approaches, does not address the remaining unmapped reads, or "microbial dark matter". We extend our approach by analyzing sub mappings with different error-tolerance and integrating dark matter variables in an effort to create a more appropriate GLM. This new idea has the potential to provide more accurate estimates of taxa abundance and inherent variation; this in turn can lead to improved taxa quantification and differential testing.

54: Deep neural networks predict drug-induced histopathology based on gene expression

Jitao David Zhang

F. Hoffmann-La Roche

Prediction and the molecular-level understanding of compound-induced organ toxicity are essential for preclinical safety evaluation in drug discovery. We leveraged gene-expression data in the Open TG-GATEs database and systematically compared predictive powers of machine-learning models of increasing complexities, including logistic regression, support vector machine, and deep neural network (DNN). We found that DNNs consistently and substantially outperformed other models for almost all types of liver histopathology. We applied the DNN models to independent datasets and confirmed high predictive power across technological platforms of gene expression profiling (microarray and RNA sequencing) and across rodent species (rat and mouse). Finally, we investigated ways to either simplify or enhance the best-performing DNN models. The present study demonstrates the feasibility and advantage of applying deep-learning techniques to predict drug-induced liver histopathology based on gene expression data in the context of preclinical studies of drug candidates. More applications and iterative refinement of our approach may reduce both attrition rates of drug candidates and animal use.

55: Fitness landscapes and cluster partitions for the Drosophila fly microbiome

Lisa Lamberti

ETH Zürich

Fitness landscapes, i.e. genotype-phneotype mappings, are an important concept in evolutionary biology. They are used to study origins and progressions of diseases such as HIV and cancer, to understand bacterial resistances to antibiotics, microbiome-host relations and much more. In these studies properties of fitness landscapes, such as epistasis, have important biological consequences. To fully grasp properties of fitness landscapes mathematical approaches are needed. Beerenwinkel et al.(2007) suggested studying fitness landscapes via regular subdivisions of convex polytopes. Building on their approach we propose cluster partitions and cluster filtrations of fitness landscapes as a new mathematical tool. In this way, we provide a concise combinatorial way of processing metric information from epistatic interactions. Using existing Drosophila microbiome data, we demonstrate similarities with and differences to previous approaches. As one outcome we locate interesting epistatic information where the previous approaches are less conclusive.

Based on joint work with H. Eble (TU Berlin), M. Joswig (TU Berlin) and W. Ludington (Carnegie Institution for Science) and work with A. Gould (University of California), V. Zhang (University of California), E. Jones (University of California), B. Obadia (University of California), N. Korasidis (ETHZ), A. Gavryushkin (University of Otago), J. Carlson (University of California), N. Beerenwinkel (ETHZ), W. Ludington (Carnegie Institution for Science).