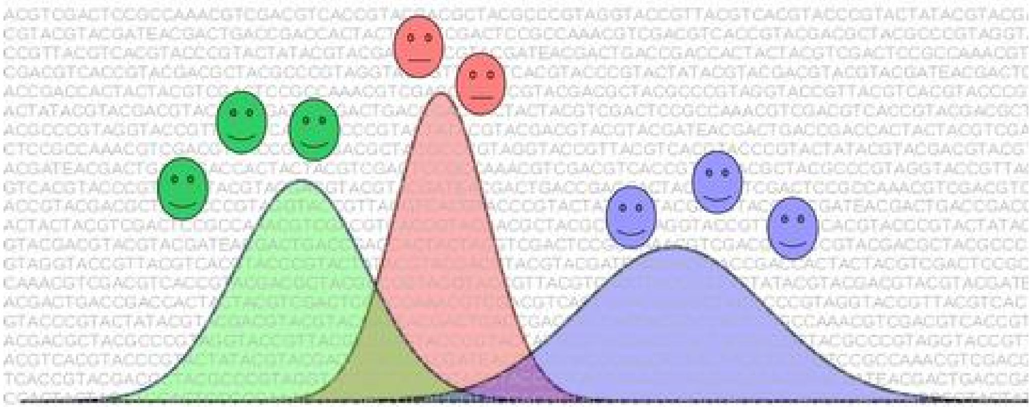
Centro Stefano Franscini

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Statistical Genomics and Data Integration for Personalized Medicine

Ascona, 12-17 May 2013



Organized by:
Niko Beerenwinkel
Peter Bühlman
Darlene Goldstein

Abstract

Advances in biotechnology have made genome-scale measurements routine, but using the vast amount of generated data to improve human health remains challenging. Medicine is moving from a ‘one-drug-for-all’ paradigm to more individualized treatment. This interdisciplinary workshop aims at the dissemination of cutting-edge biotechnological and statistical developments and at the identification of open data analysis problems in personalized medicine.

I progressi biotecnologici hanno reso oramai standard le analisi genomiche, ma ancora difficile utilizzare i dati ottenuti nella pratica medica. Dal modello ”un farmaco per tutti”, la medicina sta passando a trattamenti individuali. Questo convegno si propone di divulgare i pi recenti progressi biotecnologici e statistici, e di discutere le sfide della medicina personalizzata.

NOVARTIS

Roche

MerckSerono
Living science, transforming lives

SIB
Swiss Institute of
Bioinformatics

Centro
Stefano Franscini

Venue

Monte Verità
Via Collina 84
CH-6612 Ascona
tel. +41 91 785 40 40 - fax +41 91 785 40 50

About the Centro Stefano Franscini (CSF)

The Centro 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 Verita. 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 organized 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.) organized in the context of its international conferences and/or Monte Verità's cultural programme.

Travel directions

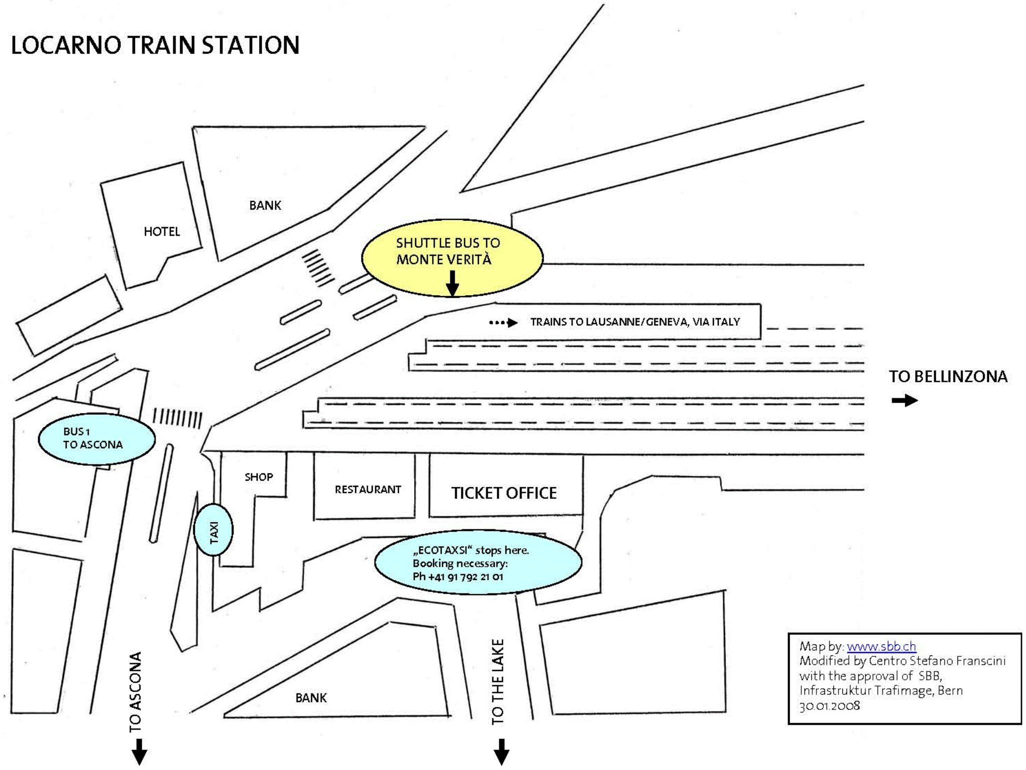
Please refer to the directions reported on the webpage of the CSF.

External hotel

Some of the participants have been accommodated in Hotel Luna. It is located in Ascona, approximately 1km from the Centro. A shuttle service will drive the hotel guests to the conference, as detailed below.

Shuttle service from Locarno Station

A free 11-seater shuttle bus to Monte Verità will leave from Locarno railway station Sunday 12th May according to the following schedule:



Map by: www.sbb.ch
Modified by Centro Stefano Franscini
with the approval of SBB,
Infrastruktur Trafimage, Bern
30.01.2008

Figure 1: Shuttle bus location

Sunday 12.05.2013 **Stazione di Locarno - MV**	note
15.20	stop Luna
16.00	stop Luna
16.40	stop Luna
17.20	direct MV
18.05	direct MV
19.05	direct MV

This shuttle bus will be filled on a first-come-first-serve basis.
Moreover, two additional runs will connect hotel Luna to MV and back.

18.00	hotel Luna → MV
21.15	MV → hotel Luna

The shuttle meeting point is on the right side of the train platforms in Locarno (see image). During the workshop, the shuttle will run according to the following program

Monday 13.05.2013	
08.40	hotel Luna – MV
21.15	MV – hotel Luna
Tuesday 14.05.2013	
08.40	hotel Luna – MV
21.15	MV – hotel Luna
Wednesday 15.05.2013	
08.40	hotel Luna – MV
PM	ristorante Grotto Broggini – hotel Luna
Thursday 16.05.2013	
08.40	hotel Luna – MV
21.15	MV – hotel Luna
Friday 17.05.2013	
08.40	hotel Luna – MV
after lunch	MV – stazione Locarno

Excursion and social dinner

The excursion will take place Wednesday 15th May in the afternoon: we will visit Valle Verzasca. Afterwards, we will go to Ristorante Grotto Broggini for the social dinner.

Keynote lectures

Removing Unwanted Variation in Personalized Medicine

Terry Speed

WEHI, jointly with Johann Gagnon--Bartsch and Laurent Jacob, Department of Statistics, University of California at Berkeley

In this talk I shall assume that the version of personalized medicine under consideration involves a molecular assay on a sample from a relevant tissue, say a microarray gene expression profile on a tumor biopsy. Such assays are prone to being affected by a variety of forms of unwanted variation, such asthose due to sample quality, sample processing and storage, and assay reagents, operators, equipment and environmental conditions. As a result, steps will usually be taken at the time the assay results are analyzed and interpreted, to adjust for or remove unwanted variation. We have recently discussed a variety of methods of doing this in the context of differential expression of molecular species, methods which are equivalent to the use of certain random effects models. In this talk I will briefly describe these methods, and then outline our efforts to adapt them to the situation most likely to be met in the context of personalized medicine, where samples are processed individually or in batches.

Towards personalized signaling networks - Modeling with unknown unknowns and ignored knowns

Rainer Spang

University of Regensburg

In cancer, mutations interrupt and modulate the propagation of signals in both signalling and gene regulation networks. The modulation can be different from patient to patient. Importantly, it can determine whether a targeted drug will be effective for this patient or not. In a personalized medicine setting we can not reconstruct entire networks but need to focus on some key features of them. Moreover we can only generate very limited amounts of data for an individual patient. I will present a simple inference method 'No-Conan' that partially reconstructs network features in a way that can not be confounded by unobserved confounders.

A Clinical Trial Design for Constructing and Evaluating Individualized Real-Time Treatment Policies

Susan A. Murphy

Mobile devices, including mobile phones, are increasingly used to both passively and actively collect patient symptoms, where the patient is, who the patient is with, level of social activity. At the same time, mobile devices are beginning to be used to deliver a variety of real-time behavioral interventions (motivational assistance, cognitive assistance, suggestions concerning social interactions). However only a few researchers have begun to use the real-time patient information to adapt and re-adapt the behavioral interventions to the patient. And for the most part, this adaptation is primarily based on behavioral theory, clinical experience and expert opinion. Data-based evidence is, at best, indirectly used in this process. In this talk we sketch out the outline for, and solicit feedback on, a new clinical trial design for the purpose of providing/using patient data to inform the development of Individualized Real-Time Treatment Policies.

