Contributed afternoon talks

Cheng Li (Msc. ETHZ) 2ETCS
Using Bayesian Networks to Analyze Expression Data

Leonardi Carlo (Msc. ETHZ) 2ETCS
Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data

Sophie Seidel (Msc. ETHZ) 2ETCS
Methods for causal inference from gene perturbation experiments and validation

Joan Badia (Unige) 2 ETCS
Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge.

Tao Fang (MSc. ETHZ BSSE) 2ETCS
Machine learning based prediction of drug-induced histopathology with gene expression data and compound-target binding information

In silico prediction of drug-induced organ toxicity can empower preclinical safety evaluation of drug candidates. Gene expression profiling provides unbiased and comprehensive understanding of changes on pathway and biological-network levels, and therefore can be of great value for toxicity prediction. However, existing tools fail to leverage the large amount of available data, and more importantly, fail to consider other important data types such as compound-target binding information. In this study, we address the issue by developing a machine-learning algorithm training data in TG-GATEs, the largest public pharmacogenomics database, and data in CAT, the most comprehensive compound-target activity database that is proprietary to Roche, to predict drug-induced rat liver and kidney histopathology as recorded in TG-GATEs. To this end, we compared two cutting-edge machine-learning approaches: support-vector machines (SVMs), and deep neural networks (DNNs, also known as deep learning). First we used gene expression data alone to predict drug-induced toxicity with either SVMs or DNNs, and found that DNNs outperform SVMs for most types of histopathology. Furthermore, we found that the combination of gene expression data and compound-target binding information improves the prediction performance for some types of histopathology. Equipped with these knowledge, we prototyped a web-based tool that makes predictions based on user-uploaded gene expression data. Our pilot study demonstrates the feasibility and advantage of applying deep-learning techniques in the context of preclinical safety characterisation of drug candidates. It highlights the prediction power of multi-source data integration in this context. Further refinement of our approach may enhance our ability to early de-risk compounds with safety concerns and improve our capacity to deliver safe drug candidates.

Dériaz Denis (MSc. Université de Neuchâtel) 2ETCS
Estimated 24-h Urinary Sodium and Sodium-to-Potassium Ratio as Predictors of Kidney Function Decline in a Population-Based Study

The increasing prevalence of chronic kidney disease (CKD) worldwide underscores the importance of identifying risk factors for age-related kidney function decline. Sodium and potassium intakes are associated with CKD progression in the renal population, but little is known on the association between sodium and potassium intakes and changes in estimated glomerular filtration rate (ΔeGFR) in the general population. We explored the association of urinary sodium and potassium excretions with ΔeGFR in a population-based study with 5-years follow-up data. We estimated 24-h urinary sodium (eUNa), potassium (eUK) and sodium-to-potassium ratio (eUNa/K) using Kawasaki formulae. We performed multivariate linear regression models studying the association of eUNa, eUK and eUNa/K with yearly ΔeGFR, taking several covariates into account, including baseline eGFR and albuminuria.

We included 4141 participants with a mean eGFR decline of -0.6 (SD 1.7) ml/min/1.73m2 per year. In the fully adjusted model, high eUNa and eUNa/K, but not eUK taken alone, were associated with faster renal function decline in term of standardized coefficients: β=0.07 ($p<0.001$) and β=0.05 ($p<0.001$), respectively. This suggests that dietary sodium and potassium intakes may play a role in kidney function decline in the general population. Whether lowering sodium and increasing potassium in the diet may help preventing CKD needs further exploration.

Zhang Jitao David (F. Hoffmann-La Roche)
Invitation to apply PGM to model drug-induced pathology in rat

I would like to present a case study of drug-induced pathology in rat, and invite collaborations to apply probabilistic graphical models to infer molecular mechanisms underlying the observations.