Solving statistical inference problems: (1) perturbation cell biology and (2) evolutionary couplings in proteins.

Chris Sander

MSKCC, New York

(1) We perform combinatorial perturbation experiments with rich molecular readout using agents, such as targeted drugs, and then optimize predictive accuracy and model simplicity. Resulting models are used to design combinatorial interventions for investigational or therapeutic purposes, identify novel drug synergies, discover drug specificity spectrums or redesign cellular circuits for synthetic biology.

(2) Addressing the well-known challenge of translating protein sequence into three-dimensional structure, we use a powerful approach, adapted from statistical physics, to calculate evolutionarily couplings between amino acid residues in a protein family. We then use these couplings to identify functionally constrained residue interactions and, using distance geometry and simulated annealing, to predict protein 3D structure, including transmembrane proteins. Project co-leader: Debora Marks, Harvard Medical School; co-authors (alphabetical): Lucy Colwell, Thomas Hopf, Andrea Pagnani, Burkhard Rost, Robert Sheridan, Riccardo Zecchina. See http://bit.ly/tob48p (PDF) and www.EVfold.org. The discovered evolutionary couplings provide insight into essential interactions constraining protein evolution and, with the rapid rise in large-scale sequencing, are likely to facilitate a comprehensive survey of the universe of protein structures by a combination computational and experimental technology. Applications to cancer genomics relate to the interpretation of the functional impact of cancer-related mutations and the design of targeted therapeutics.

Pathway - Based Personalized Analysis of Cancer

Eytan Domany

We introduce Pathifier - an algorithm that infers pathway deregulation scores for each tumor sample on the basis of expression data. This score is determined, in a context-specific manner, for every particular data set and type of cancer that is being investigated. The algorithm transforms gene level information into pathway level information, generating a compact and biologically relevant representation of each sample. We demonstrate the algorithm's performance on three colorectal cancer datasets and two glioblastoma multiforme datasets, and show that our multi-pathway-based representation is reproducible, preserves much of the original information, and allows inference of complex biologically significant information, such as pathways that were significantly associated with survival and new cancer sub-classes.

Integrative analysis of *-seq datasets for a comprehensive understanding of regulatory roles of repetitive regions

Sunduz Kelez

University of Wisconsin, Madison

A fundamental question in molecular biology is how cell type specific gene expression programs are established and maintained through gene regulation. Main drivers of cell-specific gene expression are regulatory elements (e.g., promoters, transcription factor (TF) binding sites, chromatin/epigenomic marks, enhancers, silencers). Identifying genomic locations of these elements and unraveling exactly how they control gene expression in different cell types has been a major challenge. The ENCODE projects have generated exceedingly large amounts of genomic data towards this end. A formidable impediment to comprehensively understanding of these ENCODE data is the lack of statistical and computational methods required to identify functional elements in repetitive regions of genomes. Although next generation sequencing (NGS) technologies, embraced by the ENCODE projects, are enabling interrogation of genomes in an unbiased manner, the data analysis efforts by the ENCODE projects have thus far focused on mappable regions with unique sequence contents. This is especially true for the analysis of ChIP-seq data in which all ENCODE-adapted methods discard reads that map to multiple locations (multi-reads). This is a highly critical barrier to the advancement of ENCODE data because significant fractions of complex genomes are composed of repetitive regions; strikingly, more than half of the human genome is repetitive. We present a unified statistical model for utilizing multi-reads in *-seq datasets (ChIP-, DNase-, and FAIRE-seq) with either diffused or a combination of diffused and point source enrichment patterns. Our model efficiently integrates multiple *-seq datasets and significantly advances multi-read analysis of ENCODE and related datasets.

Statistical issues in studying tumour heterogeneity

Anestis Touloumis, John Marioni, and Simon Tavaré

It is now common to sample a solid tumour from multiple spatial locations. Analyzing the genomic and transcriptomic variation in these samples is crucial for understanding the nature of tumour heterogeneity and its implications for treatment and relapse. We have been developing statistical models for assessing expression data in this setting. "Transposable data" refer to random matrices where the rows and the columns correspond to variables of interest and dependencies occur among and between these variables. We have developed nonparametric tests for the form of a high-dimensional covariance matrix that describes the row (column) dependence structure while treating the column (row) dependence structure as a 'nuisance'. In simulations, we observed that the proposed tests preserve the nominal level when the null hypothesis is true and are powerful against the alternative hypotheses tested. We apply the proposed tests to a glioblastoma cancer dataset to study the dependence structure between genes when multiple measurements of gene expression levels are available for each subject.

Viral Genetiic Linkage Analysis in Cluster Randomized Trials for HIV prevention

Rui Wang, Ravi Goyal, Vlad Novitsky, and Victor DeGruttola

Havard School of Public Health

Currently under development are two cluster-randomized trials (CRTs) sponsored by CDC, NIH and USAID to evaluate the effectiveness of HIV prevention strategies in reducing HIV incidence in different settings in Africa. We discuss the goals and the design and analytical challenges for one of these studies-the Botswana Combination Prevention Project (BCPP). Goals include investigation not only of the effect of interventions on HIV incidence but also on patterns of viral genetic linkage, the analysis of which is essential for investigating HIV transmission dynamics. When interventions are scaled up, it will be important to tailor interventions to these patterns. To do so, it is important to identify host characteristics (e.g. compliance with interventions, disease status, and demographics)that are associated with genetically-linked infections. A complicating factor in this analysis is that the probability of obtaining a sequence may vary with personal characteristics; bias in inferences regarding viral genetic clustering that arise from sampling can be reduced by using new methods for accommodating missing data. Such analyses will be most useful if the CRTs are properly powered, but in CRTs, a major driver of power is the intraclass correlation which is not generally known. , Correlation structure for HIV infection endpoint is driven in part by the sexual network; correlation between partners would be expected to be higher than that among individuals who are far apart in the sexual network. We investigate analytically and through simulations how different sampling strategies and mechanisms for generating correlation would affect study power. We develop a new method to generate a robust collection of sexual networks utilizing both the estimated features of the mixing matrix and its sampling variability. Viral genetic linkage analysis will provide information regarding the nature of sexual networks that can be used to improve this process, both during the study (where simulation studies will be used in making decisions about study conduct) and after its conclusion.

How predictive is our DNA?

Ceceile Janssens

Emory University

The rapid and continuing progress in gene discovery for complex diseases is fuelling interest in the potential implications of this knowledge for clinical and public health practice. One of the prominent expectations is that preventive and therapeutic interventions can be more effectively administered when they are targeted to individuals on the basis of their genetic risks. An essential prerequisite for such applications is that DNA has appreciable predictive ability. The number of studies assessing the predictive ability is steadily increasing. This lecture will give a review of the recent developments in genetic risk prediction studies and a preview on the predictive ability of future DNA testing including whole genome sequencing: how well can we predict diseases when we know all genetic risk factors?

When is Reproducibility an Ethical Issue? Genomics, Personalized Medicine, and Human Error

Keith Baggerly

MD Anderson Cancer Center

Modern high-throughput biological assays let us ask detailed questions about how diseases operate, and promise to let us personalize therapy. Careful data processing is essential, because our intuition about what the answers 'should' look like is very poor when we have to juggle thousands of things at once. When documentation of such processing is absent, we must apply 'forensic bioinformatics' to work from the raw data and reported results to infer what the methods must have been. We will present several case studies where simple errors may have put patients at risk. This work has been covered both on the front page of the New York Times and by CBS' 60 Minutes, and has prompted several journals to revisit the types of information that must accompany publications. We discuss steps we take to avoid such errors, and lessons that can be applied to large data sets more broadly.

Talks

Towards the next generation malaria vaccine

Alena van Boemmel [1], *Terry Speed* [2], *Alyssa Barry* [2]

[1] *Max Planck Institute for Molecular Genetics,* [2] *Walter and Eliza Hall Institute of Medical Research*

Although an intensive international effort over the last decades has led to a discovery of several vaccine-candidate antigens of Plasmodium falciparum, a broadly effective malaria vaccine is still missing. One of the main reasons is the existence of high levels of genetic diversity of the parasite and a low prevalence of vaccine alleles in the parasite population (1-25%). Therefore, for the design of a malaria vaccine it is very important to understand the population structure and the genetic diversity of the malaria parasite.

In our study, we analyze the distribution of alleles of the Merozoite Surface Protein 1 (MSP1), which is the most studied Plasmodium falciparum antigen. Specifically, we investigate 1 indel and 11 common (minor allele frequency ¿10%) non-synonymous single nucleotide polymorphisms (SNPs) in the 42kDa MSP1 domain, which correspond to genetic regions with signatures of balancing selection. Altogether, 236 samples from infected patients in 8 worldwide locations were collected. Amongst this dataset, we identified 48 distinct parasite “haplotypes” comprised of different combinations of SNPs. Further, we have developed a strategy to find the most appropriate combination of up to 4 haplotypes, which fulfill the following 3 criteria: an overall high prevalence in studied locations, a small variance of the prevalence and a large genetic distance among the haplotypes. With these criteria we were able to detect the best combination of haplotypes, which could be included to provide the broadest possible coverage of the worldwide parasite population for the next generation of malaria vaccines to be more effective against several genetic strains of the parasite. To find the best combination of the haplotypes we optimized a linear function of the central prevalence tendency, the dispersion measure of the prevalence and of the Hamming distance. The optimal combination found with our model has an average prevalence of 67% in the parasite population and outperforms herewith the current vaccines haplotypes with maximal prevalence of 25%.

A comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer

Levi Waldron

Harvard School of Public Health and Dana-Farber Cancer Institute

Background: Numerous gene signatures of patient prognosis for late-stage, high-grade ovarian cancer have been published, but diverse data and methods have made these difficult to compare or use in a clinically meaningful way. We sought to identify successful prognostic gene signatures for ovarian cancer through systematic validation using public data.

Methods: This effort involved creating a database of uniformly processed and curated public ovarian cancer microarray data and clinical annotations, and systematic identification and re-implementation of published prognostic models. This enabled an objective assessment of 14 prognostic models published between 2007 and 2012 across 10 independent datasets totaling 1,455 late-stage, high-grade, serous ovarian cancer patients.

Results: This work addresses several important issues for the translation of genomics to clinical application: 1) the ability to independently implement and validate published prognostic models, 2) the accuracy of published prognostic models when applied to new, independent datasets, 3) similarities between independently developed prognostic models, 4) the influence of popular datasets on the literature, and 5) the prognostic ability of random gene signatures. This analysis provides definitive support for the prognostic ability of a handful of the proposed models, but also confirms that these require improvement to be of clinical value.

Conclusions: Reproducible and user-friendly frameworks and resources are provided for all analyses, to encourage independent analyses and provide a template for other such large-scale systematic evaluations. This study clarifies the literature of prognostic ovarian cancer gene signatures and addresses published controversies [1,2] by quantitatively placing these studies within their broader context.

[1] Baggerly et. al. J Clin Oncol. 2008 Mar 1;26(7):1186-7 [2] Swisher et. al. J Natl Cancer Inst. 2012 May 2;104(9):642-5

Inferring causal molecular intermediates from omics data in the context of genetic and environmental variations

Julien Gagneur

Gene Center, LMU, Munich

Dissecting the molecular mechanisms that link genotype to phenotype promises to deliver the necessary insights to develop drugs tailored to the genetic background and life circumstances of the patient. Information from interventional data is scarce, and hence the challenge resides in developing causal inference strategies to exploit the breadth of observational population-level genetic and molecular profiling data being generated.

Here we investigated to what extent environmental perturbations, combined with genetic variations, facilitate causal inference in molecular networks. Using yeast as a model system, we carried out joint profiling of fitness and gene expression of a genetically diverse population in 5 environmental contexts. We developed novel inference techniques to predict molecular functional intermediates with an environment-specific role on growth. Our approaches leverages on ubiquitous genotype-environment interactions, exploiting the rich statistical independencies they imply. Technically, we build on Bayesian model comparisons, assessing the statistical evidence that a particular transcript carries a mediating role between genetic signal and its environment-specific effect on phenotype. We applied the approach to genome-wide identified transcripts specific for each environment-specific growth QTLs. Comprehensive independent test using the genome-wide deletion collection confirmed the majority of the 400 top-ranking model predictions. Our results show that exploiting condition-specific genetic effects substantially increases the predictive accuracy over approaches based on genetic or environmental variations alone.

Together, these results have wide-ranging implications for the design of clinical omics studies and their integrated analysis across multiple contexts. Furthermore, the dataset is a unique resource to test different inference strategies as, for the first time, large-scale perturbational data for matching conditions is available.

Extensive Variation in Chromatin States Across Human Individuals and Populations

Judith Zaugg

Stanford University

The vast majority of disease-associated variants lie outside protein-coding regions, suggesting that variation in regulatory regions may play a major role in disease predisposition. However, despite extensive studies of gene expression differences between individuals, little is known about the regulatory mechanisms responsible for these differences. Here, we study differences in regulatory regions across 19 diverse individuals by systematic mapping of chromatin state using six histone modifications, cohesin, Pol2 and CTCF in lymphoblastoid lines. We find extensive regulatory region differences in both activity (strong vs weak vs poised) and identity (enhancers vs promoters vs repressed regions). Differences in modifications are inherited in trios and correlate with gene expression differences, indicating that they have functional consequences. Overall, our results provide fundamental insights into genetic and epigenetic differences of humans.

Systematic inference of transcriptional and network markers of drug sensitivity in cancer cell lines

Francesco Iorio

EMBL - European Bioinformatics Institute and Wellcome Trust Sanger Institute

Oncogenic lesions tend to occur in a restricted number of pathways, and alterations in the same pathway tend not to co-occur in the same patient. This leads to the idea that if a gene involved in an oncogenic/crucial pathway is altered, a second alteration in the same pathway does not provide additional selective advantages to the cancer cell.

Based on this we sought to identify pathways or pathway proxies showing statistically significant mutual exclusivity in the corresponding patterns of mutation across the panel of cell lines of the Genomic of Drug Sensitivity in Cancer (GDSC) project. We then correlated their mutational status (at a network level) with drug responses. This unveiled new network markers for drug sensitivity/resistance with a significant improvement of interpretability and predictive power with respect to single-oncogene markers.

This was obtained by integrating a set of high-quality (based on sequencing depth and mutant allele frequency) genomic variants comprising non-synonymous mutations, homozygous deletion and amplifcations, for a total number of 34,799 variants, with large manually curated signalling maps.

To this aim we made use of a state-of-the-art randomisation algorithm preserving genomic-event distributions both across genes and samples. We analytically derived and used a new upper bound for this algorithm, which is hugely lower than the currently used one, through an R package developed on purpose.

Additionally we designed DoRothEA (Discriminant Regulon Enrichment Analysis): a method that, from a set of genome-wide gene expression profiles, infers transcription factors whose basal activity is associated with underlying factors. Through this, we inferred associations between 2,138 (known or putative) human transcription factors and the response of the cell lines in the GDSC project to 130 drugs. Results provide the basis for the identification of up-streaming pathways modulating drug response, thus dissecting the molecular mechanisms involved in drug resistance. They will be made publicly available via a web resource providing interactive statistical plots and data downloaded.

Metagenomic Diagnostics on the Species Level

Bernhard Renard

Research Group Bioinformatics, Robert Koch-Institut

The reliable detection of a bacterial or viral pathogen in a complex sample is a major challenge in metagenomic based diagnostics. Reference genomes are subject to constant genomic change and rarely complete. Further is the accurate detection hindered by the high genomic similarity between species. The majority of existing approaches either quantify at low resolution (e.g. at phylum level), or rely on the existence of specific genes, or have severe problems discerning species with highly similar genome sequences. We developed Genome Abundance Similarity Correction (GASiC), a versatile method to estimate true genomic abundances in metagenomic datasets on the species. Within our approach, metagenomic sequence reads are first mapped against a set of reference genomes of species potentially present in the dataset. Then, we estimate the pairwise similarities of the reference genomes using a simulation approach. The similarities are then used to correct the mapping results from the first step and to obtain estimates of the true genomic abundances in the dataset. To this end, we formulate the problem as a non-negative LASSO. Further, we study how changes in the sequence coverage of a specific genome can serve as indicator for incorrect metagenomic assignments and propose a mixture model approach for detecting these artifacts. We applied our method to the standard benchmark datasets and compared its performance to existing methods, showing that reduces the quantitative error by up to 60% even in the presence of sequence similarities of 95% and above.

Trio- and population-aware variant calling in the 'Genome of the Netherlands'

Alexander Schönhuth

Centrum Wiskunde and Informatica

The 'Genome of the Netherlands' (GoNL) is a Dutch collaboration that aims at characterizing and cataloging genetic variation in the Dutch population. The resulting variant catalogue will decisively add to our understanding of genetic variability in the context of population history, acquisition of de novo variation and GWAS interpretation. 250 Dutch families, consisting of trios and mono-/dizygotic twin quartets, have been sequenced to 12x coverage per individual, which, in terms of abundance and arrangement of data, is unique. Here, we will report on results of the analysis of structural variants and indels. We provide an overview of findings and report on key indices. We focus on ancestry- and population-aware variant calling in particular and present novel, statistical approaches by which to combine high-precision variant calling with Mendelian statistics and linkage disequilibrium. Indels of size 20 to 100 bp pose particular statistical challenges, both in ancestry-related and generic settings. We present MATE-CLEVER (Multiple-Allele-Attentive Clique-Enumerating Variant Finder) as a tool that tackles this challenge by combining enhanced algorithmic engineering with Bayesian statistics.

Joint analysis of multiple cancer molecular profiles: connecting the needles in the haystack

Renee X. de Menezes

Department of Epidemiology and Biostatistics, VUmc, Amsterdam, The Netherlands

Consider a study in which samples have whole genome profiles produced, such as genomic copy number and gene expression. Such data can help understand which changes in gene dosage affect gene expression and can, thus, help explain phenotype. Statistical models have been proposed to perform such an analysis (Louhimo et al., NatMeth, 2012). In addition to gene dosage, there is also interest in including other types of molecular profiles in the model, such as methylation and microRNA expression, as these can also be involved in gene expression regulation. Methods proposed so far to handling more than two high-dimensional data sets simultaneously are very computational intensive, making them impractical. In addition, no methods have so far been proposed to model conditional effects, for example to study the conditional impact of methylation on gene expression, given copy number change.

We extend the model first proposed by Menezes et al (BMC Bioinf, 2009) to handling multiple molecular profiles as explanatory sets of variables. This extension involves generalizing the score test first proposed by Le Cessie and van Houwelingen (Biometrics, 1995) to the case where there are multiple sets of covariates, which can be efficiently calculated. In addition, as a regression framework is used, it is possible to correct for confounders as by Menezes et al (2009). We also extend the method to test for conditional associations between molecular profiles, via a two-step approach involving a correction of the association pattern of given molecular profiles using ridge regression, followed by a score test for the conditional effect of the measurements of interest. We apply this approach to a subset of the publicly available colon cancer data set from The Cancer Genome Atlas. Our joint analysis of gene expression, copy number, methylation and microRNA confirms already known associatons, and identifies new ones.

Integrating multi-layered data for biomarker discovery in Melanoma

Jean Yee Hwa Yang

University of Sydney

Over the last decade, several statistical techniques have been proposed to tackle genome-wide expression data. However, with the advancement of many other high-throughput biotechnologies, the interest of researchers has been focusing on utilizing multiple data sources together with the clinical data, to improve the prognosis of disease outcome.

Integrating the components from different platforms has become a crucial step to better understand the relationships between clinical and -omics data and the information they provide to explain/predict some response. The statistical task to preserve the stability and interpretability of the model has become more challenging in this integration framework. One major issue is that the large dimension of -omics data can completely dominate the modelling procedure and it is an open question how to best combine different types of variables.

This work will outline, the integration of clinical and the availability of other high-throughput omics data to improve the prognosis capabilities and determination of biomarkers in stage III melanoma. This is based on a framework, which combines bootstrap sampling and multiple imputation (B-MI) to produce a model with good predictive properties. More specifically, we combined gene expression data, protein data and microRNA data, to explore methods in integrating -omics data using Lasso based methods.

Investigating a role for rare functional variation in HIV-1 control through exome sequencing

Jacques Fellay and Paul McLaren

EPFL School of Life Sciences

Genome-wide association studies have identified common variants associated with HIV-1 viral load and disease progression in the MHC and CCR5 regions, which together explain about 20% of inter-patient variability. A more comprehensive understanding of host genetic influences on retroviral control is required for more personalized medical care. Therefore, we here investigate the role of rare and functional variants in mediating HIV-1 control using exome sequencing. We captured and sequenced all coding exons to high coverage (¿ 70x) in 131 HIV-1 infected individuals of European ancestry using the SureSelect Human All Exon 50Mb enrichment kit and the Illumina HiSeq2000. Paired-end, 120 bp reads were aligned using the Burrows-Wheeler Aligner (BWA) and quality control and variant calling were performed using a combination of Picard and Samtools. Variant functional annotation was performed using snpEff version 2.1. As a correlate of HIV-1 control, we used viral load at set point (spVL), calculated as the average of at least 3 longitudinal measurements obtained during the chronic phase of infection, in absence of antiretroviral therapy. An average of 10,109 non-synonymous variant alleles were observed per individual. After accounting for genetic ancestry, age and sex, no association was observed between spVL and genome-wide burden of non-synonymous variants. In total, 358 genes were found to carry homozygous stop gain mutations in at least 1 individual. Comparing these to a list of genes with a reported HIV-1 protein interaction showed an overlap of 22 genes. No association was observed between presence of a homozygous stop-gain mutation in a known HIV-1 interacting gene and spVL. Addition of further samples and other testing methods (e.g. individual variant testing, collapsing of variants by gene and pathways) will be required to fully address the role of rare, functional sequence variants in controlling HIV-1 infection.

Challenges in development of practical omics biomarkers for risk prediction

Pratyaksha Wirapati

Bionformatics Core Facility, Swiss Insitute of Bioinformatics

In the past decade, many gene expression signatures had been proposed as biomarkers in cancer. However, only a handful managed to survive more extensive validations and used in practice. For some type of cancer, large sample size can be obtained by combining publicly available datasets to obtain more reliable signatures. Nevertheless, there are still challenges in translating such signatures into a specific, customized diagnostic platforms for daily clinical practice. Some of clinical requirements are: (1) ability to process single-sample data, rather than batches, with proper accounting of measurement uncertainty (2) inclusion of already well-established conventional clinical information and biomarkers into the prediction system, (3) relating the assay score to actual risk in the target population (4) relating the projected risk to clinical utility measures, for more flexible and individualized clinical decision making.

I will present a prototypical system that attempts to address the above issues, based on a current project in development of early-stage lung cancer prognostic marker, where signatures learned from a compendium of diverse expression microarray platforms need to be translated into practical FFPE assay technology, and incorporated into an “electronic cancer nomogram”.

An investigation of confidence intervals for binary prediction performance with small clinical trial datasets

Xiaoyu Jiang, Steve Lewtzky, Martin Schumacher

Novartis Institutes for BioMedical Research

Identifying predictive biomarkers has become a key subject in personalized medicine. The most common case is to predict for binary response, for instance, responders and non-responders to a given treatment, based on various -omic biomarker data. Building a predictive model in this context often suffers from small sample sizes in conjunction with unbalanced sizes of the two classes, as well as a large number of candidate features. This can cause resampling-based point estimates for predictive performance to be highly variable; hence, it is important to assess the characteristics of confidence interval estimates for predictive performance statistics as a function of several relevant parameters.

We use a double resampling scheme, namely, the Bootstrap Case Cross-Validation with Bias Reduction (BCCV-BR) method proposed in Jiang, Varma and Simon (2008), to construct confidence intervals for commonly reported predictive performance metrics, including positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, accuracy, Matthew's correlation coefficient (MCC) and area under the ROC curve (AUC). We perform an extensive simulation analysis to study the behavior of confidence interval estimates generated by BCCV-BR with different sample sizes, ratio of the sizes of the two classes, biomarker effect sizes, and number of candidate features. The empirical coverage probability, empirical power and confidence interval width are investigated. The results provide guidance regarding critical issues and decision making in predictive modeling with small clinical trial datasets.

Identification of cancer driver mutations based on recurrent mutation timing

Thomas Sakoparnig

ETH Zurich

Carcinogenesis is an evolutionary process which is driven by the accumulation of advantageous mutations in single cells and the subsequent outgrowth of those cells due to clonal expansion. Mutations in certain genes are present in a large fraction of cancers such as TP53 mutations, others exhibit high mutation rates in cancers of the same type such as BRCA1 in breast cancer. However, cancer genomes contain many more mutations which do not show high degrees of recurrence. There are several reasons for these low rates of mutations: (I) Mutations in some loci depend of the presence of mutations in other loci. However, cancer diagnosis and genotyping is usually not done at one specific stage of the carcinogenesis and hence mutations which are highly dependent on other mutations often have not yet occurred at the time of diagnosis. (II) Mutations in single members of functional entities (i.e. pathways) are often sufficient to disturb the pathways function and mutations within those pathways then display mutual exclusivity and low mutation rates across cancer samples. (III) Furthermore, cancer cell collect a large number of random passengers mutations in the process of carcinogenesis. These are mutations which randomly occur, however which are also manifested within a cancer cell population due to the co-occurrence with advantageous driver mutations. Here, we describe a method which aims to identify driver genes which have low mutation frequencies due to the mutational suppression in early cancer stages.

An empirical Bayesian ridge approach to modeling the transcriptional effects of DNA copy number aberrations

Gwenael G.R. Leday, Aad W. van der Vaart, Mark A. van de Wiel

VU University

DNA copy number aberrations are a hallmark of cancer cells. These aberrations, focal or broad, consist in gains and losses of chromosomal DNA. These may alter directly expression levels of mRNA transcripts that map to the aberration or indirectly those that are located outside. We here present a Bayesian multivariate model for the joint estimation of direct (in cis) and indirect (in trans) transcriptional effects of DNA copy number aberrations. Gene-wise, the model contains two parts: a low-dimensional vector of covariates for the cis-effects and a high-dimensional one for the trans-effects. Cis-acting effects are modeled by piecewise linear regression splines (Leday et al., 2013). This class of models combines copy number data from various steps of the preprocessing (namely the continuous segmented and discrete called data) and hence allows the effect of DNA on mRNA to differ across types of aberrations (e.g. loss, normal, gain and amplication).

The model thus provides good interpretability as to how the gene copy number affect its expression and for which samples. It improves power for detecting cis-effects that are specific for possibly small patient subgroups. The corresponding genes may be potential targets for personalized medicine. We argue for the shrinkage estimation of cis-effects to improve reproducibility. For the modeling of indirect trans-effects, we impose ridge priors to the high-dimensional vector of parameters. A novelty of this work is that parameters of priors are estimated empirically with the approach described by van de Wiel et al. (2013). In all, our model can be seen as a Bayesian graphical ridge that accounts for perturbation effects (DNA copy number). The amount of regularization is learned empirically, and may vary across genes. Sparsity is determined a posteriori through a model selection procedure

---

Accurate detection and clinical consequences of the clonal architecture in myelodysplastic syndromes

Moritz Gerstung

Wellcome Trust Sanger Institute

Myelodysplastic syndromes (MDS) are clonal haematological malignancies caused by mutations of the bonemarrow, and frequently transform into acute myeloid leukaemia. Here we present a study in which we sequenced 111 cancer genes in 738 MDS patients at an average coverage of 250x. We have developed a robust Bayesian approach exploiting sequencing depth, cohort size and prior knowledge for accurately detecting clonal and subclonal point mutations in each sample. By aggregating the reconstructed clonal architectures of each individual patient we show that mutually exclusive mutations in four different components of the spliceosome are among the earliest events in MDS pathogenesis and lead to phenotypically different subgroups. We demonstrate that subclonal mutations are frequent and generally as informative for survival predictions as clonal mutations. As the majority of mutations occurs in infrequently mutated genes, it is interesting to observe that the total number of point mutations in each sample is a powerful predictor of progression-free survival and independent of common survival scores. By training multivariate survival models using stability selection and a robust empirical Bayes estimator, we show that the cumulative prognostic value of point mutations and peripheral blood counts is as precise, but easier to obtain as conventional cytogenetics from bone marrow biopsies.