Philipp Mekle (Uni Basel)
Epidemics Nowcasting Using Open Source Data

As an extension to existing classical public health monitoring schemes the internet, social and news media have increasingly been used to explain patterns in disease activity and progression. Out of a real-world private equity investment context, the presentation shows the main elements of an exploratory study on infection forecasting, monitoring and nowcasting, based on available literature, public health and company data. In particular, the intent was to better understand open-source data indicators (OSDI) used for epidemiological nowcasting; nowcasting being defined as employing OSDI to compensate for the 1-4 week lag needed to collect and disseminate information via public health reports. The demonstration cases chosen are from the areas of influenza, dengue and foodborne epidemics. Among the topics addressed are issues about data relevance, use of hidden variable models, model forecasting capacity, modelling with scant/uneven data, and interpreting ex-post information.

Padvitski Tsimafei (Cologne)
Improving transcription factor targets list by using coexpression information

Inferring direct targets of transcription factor (TF) from Chip-seq data is an important step of reconstructing gene regulatory networks. However linking TF binding sites to target genes is known to be a difficult task. Many computational methods that address peak to gene assignment were developed, but improvement of list of already predicted targets is less explored field. We suggest that improving list of targets can be done by using external gene co-expression information. We tested our hypothesis by applying Network Smoothing to the results of TF target gene calling, smoothing scores over a co-expression network derived from large-scale single-cell RNA-seq datasets. We then evaluated performance of the approach by using TF knock-out expression data, assuming that true TF targets should be stronger affected upon TF perturbation. We were able to show that the method allows to improve a list of predicted targets.

Davidson Natalie (ETHZ)
Integrating Diverse Transcriptomic Alterations to Identify Cancer-Relevant Genes

BACKGROUND Previous multi-cancer genomic studies have focused on the analysis of somatic mutations as the driver of phenotypic changes. Here, we propose and apply a novel method to integrate a wide variety of transcriptomic aberrations in combination with DNA-level changes to redefine the concept of driver events and account for the role of the transcriptome in tumorigenesis. We present a novel analysis that 1) identifies cancer driver genes through a recurrence analysis over diverse types of transcriptomic alterations 2) identifies frequent and heterogeneous transcriptomic alteration signatures in 1,188 samples across 27 histotypes as part of the PanCancer Analysis of Whole Genomes (PCAWG) of the International Cancer Genome Consortium (ICGC). We integrated the following alteration types: expression outliers, alternative splicing outliers, gene fusions, alternative promoters, non-synonymous variants, RNA-editing, and allele-specific expression. RESULTS To identify the cancer relevant genes, we created a new method for performing recurrence analysis on transcriptomic features. Our method has three main strengths: flexibility to handle any number or type of alteration, sensitivity to different

frequencies of alterations, and ability to prioritize genes with heterogeneous alterations. The method has four main steps: 1) binarize each alteration type by identifying rare/outlier events; 2) sum alteration events over samples for each gene; 3) transform counts to ranks for alteration comparability; 4) combine ranks across alterations to create ranking scores. We performed >1M permutations to identify a cut-off for informative scores; our analysis yielded 1,012 genes with an empirical p-value <0.05. These genes show a 2.82-fold enrichment for cancer census genes (CGC) and driver genes (provided by ICGC Driver Working group;PCAWG-9) with a p-value of 5x10-26 (hypergeometric test), signifying that our analysis is identifying cancer-relevant genes (Fig. 1). Among the top 5% of our ranked genes is CDK12, which is impacted by multiple, but non-overlapping, types of alterations within a protein kinase domain associated with dysregulation of DNA repair in cancer. In our cohort, we find 87 samples that have an alteration in this domain, with 64 (74%) samples having only a RNA alteration in the domain. The most frequent alteration is an alternative promoter event that leads to a truncated transcript of CDK12, removing a majority of the kinase domain. Fusion and splice events lead to additional disruptions of the same domain. CDK12 exemplifies the value of integrating RNA and DNA alterations. To understand the functional impact of these alterations, we compare alteration patterns across cancer types and cancer-specific pathways. We notice that alteration patterns vary between the cancer types (Fig. 2): Chromophobe renal cell carcinoma (Kidney-ChRCC) in comparison with Skin-Melanoma has significantly different numbers of non-synonymous variants (t-test; p-adj.: 1.42x10-5), copy-number alterations (p-adj.: 6.70x10-4), fusions (p-adj.: 1.56x10-4), and splice outliers (p-adj.: 7.05x10-10). In contrast, similar cancers like Kidney-RCC and Kidney-ChRCC only differ in the amount of non-synonymous variants (t-test; p-adj.: 5.50x10-25). Comparing across cancer-specific pathways, we find that the TOR and metabolism pathways are more impacted by RNA alterations. We also find that of the 578 samples with an altered p53 pathway, typically associated with high non-synonymous variants, 131 (22.7%) of them carried only RNA alterations. This is further evidence that neglecting transcriptomic alterations could underestimate the degree of cancer pathway disruption. CONCLUSIONS Through our joint analysis, we integrated both DNA and RNA aberrations in a recurrence analysis that yielded a list of promising genes highly enriched for known cancer driver genes. Furthermore, we identified associations of transcriptomic alterations to cancer type and DNA-level aberrations, helping to broadly explain the influence of the transcriptome on various cancer-related processes.