Posters

Significance of gene teams in cancer

Ewa Szczurek

ETH Zurich

Since the discovery of the causal, 'driver' genes mutated in several cancers, substantial scientific effort was invested to identify such drivers from cancer patient data and to distinguish them from their numerous, randomly mutated 'passenger' counterparts. One group of approaches is based on the fact that drivers come in groups, or 'teams' and are often connected by specific relations in the data. All of existent methods propose driver gene teams without assessing their statistical significance. In this work we formalize the relations of mutual exclusivity and concurrence that connect such driver gene teams. Based on this formulation, we next construct factor graph models in which likelihood of being a driver gene team is straightforward to compute. Finally, we propose two likelihood ratio-based tests for assessing the significance of a given gene team.

Classifiers learnt on microarray gene expression data - an illustration of the uses of cross-validation

Sarah Gerster[1] **, Charlotte Soneson*[1] ** and Mauro Delorenzi*[1] [2]

[1] *SIB Swiss Institute of Bioinformatics,* [2] *University Lausanne*

Background: Finding a common language and ensuring good communication is essential in multidisciplinary fields. To allow for effective collaborations, it is important to discuss how specific methods and tools from statistics can be used and understand their properties, advantages and disadvantages. The easiest way to do this is by illustrating a method on controlled data sets with known ground truth.

Focus: The current project focuses on applying cross-validation (CV) in the context of training classifiers on microarray gene expression data. The aim is to illustrate how CV should be applied in order to accurately estimate the prediction error of a classifier. We also show how a wrong design of the CV scheme or a bad choice of the error measure can lead to biased error estimates. A further aim of the study is to illustrate for which tasks CV is useful, and especially also which typical biases/confoundings in gene expression data cannot be revealed/alleviated by CV.

Data: We work on simulated data sets. To have intensity distributions similar to real data, we use a gene expression data set (PETACC3 cohort, 853 samples, 61528 features) as input. Random samples from this expression matrix are selected and assigned to group 1 (control) and group 2 (treated). After setting the mean of both categories to the same value, we add batch effects and select some features to be differentially expressed. In addition, a noise term is added to all features. We consider several scenarios for our study, including balanced and unbalanced data as well as different levels of confounding between groups and batch effects. Ten data sets (replicates) are simulated for each scenario.

Methods: We focus on well-known classifiers: logistic regression, SVM (Support Vector Machines), KNN (k-nearest-neighbors) and RF (Random Forests). Feature selection is performed with the Wilcoxon test or the Lasso. Parameter tuning and feature selection are embedded in the CV scheme. The prediction error of the each classifier is estimated within the CV scheme. This value is then compared to the error obtained when applying the classifier to independent data.

Feature Selection and Grouping for CBN Progression Structures

Simona Constantinescu

Department of Biosystems Science and Engineering, ETH Zurich

The evolution of biological systems is often subject to order constraints on the space of all possible events through which their progression is described. For example, it is widely known that the stages of cancer can be well characterized by the alteration of different sets of genes, relating to eachother through temporal constraints. The probabilistic graphical model Conjunctive Bayesian Networks estimates exactly this type of progression structures, by enforcing order constraints on the level of genetic events. The problem of optimal structure search is however computationally intensive, making it infeasible to be applied on a large number of events. In this work, we are focusing on selecting features (events) and grouping them in a biologically meaningful way. The groups consist of genes whose alterations bring the same contribution to the observed phenotype (also known as exclusive mutations). Under a constrained progression framework, only one of the events in each group needs to happen in order for the progression to move forward. The groups are selected and identified through a regularized linear regression method, termed pairwise fused lasso or clustered lasso. The independent variables are the binary - encoded events (genes), with a "1" representing alteration, while the dependent variable is the observed phenotype. In addition to the usual L1 penalty enforced on the absolute sum of the coefficients, the pairwise fused lasso penalty has an extra term, which sums over all the absolute pairwise differences between the coefficients of the regression. By weighting the difference terms with e.g. prior biological knowledge, certain coefficients become more likely to be grouped together. In this way, both sparsity and smoothness in the solution coefficients are enforced. The method is applicable on any large dataset, which can be meaningfully binarized. Moreover, it reveals different groupings and different order structures for patients with specific expression or mutation signatures, contributing with important knowledge in the improvement of patient - specific treatments.

Detecting synthetic genetic interactions in isogenic cell lines

Felix Klein

EMBL Heidelberg, Genome Biology Unit

The term synthetic genetic interaction describes the fact that only the combination of two genetic perturbations gives rise to a distinct phenotype, while each genetic perturbation alone does not. As it is often difficult to target the known oncogenes of cancer, the use of synthetic genetic interactions, especially synthetic lethality, provides a new approach for cancer therapy.

With the availabilty of high-throughput methods, genetic interactions can be systematically studied. We use a microscopy based approach to screen for synthetic genetic interactions in a panel of isogenic colorectal carcinoma cell lines. The first genetic perturbation in this case is manifested by the genetic background of each isogenic cell line and drug treatments are used as a second perturbation.

From the image data set we extracted summaries of geometric, morphologic and texture features and selected the most informative features for further analysis. To detect syntetic drug-cell line interactions we modeled the observed drug-cell line features using an additive model of overall effect, drug effect, cell line effect and residual term. Under the assumption that the residual term, representing the pairwise drug-cell line interactions, is mainly zero indicating no pairwise drug-cell line interactions, the overall effect, cell line effect and treatment effect were estimated using robust regression minimizing the residual terms.

On a single drug level we found several drugs that only show an interaction in one of the isogenic cell lines. Using hierarchical clustering we obtained clusters of drugs with similar interaction profiles. Some of these clusters were enriched for drugs that target similar pathways.

Further improvements in detecting synthetic genetic interactions are important. As a step towards personalized medicine synthetic genetic interactions might be used to predict optimal drugs for a given genetic background.

Probing of viral diversity by global haplotype prediction

Armin Töpfer

BSSE, ETH Zürich

The focus in clinical virology shifts from well-established identification of single nucleotide variants (SNV) to genome-wide probing of full-length viral RNA strains, called haplotypes. The success of antiretroviral HIV treatment heavily depends on the knowledge of an intra-patient's viral population heterogeneity, because diversity and in particular, low frequency variants affect virulence, immune escape, and drug resistance. With its poor reverse transcriptase enzyme fidelity, HIV produces daily up to 10 billion viruses, covering all possible mutations along the genome. In addition, the recombination rate of HIV is assumed to be ten times higher than the mutation rate, which leads to an even higher degree of genetic diversity. This combination of genetic change and selection, gives rise to a heterogeneous population, called quasispecies. Such swarm of closely related haplotypes is assumed to emerge from a few dominating haplotypes, called generators, subjected to mutation and recombination. We have devised a descriptive jumping hidden Markov model, capable of inferring an underlying quasispecies from error-prone second and third-generation sequencing data and able to predict the haplotype distribution. This models accounts for SNVs, by position-wise probability tables to explain the observed diversity, and recombination, by allowing a single observed read to originate from a mosaic of generators. We have implemented a Variational Bayes modified EM algorithm to compute maximum a posteriori estimates of the model parameters, model selection, and prediction of the haplotype distribution in a Java program called QuasiRecomb. QuasiRecomb is validated by simulation studies, to assess the advantage of explicitly taking the recombination process into account and furthermore validated on 454/Roche, Illumina, and PacBio sequencing data of a mix of 5 molecular HIV-1 viral strains. QuasiRecomb has also been applied to clinical studies of HIV-infected samples.

Bayesian inference of Quasispecies fitness

David Seifert

D-BSSE ETH Zürich

Manfred Eigen and Peter Schuster have developed quasispecies theory 35 years ago. They refer to a quasispecies as a cloud of evolving RNA species interconnected by mutation. While the notion of a viral quasispecies finds ubiquitous use in virology, the theory itself has never been supported by any significant experimental evidence. The central building block of quasispecies theory is the quasispecies differential equation, a non-linear equation describing the evolutionary trajectory of an infinite population of viral haplotypes. We present a first model to analyse clinical deep sequencing data of HIV positive patients within the framework presented by quasispecies theory, assuming stationarity of the viral ensemble. We have devised a Bayesian inference scheme to estimate the aforementioned viral haplotype fitness. The results of this analysis point to frequency as a weak predictor of fitness, that is, strains with a high fitness might not necessarily show a high abundance. This is in stark contrast to classical Darwinian evolution according to "survival of the fittest", where abundance predicts fitness and vice versa. Furthermore, due to varying host factors, HIV haplotype fitness can differ between hosts and conclusions for one patient might not be applicable to the next. Should assumptions of our model prove to be reasonable, this model could be an improvement in the toolbox of personalized medicine as it predicts HIV fitness from individuals' in-vivo data and not from in-vitro assays as is still mostly the case today.

Relating genomic variation to drug response in childhood acute lymphoblastic leukemia

Agata Wesolowska-Andersen

Technical University of Denmark

Objectives: Acute lymphoblastic leukemia (ALL) is the most common childhood cancer with high inter-individual variability in treatment resistance and toxic side effects. The aim of this study is to define pharmacogenomic single nucleotide polymorphism (SNP) profiles predictive of drug response and susceptibility to severe side-effects, as well as understanding of the underlying molecular mechanisms. Investigating the complete landscape of childhood ALL is achieved by complementing the current knowledge of disease mechanisms with systems biology driven SNP panel selection.

Methods: Ca. 900 childhood ALL patients comprising two Danish and one German cohorts were genotyped for a carefully selected panel of ca. 25,000 functional SNPs by means of multiplexed targeted sequencing using Agilent SureSelect target enrichment together with next-generation sequencing. The selection of genes and polymorphisms involved previously known determinants of treatment response, as well as genes from domains of potential importance for ALL treatment. The contribution of inherited genetic variation to treatment response was investigated by associating germline SNPs with risk of high minimal residual disease levels after remission induction chemotherapy, clearance of administered chemotherapeutic drugs, susceptibility to various side-effects and risk of relapse. Besides investigating contribution of individual variations to phenotype, the effects of multiple SNPs acting in the same pathway or protein-protein complex were investigated using neural network models.

Results: Predictive SNP profiles for a range of drug response phenotypes were developed and validated in an external cohort treated on a similar protocol. Since the SNPs assayed in this study were selected for their functionality, the molecular mechanisms underlying each phenotype are easily explained. The complementary analyses of collections of SNPs grouped by metabolic pathways and protein-protein complexes provide a comprehensive review of the molecular processes influencing the drug response.

An integrated genomic analysis: Effect of grouping explanatory variables

Nimisha Chaturvedi

Epidemiology and Biostatistics, Vrije Universiteit Medisch Centrum

Consider a study where each sample has genome-wide gene expression and genomic copy number profiles. A number of statistical models have been proposed for studying the association between gene expression and copy number data while performing integrated analysis (Louhimo et al, NatMeth, 2012). Sometimes there is also interest in comparing association patterns found for subgroups of samples defined by explanatory factors. Such results could be helpful in locating differences in activated pathways. However, no method has been proposed so far to find such differences, to the best of our knowledge.

We propose a two-step approach to find differences in association between copy number and gene expression, when comparing subgroups of samples. Firstly, we use ridge regression to correct for the baseline associations between copy number and gene expression. Secondly, the global test (Goeman et al, JRSS B, 2006) is applied to the corrected data in order to find differences in association patterns between groups of samples. The ridge penalization can be estimated efficiently, in particular if the same gene sets are used for multiple outcomes.

In a simulation study it is shown that the method picks up differences even in small regions. We also apply our method to two publicly available breast cancer datasets, for which we identify chromosome arms where copy number led gene expression regulation differs between estrogen receptor positive and negative samples. In spite of differing genomic coverage, some selected arms are identified in both the datasets.

Our approach can be used with most types of microarray/sequencing dataset, including methylation and microRNA expression. The method is implemented in R.

Visualization of Gene re-arrangements in RNA-Seq

Jan Oosting

LUMC, Department of Pathology

DeFuse is a tool to identify gene fusion events in RNA-Seq data. We present an R-package that can show the genetic re-arrangements directly from the output of DeFuse without any manual intervention.

Intra-individual reconstruction of tumor evolution in multiple myeloma

Julian Gehring

EMBL Heidelberg

Multiple myeloma is a plasma cell malignancy without any available curative therapy. Many patients develop asymptomatic precursor states, of which a fraction of about 1% per year experience the transition to the final symptomatic cancer. Both key drivers and reliable clinical predictors for this transitions to multiple myeloma are largely unknown and focus of current research.

For a cohort of three patients, we performed high coverage (50x) whole genome sequencing for three time points throughout tumor progression, including a rarely investigated asymptomatic stage. Based on the identification of single-nucleotide and large-scale structural variations in all stages, we detected the varying dominance of subclones over time and reconstructed the chain of genomic alterations throughout tumor progression on an intra-individual basis.

The work presented here offers new insights into the mechanisms of multiple myeloma formation, as well as a statistical graph theory based framework for the inference of tumor development from short read genome sequencing data sets, applicable also to single case studies.

Improved genotype calling for rare variants

Matthew Ritchie

The Walter and Eliza Hall Institute of Medical Research

Over the past decade, SNP genotyping microarrays have revolutionised the study of complex disease. The current range of commercially available genotyping products contain extensive catalogues of rare variants (those with frequency less than 5% in a population). Existing SNP calling methods have difficulty with rare variants, as the models rely on each genotype having a minimum number of observations to ensure accurate calls. We have developed a new unsupervised method for converting raw intensities into genotype calls that aims to overcome this issue. Our method allows a variable number of clusters (1, 2 or 3) for each SNP that is predicted using the available data, and offers improved performance over current approaches. We test our method against four competing genotyping algorithms on several Illumina data sets that include samples from the HapMap project where the true genotypes are known in advance.

Manifold dimension estimation

Kerstin Johnsson

Lund University

Omics data sets such as gene expression data are very high-dimensional, i.e. they have a large number of variables. However functional relations between the variables (e.g. co-regulated genes) means that the data has a structure so that the number of independent variables is actually much lower. When the functional relationships are non-linear a manifold with noise can be a good model for the data. I will give an overview over methods for estimating the appropriate dimension for this manifold model, and present a new, simple model for dimension estimation.

A formal definition for syntenic blocks

Cristina Gabriela Ghiurcuta

IC LAboratory of Computational Biology and Bioinformatics, EPFL

The latest advances in technology made it possible to produce tremendous amounts of genomic data, yet genome-scale analysis still presents formidable challenges: even simple pairwise comparisons are hard, since we lack good models for genome structure and evolution. Current approaches are based on the identification of so-called syntenic blocks (genome fragments that present highly similar collections of markers in most of the genomes under study). The identification of such blocks is crucial in comparative studies, yet its effect on final results has not been well studied, nor has any formal, biologically meaningful definition been proposed.

Syntenic blocks are in many ways analogous to genes – mostly, the markers used in constructing them are genes. Like genes, they can exist in multiple copies, thus we could define analogs of orthology and paralogy. However, whereas genes are typically studied at the sequence level, syntenic blocks are meant for genome-scale studies. It is their arrangement within the entire genome that is the main object of study. Thus the definition and construction of syntenic blocks involves both large- and small-scale evolutionary models.

We focus on an abstract framework and definition that applies to any type of markers and their collection of homology statements. Homology statements come from sequence similarity analysis; but they can also come from other types of analyses as well as from existing databases. Our framework focuses on providing a means of comparison for a wide range of implementations as well as for existing tools. Our definition for syntenic blocks is given in general terms and is refined by including characteristics specific to groups of organism. Thus, one of the key points of this framework is to include those genomic signatures that are relevant for such comparative studies, by integrating the existing heterogeneous, high-dimensional data types. Part of the challenges of this task is a better understanding of the data, and implicitly of the issues arising with its analysis, including statistical studies of biomarkers, genomic signatures, microarrays, genotypes.

Dynamics of HIV latency and reactivation at population and single-cell level

Pejman Mohammadi

D-BBSE, ETH Zurich

Successful antiretroviral therapy is able to reduce plasma viremia to undetectable levels in HIV+ patients; however, the virus is not eradicated. One of the current hypotheses to explain this phenomenon includes latently infected cells. These cells represent the reservoir from where the virus will reinitiate rounds of infection. HIV establishes latency in a minority of cells that are infected. The process leading to latency, and to reactivation are important steps for the characterization of pharmacological agents targeting this reservoir. We have put together a dynamic experiment in order to assess the process of latency, maintenance, and reactivation from the latent state from a transcriptional and epigenetic point of view. We further study the heterogeneity of the process at the cellular population level, and examine the key features at the single-cell level.

A Framework for Data Integration from Heterogeneous Sources in Genome Sciences

Christian X. Weichenberger

European Academy of Bozen/Bolzano

Analysis of data resulting from genome wide association (GWA) studies and next generation sequencing (NGS) experiments frequently involves a series of distinct processing steps, requiring a pipeline that annotates the generated data based on reference databases and that combines these different types of experimental results within a knowledge generation process.

Given the dynamic nature of the reference databases and the large volume of data generated by these genome-wide experiments there is an urgent need to organize storage, access, and manipulation of any dataset involved in these pipelines, including detailed information on the software used during its invocation. Furthermore, pipelines are characterized by their flexibility and are therefore often modified to address different scientific questions based on the same basic data. Taken together, these issues call for a framework that links usage and versioning of data sets and programs utilized in pipelines, guaranteeing reproducibility of previous results, but also allowing to rerun analyses on the newest databases.

We have thus developed the Data Integrator Framework (DIF), which fulfills these requirements and additionally implements a suite of tools, which are used regularly within our research group for GWA, NGS, and other types of genome wide experimental data. The DIF suite comprises modules for converting gene and protein IDs, finding orthologs in model organisms, listing genes in regions on the human genome, lifting of outdated genomic coordinates to the current build, assigning consequence types and predicting functional impact of single base pair substitutions, and for calculating linkage disequilibrium for a pair of dbSNP entry / human gene. DIF has been added to an in-house instance of the Galaxy web-based platform for graphical manipulation of workflows and is highly extensible. Currently, one of our main priorities is the development of a statistical module for integration of heterogeneous datasets for exploratory analysis and gene prioritization.