Sima Ana Claudia (Zurich University of Applied Sciences and University of Lausanne)
Bio-SODA: a unified search engine across heterogeneous biological databases

Biological databases are growing at an exponential rate, currently being among the major producers of Big Data, almost on par with commercial generators, such as YouTube or Twitter. While traditionally biological databases evolved as independent silos, each purposely built by a different research group in order to answer specific research questions, more recently significant efforts have been made toward integrating these heterogeneous sources. However, searching through biological databases still remains a largely manual and time-consuming process. We argue that an integrated biological data access system, coupled with a powerful, multi-modal search interface that spans multiple data sources, similar to what web search engines provide today, can open the path for novel insights into the unified data, beyond the capabilities currently offered by each individual database.

Kaie Kubjas (MIT)
Mathematics of Chromosome Packing in Diploid Organisms”

The 3D organization of the genome plays an important role for gene regulation. Chromosome conformation capture techniques allow one to measure the number of contacts between genomic loci that are nearby in the 3D space. In this talk, we study the problem of reconstructing the 3D organization of the genome from whole genome contact frequencies in diploid organisms, i.e. organisms that contain two indistinguishable copies of each genomic locus. This talk is based on joint work with Lawrence Sun and Caroline Uhler.

Lisa Lamberti (ETHZ)
Higher order interactions and polyhedral subdivisions

Understanding the effects of interactions for instance between modified species or mutated genes is a central problem in biology. In this talk I explain how polyhedral subdivisions help to describe and quantify interactions and present recent developments and open problems in the field.

Duarte Eliana (Otto von Guericke Universitat Magdeburg)
Defining equations of probability tree models

Staged are a new and exciting class of statistical models that generalize the well known Bayesian Networks. In this talk I will explain the algebraic and statistical properties of these models and characterise the case in which they are toric varieties trees.

Orlando Marigliano (MPI)
A Graphical Model for Soil Stability

This talk is about an application of graphical models to ecology. I walk through a software implementation of a structural equation model described in a study linking plant community properties to soil aggregate stability. Basic theoretical concepts of graphical models and causality are discussed using this example. I compare the model to experimental data and explain how causal interventions can be used to suggest further experiments.

Kandasamy Saravanan (Indian Institute of Science)
Learning and Testing Causal Models with Interventions

We consider testing and learning problems on causal Bayesian networks as defined by Pearl. Given a causal Bayesian network M on a graph with n discrete variables and bounded in-degree and bounded ``confounded components'', we show that O(log n) interventions on an unknown causal Bayesian network X on the same graph, and n/eps^2 samples per intervention, suffice to efficiently distinguish whether X=M or whether there exists some intervention under which X and M are farther than eps in total variation distance. We also obtain sample/time/intervention efficient algorithms for: (i) testing the identity of two unknown causal Bayesian networks on the same graph; and (ii) learning a causal Bayesian network on a given graph. Although our algorithms are non-adaptive, we show that adaptivity does not help in general: log n interventions are necessary for testing the identity of two unknown causal Bayesian networks on the same graph, even adaptively. Our algorithms are enabled by a new subadditivity inequality for the squared Hellinger distance between two causal Bayesian networks.

Machlab Dania (University of Basel)
Computational Identification of Transcription Factors Driving Changes in Chromatin Accessibility

ATAC-seq data offers the opportunity to identify accessible chromatin regions, which are believed to be created by transcription factors binding to DNA, potentially in concert with chromatin remodeling enzymes. Comparing two conditions that express a different set of transcription factors, for example developmental stages or different genotypes, allows us to view changes in chromatin accessibility that reflect the underlying differential activity of these factors. We applied the stability selection method developed by Meinshausen and Bühlmann, to select for transcription factors that are likely to explain the observed changes. Randomized Lasso Stability selection involves subsampling randomly from the response vector (changes of accessibility) and performing lasso regression using transcription factor binding sites in accessible regions as predictors. This results in each transcription factor obtaining a selection probability to explain the observed changes in accessibility. When we apply the method on a simulated data set, we see increased accuracy in selecting for true signal and more correlated variables compared to other linear regression methods.

Coons Jane (North Carolina State University)
The Cavender-Farris-Neyman Model with a Molecular Clock

We give a combinatorial interpretation of the toric ideal of invariants of the Cavender-Farris-Neyman model with a molecular clock on a rooted phylogenetic tree and prove results about the polytope associated to this toric ideal. For instance, the number of vertices of this polytope is a Fibonacci number and the facets of the polytope can be described using the combinatorial structure of the underlying rooted tree. The toric ideal of invariants of this model has a quadratic Groebner basis, and we use this Groebner basis in special cases to give a unimodular triangulation of the associated polytope with number of simplices equal to an Euler zig-zag number. Finally, we show that the Ehrhart polynomial of these polytopes, and therefore the Hilbert series of the ideals, depends only on the number of leaves of the underlying tree, and not on the topology of the tree itself.

Arthur Bik (University of Bern)
Polynomials of bounded strength

Storing a general homogeneous polynomial f in K[x_1, ... ,x_n] of degree d is difficult as n -> oo. The number of values we need to remember is a polynomial p(n) of degree d. However if we only consider polynomials f that can be expressed as the sum of at most k products of polynomials of lower degree, then the number of values that we need is at most a polynomial of degree d-1. In this talk, I will discuss joint work with Jan Draisma and Rob Eggermont stating that we can bound the number of products we need uniformly if f comes from a family of nicely related Zariski-closed subsets of homogeneous polynomials of degree d indexed by finite-dimensional vector spaces.

Karren Dai Yang (MIT)
Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions

We consider the problem of learning causal DAGs in the setting where both observational and interventional data is available. This setting is common in biology, where gene regulatory networks can be intervened on using chemical reagents or gene deletions. Hauser and B\"uhlmann (2012) previously characterized the identifiability of causal DAGs under perfect interventions, which eliminate dependencies between targeted variables and their direct causes. In this paper, we extend these identifiability results to general interventions, which may modify the dependencies between targeted variables and their causes without eliminating them. We define and characterize the interventional Markov equivalence class that can be identified from general (not necessarily perfect) intervention experiments. We also propose the first provably consistent algorithm for learning DAGs in this setting and evaluate our algorithm on simulated and biological datasets.

Max Pfeffer (MPI MiS Leipzig)
Joint PCA and Clustering of Multiple Biomedical Datasets

Large amounts of data arise in many biological experiments. Often, several different measurements are performed in order to explain one phenomenon. For instance, cancer types are classified by looking at the gene expression, the protein expression, mutations in the genome, copy numbers and more. These datasets can be of very different kind, i.e. they can vary in size, scale differently and so on. Therefore, classification of cancers is often done by investigating the datasets individually and then combining the obtained knowledge with much biological intuition. We aim at analyzing the datasets simultaneously, by performing a generalized PCA. This method essentially finds a subspace that is most common to all datasets. Differences in scale and size are not prohibitive. We then perform a simple k-means clustering to obtain first preliminary results. This is done on data for the uveal melanoma, as this allows for easy comparison with existing results.