Estimation of Individual Gene Knockdown Effects from Mixed Effects Phenotypes in siRNA Screens

Fabian Schmich

Computational Biology Group, D-BSSE, ETH Zurich

Small interfering RNA (siRNA) screening is a technology that is prone to a high rate of false-positives, predominantly due to siRNA off-target effects, which have been shown to be mediated through partial complementary of siRNAs to the 3' UTRs of gene transcripts. This can lead to misinterpretation of the measured phenotypes with the consequence of low agreement between separate screens and decreased success rates in follow up validation experiments. We assume that the measured phenotype for a particular siRNA is in fact a combination of knockdown effects from on- and off-targeted gene transcripts. The model we developed de-convolutes this mixed effects signal to yield single gene effects, integrating prior knowledge of siRNA off-targets. Moreover, a statistically sound selection of effect genes, e.g. from a genome-wide screen, is performed during model fitting. We applied our approach to data from genome-wide RNAi screens to identify factors for Bartonella and Salmonella host cell invasion. Estimated single gene effects from multiple separate screens, where siRNAs with different nucleotide composition were used to knockdown each gene, showed a significant increase of correlation compared to uncorrected phenotypes. For each pathogen, lists of expected genes, provided by experts of the field, were succesfully selected and prioritised together with previously unknown factors. The proposed model can be applied to any phenotypic readout from siRNA screens in order to yield interpretable and reproducible gene prioritisations. We therefore expect that our approach will serve as an important component in the process of robust hit selection within the siRNA community.

Approaches for high throughput compound testing at single cell resolution

Wolfgang Raffelsberger

IGBMC (Illkirch, France)

Transfected cell arrays (TCA) provide a platform for high throughput compound testing, like RNAi screens. Recent developments of scanning microscopes make evaluation of multiple cellular markers at single cell resolution possible. To better interpret such data, we have made use of several different statistical approaches to establish a method for automatic evaluation of single cell resolution data. To obtain a universal and robust procedure, we first performautomatic detection of aberrant replicates and then compare each compound topositive and negative controlsin a plate-wise manner to obtain test-statistics.

Nested effects models for probabilistic combinatorial perturbation experiments

Juliane Siebourg-Polster

ETH Zürich, D-BSSE

Nested effects models (NEMs) are probabilistic graphical models for learning dependency networks from intervention data. The network nodes are perturbed experimentally and their connectivity is inferred from the nested structure of a number of observed effects. These models have been applied successfully to gene expression experiments. Since knock-down experiments are rather noisy, we generalize the framework to handle probabilistic perturbations. In particular we address the problem of a noisy input signal. We furthermore extent the model for combinatorial interventions and apply it to image based gene silencing experiments. These are part of 'InfectX' an RTD project from the Swiss SystemsX initiative and consist genome wide RNAi screens of human cells under infection of different pathogens. Aims of the project are firstly to get a better understanding of the human 'infectome' and identify novel drug targets among the involved human proteins. Secondly, we want to compare the learned networks for different diseases to search for shared shared pathway components.

Statistical inference in single-cell RNA-Seq experiments

Simon ANDERS

Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

Advances in wet-lab protocols now allow to sequence the mRNA content of single cells. Valuable biological insights can be gained not only from comparing different cell types but also from studying the variability within a population of seemingly homogeneous cells. For example, identification of clusters of covarying genes may help to infer regulatory networks in a more direct way than is possible with bulk (i.e., many-cell) RNA-Seq. For such applications, it is crucial to ascertain by apropriate inferential procedures to which extend an observed difference in read counts between cells is really due to differences in these cells' expression of the gene rather than due to technical noise.

Our studies of several current single-cell RNA-Seq protocols showed that technical noise always shows a particular mean-variance relationship, with very strong noise for low and medium read counts and hardly more noise than in conventinal many-cell experiments for high counts. We present a model that explains this mechanistically as a consequence of sampling on the level of individual mRNA molecules. We devised a method that uses spike-in data to fit the mean-variance relationship of technical noise for each sample. This is then used in a hypothesis test to assess whether the observed variance exceeds the predicted technical noise sufficiently to call the gene highly variable.

Our work not only provides a more reliable means to perform inference in single-cell RNA-Seq but also sheds rather critical light on recent claims about the sensitivity of state-of-the-art single-cell transcriptomics assays.

Anna Drewek

ETH Zurich

We are working on data from InfectX (www.infectx.ch). The goal of this consortium is to experimentally identify the pathogen entry mechanism for a set of bacterial and viral pathogens. In experiments HeLa cells were infected with seven different pathogens. In each screen one gene was knocked down using a siRNA targeting it. To obtain biological replicates per gene, these screens were performed with several siRNAs. We are building a model which brings the data of all pathogens together and incorporates a reliable hit ranking of genes. For this we apply linear mixed modeling with infection score as response and gene as random effect.

Identification of relationships between gene expression in mature adipocytes and whole fat biopsies in patients undergoing weight loss

Lennart Opitz

Functional Genomics Center Zurich, University Zurich/ETH Zurich

Obesity is a paramount problem for people in industrial countries in which the caloric intake exceeds the dietary needs, because of associated risks of secondary disorders including diabetes, atherosclerosis and hypertension. It is estimated that more than 50% of the population in developed countries can be characterized as obese according to the standards of the WHO.

Detailed understanding of the molecular mechanisms related to modification of the adipose tissue metabolic phenotype and endo/paracrine function at the levels of gene expression, posttranslational modification and the lipid droplet is essential for development of an effective treatment for the obesity related health complications. I will present methods and results of the analysis of changes in adipocyte and whole fat gene expression patterns and its correlation with multiple clinical parameters obtained as part of the weight loss study Optifast52 at the University of Heidelberg.

Subclass and biomarker discovery in peripheral T cell lymphomas

Maria Pamela Dobay

Bioinformatics core facility, Swiss Institute of Bioinformatics

---

Gene-subnetwork priorization for exome-sequencing of Mendelian diseases using a phenotype-interaction network approach

Daniela Beisser

Department of Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen

Whole-exome sequencing is a promising technique used in genetic diagnosis to identify sequence variations and their functional consequences for Mendelian diseases. However, a major difficulty of the approach is the differentiation between pathogenic and non-pathogenic mutations out of tens of thousands of variants per exome. Despite several filtering steps that can be applied to reduce the number of variants, e.g. removal of variants that fall into non-coding regions, mutations that yield silent substitutions and filtering of known variants, still hundreds of candidate variants remain. In the absence of affected family members or in the analysis of rare diseases, gene priorization has been applied recently to rank candidate genes.

We propose a novel network approach for gene priorization combining gene-similarities based on phenotype information with an interaction network. This allows (i) to score candidate genes according to the phenotypes of interest and (ii) to find combinations of candidate genes in high-scoring subnetworks. The methodology is tested on exome sequencing data of patients with well-defined syndromes, such as Nager syndrome and Coffin-Siris syndrome, in which genetic heterogeneity is suspected. The objective is to measure the performance of the technique with regard to the identification of known gene mutations as well as to identify novel disease-causing variants.

Joint Bayesian Inference of miRNA and TF Activities from miRNA and mRNA Expression Data

Holger Fröhlich

University of Bonn

Motivation: There have been many successful experimental and bioinformatics efforts to elucidate transcription factor (TF)- target networks in several organisms. For many organisms, these annotations are complemented by miRNA-target networks of good quality. Attempts that use these networks in combination with gene expression data to draw conclusions on TF or miRNA activity are, however, still relatively sparse. Results: We propose Bayesian inference of regulation of transcriptional activity (BIRTA) as a novel approach to infer both, TF and miRNA activities, from combined miRNA and mRNA expression data in a condition specific way. That means our model explains mRNA and miRNA expression for a specific experimental condition by the activities of certain miRNAs and TFs, hence allowing for differentiating between switches from active to inactive (negative switch) and inactive to active (positive switch) forms. Extensive simulations of our model reveal its good prediction performance in comparison to other approaches. Furthermore, the utility of BIRTA is demonstrated at the example of Escherichia coli data comparing aerobic and anaerobic growth conditions, and by human expression data from pancreas and ovarian cancer.

Inferring Causal Structure from Observational Data Using Structural Equation Models

Jonas Peters

ETH Zurich

Causal inference tackles the following problem: given iid observational data from a joint distribution, one tries to infer the underlying causal graph. This graph contains a directed arrow from each variable to its direct effects and is assumed to be acyclic.

Independence-based methods like the PC algorithm assume the Markov condition and faithfulness. These two conditions relate conditional independences and the graph structure; this allows us to infer properties of the graph by testing for conditional independences in the joint distribution. Those methods, however, can discover causal structures only up to Markov equivalence classes. Furthermore, conditional independence testing is

very difficult in practice.

In structural equation models (SEMs) each variable is assumed to be a deterministic function of its direct causes and some noise variable (e.g. Z=f(X,Y,N)), and the noise variables are assumed to be jointly independent. SEMs display their full strength when making constraints between causes and noise variables. For certain restrictions we obtain full identifiability, i.e. given an observational distribution P, we can recover the underlying causal graph. In additive noise models the structural equations are of the form Z=f(X,Y)+N. The subclass of linear functions and additive Gaussian noise does not lead to identifiability. This, however, constitutes an exceptional setting. If one assumes either (i) non-Gaussian noise, (ii) non-linear functions in the SEM or (iii) all noise variables to have the same variance, one can show that additive noise models are identifiable.

We develop a causal inference method that does not require faithfulness and can identify causal relationships even within an equivalence class. One can further reason that this procedure allows for a model check. Therefore, the method may remain undecided. We present results on both synthetic and real data sets.

Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription

Andreas Gschwind

University of Lausanne

Recent advances in genome-wide profiling of transcription factor (TF) binding and histone modifications have identified specific chromatin signatures related to various classes of functional elements in different cell types. However, their genetic basis and degree of variability across individuals remain largely unknown. We generated genome-wide enrichment profiles of TF binding, chromatin marks, and different measures of transcription in lymphoblastoid cell lines from two trios and 8 unrelated individuals sequenced in high depth as part of the 1000 Genomes project. Inter-individual variability of these phenotypes was quantified to understand both DNA sequence dependent and independent variation on transcription, TF binding, chromatin state, and their interplay in an allele-specific framework. Different organizational layers of the genome show abundant allelic effects and strong allelic coordination between layers, with the genetic control of this coordination acting primarily through transcription factor binding. Our findings support the notion of transcription factors being the primary determinants of gene expression programs, with the overall chromatin state reflecting, but not necessarily driving gene expression activity. We extended our analysis to 54 unrelated individuals for most assays to identify genetic effects affecting chromatin properties on a population level, and continue exploring the combinatorial patterns of all assays to further dissect the components of the general transcriptional state of the cells. This study will improve our understanding of the biological landscape around regulatory and other functional elements of the genome, and provide better means to interpret the heritability and molecular basis of phenotypic diversity, such as disease susceptibility, in humans.

3D structures of antibiotic resistant proteins from maximum entropy analysis of genomic information

Sikander Hayat [1], *Chris Sander* [2], *Debora S. Mark s*[3], *Arne Elofsson* [1]

[1] *Stockholm University, Sweden,* [2] *Sloan-Kettering Cancer Center, New York, USA* [3] *Harvard Medical School, Boston, USA*

Recent work showed that one can predict 3D protein structure from sequences alone by using a maximum entropy formalism to find causal evolutionary correlations in the sequences and avoid signals arising from transitivity (1-4). A crucial challenge/bottleneck in bacterial antibiotic resistance biology is the elucidation of the 3D structure and functional residues of beta barrel membrane proteins, which is extremely hard with current experimental techniques (9). Significant contributions have been made recently in predicting beta-barrel protein topologies, where machine learning methods and analytical methods based on geometric construction of beta-barrels have been developed to generate idealized beta-barrel models (10-11). However, high-resolution 3D modeling of beta-barrels and identification of functionally relevant residues is still an open problem.

I will present a maximum entropy approach to identify causal evolutionary correlations between residues allowing us to compute the 3D structure and identify crucial residues of these important proteins. This work and others shows the potential of using global probability models to reveal evolutionary constraints when enough data is available and suggests the approach may be useful for any area of biology where there is large amounts of data and many non-causal correlations (1-7). If successful, this will enhance our understanding of assembly of translocation machineries in the outer-membrane of bacteria, chloroplast and mitochondria. Further, with high-resolution 3D models and functionally relevant residues identified as described above, we will be able to better understand the translocation mechanism of beta-barrel proteins involved in the secretion of virulence factors in Chlamydia (12).

Inferring microbial interaction networks and susceptibilities to external perturbations from longitudinal metagenomic data

Richard R. Stein, Vanni Bucci, Nora C. Toussaint, Gunnar Rtsch, Joao B. Xavier, Chris Sander

Memorial Sloan-Kettering Cancer Center New York City

Recent advances in metagenomics revealed the fundamental role of the intestinal microbiota in human health and disease. In healthy individuals, the intestinal microbiota resides in relatively stable but seemingly fluctuating conditions; shifts in this composition are indicators of external perturbations, such as changes in diet or antibiotic administration, and imbalance in composition has been linked to the progression of obesity, Crohn's disease and type 2 diabetes.

In this work, we present an ecological data-based method, in which the measured temporal variations of species' densities are gradually adjusting an underlying microbial community model. As a result, we obtain global parameters describing the growth, the interactions and the susceptibilities to external stimuli of the species in the studied community.

These parameters characterize not only the ecological system but provide us with experimentally testable hypotheses. Moreover, they can be used to predict the system's temporal dynamics. Applied to the bacteria in the gut, such predictions may allow us to improve the rational design of therapies based on the better understanding of the antibiotic impact on the microbial community. Furthermore, the system's steady states may help explaining experimentally observed catastrophic shifts induced by external perturbations.

Joint estimation of modular networks of multiple genomic data types

José Sánchez

Chalmers University of Technology

In the study of transcriptional, genetic or epigenetic data for different groups (cancer types, for example) it's reasonable to assume that some links in the networks are common across groups, and that this property is preserved locally, thus defining a modular structure. For ease of interpretation, sparsity in the resulting model is also desirable. We assume the genomic data to have a multivariate normal distribution and propose to estimate the networks by optimization of a fused adaptive lasso penalized likelihood function for the inverse covariance matrices. To achieve modular topology we propose a novel adaptive penalty, which is computed from an initial zero-consistent solution. We also propose a generalization of the fused lasso penalized likelihood to higher-order dimensions, where equality is not only encouraged acrossgroups but also with respect to an ordered variable like survival. For integration of data types we define a prior distribution for the plausible links. We optimize the penalized log-likelihood using ADMM (Alternate ....). By simulation we show that our method performs better than competitors and is faster. We apply the method to the study of regulators or disease driving genes in glioblastoma, breast and ovarian cancer and integrate mRNA, miRNA, CNA, methylation, lost of heterozygosity and clinical data (survival).

Participants

Ali Alfaiz University of Lausanne
Simon Anders European Molecular Biology Laboratory
Keith Baggerly UT MD Anderson Cancer Center
Alfred Balch University of Utah
Niko Beerenwinkel ETH Zurich
Jonas Behr ETH Zurich
Daniela Beisser University of Duisburg-Essen
Hans Bitter Novartis Institutes for Biomedical Research
Andreas Buness Novartis
Laura Monika Buzdugan ETH Zurich
Peter Bühlmann ETH Zurich
Nimisha Chaturvedi Vrije Universiteit Medisch Centrum
Simona Constantinescu ETH Zurich
Victor DeGruttola Harvard SPH
Mauro Delorenzi CHUV-UniL-SIB
Maria Pamela Dobay University of Lausanne
Eytan Domany Weizmann Institute of Science
Anna Drewek ETH Zurich
Jan Ernest ETH Zurich
Jacques Fellay EPF Lausanne
Holger Fröhlich University of Bonn
Julien Gagneur LMU Munich
Swann Gaulis Novartis Pharma AG
Julian Gehring EMBL Heidelberg
Sarah Gerster SIB Swiss Institute of Bioinformatics
Moritz Gerstung Wellcome Trust Sanger Institute
Cristina Gabriela Ghiurcuta EPF Lausanne
Darlene Goldstein EPFL
Andreas Gschwind University of Lausanne

Sikander Hayat Stockholm University
Ariane Hofmann ETH Zurich
Thomas Hopf TU München
Francesco Iorio EMBL-European Bioinformatics Institute and Wellcome Trust Sanger Institute
Cecile Janssens Emory University
Xiaoyu Jiang Novartis Institutes for BioMedical Research
Kerstin Johnsson Lund university
Sunduz Keles University of Wisconsin
Edward Khokhlovich Novartis
Felix Klein EMBL Heidelberg
Jean-Pierre Kocher Mayo Clinic
Alix Leboucq EPFL
Gwenael Leday VU Amsterdam
Greg Lefebvre Nestle Institute of Health Science
Steve Lewitzky Novartis
Marloes Maathuis ETH Zurich
Douglas Mahoney Mayo Clinic
Debora Marks Harvard Medical School
Tobias Marschall CWI Amsterdam
Paul McLaren EPF Lausanne
Renee Menezes VU University
Eugenia Migliavacca University of Lausanne
Pejman Mohammadi ETH Zurich
Susan Murphy University of Michigan
Jan Oosting LUMC Leiden
Lennart Opitz ETH Zurich
Klea Panayidou University of Frederick
Jonas Peters ETH Zurich
Wolfgang Raffelsberger Universit de Strasbourg
Bernhard Renard Robert Koch-Institut
Matthew Ritchie The Walter and Eliza Hall Institute of Medical Research
Volker Roth University of Basel
Thomas Sakoparnig ETH Zurich
Chris Sander Memorial Sloan-Kettering Cancer Center
Fabian Schmich ETH Zurich
Martin Schumacher Novartis Pharma AG
Alexander Schönhuth Centrum Wiskunde and Informatica

Frdric Schütz SIB Swiss Institute of Bioinformatics
David Seifert ETH Zurich
Juliane Siebourg-Polster ETH Zurich
Charlotte Soneson SIB Swiss Institute of Bioinformatics
Rainer Spang University of Regensburg
Terry Speed Walter and Eliza Hall Institute
Richard Stein Memorial Sloan-Kettering Cancer Center
Daniel Stekhoven University of Zurich
Ewa Szczurek ETH Zurich
José Sánchez Chalmers University of Technology
Simon Tavar Cambridge
Armin Töpfer ETH Zurich
Levi Waldron Harvard School of Public Health
Christian X. Weichenberger EURAC Bozen/Bolzano
Agata Wesolowska-Andersen DTU
Pratyaksha Wirapati SIB Swiss Institute of Bioinformatics
Jean Yee Hwa Yang University of Sydney
Judith Zaugg Stanford University
Giovanni d'Ario SIB Swiss Institute of Bioinformatics
Alena van Bömmel Max-Planck Institute for Molecular Genetics