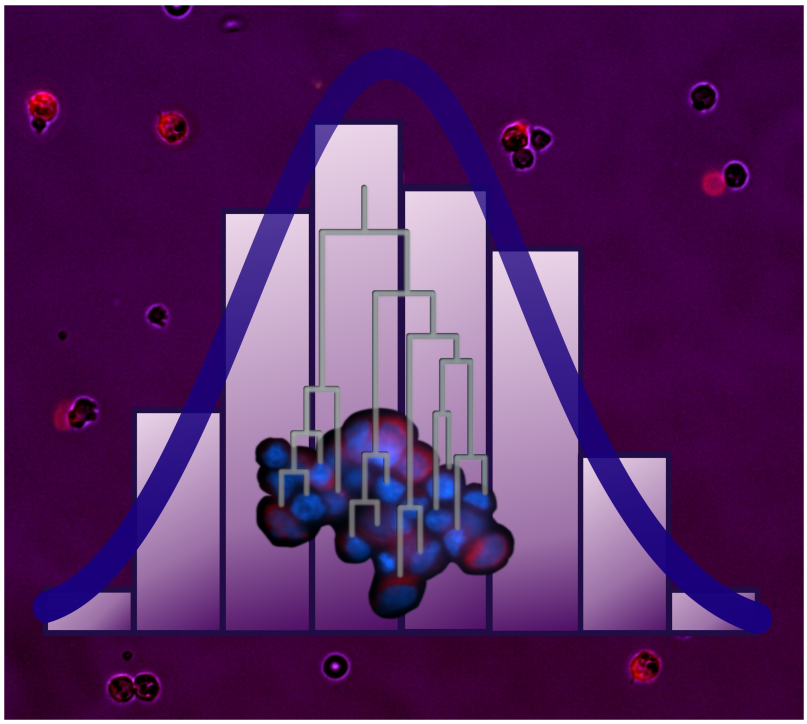
Congressi
Stefano Franscini
Swiss Federal Institute
of Technology Zurich

Statistical Challenges in Single-Cell Biology

Ascona, April 30 - May 5 2017



Organised by Niko Beerenwinkel, Peter Bühlman, Wolfgang Huber

Driven by biotechnological developments that enable single-cell handling and measurements on a genome-wide scale, single-cell profiles are now generated widely and rapidly. Single-cell approaches allow for probing biological systems at an unprecedented level of detail. For the first time, we can manipulate single cells and investigate variation among individual cells and inter-cellular interactions on the molecular level.
This interdisciplinary workshop is intended to be a forum for the dissemination of cutting-edge biotechnological and computational developments and the identification of open data analysis problems and solutions. Targeted areas for the workshop include: novel experimental techniques for single-cell analysis, statistical models of cell-to-cell variation, data integration, and applications of single-cell genomics to somatic variation in development and disease.

Sponsors


Congressi
Stefano Franscini
Swiss Federal Institute
of Technology Zurich
FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation
Kontaktgruppe für Forschungsfragen (KGF)
BASF
syngenta
NOVARTIS
Roche
SIB
Swiss Institute of
Bioinformatics
SOUND
PLOS
COMPUTATIONAL
BIOLOGY

Venue

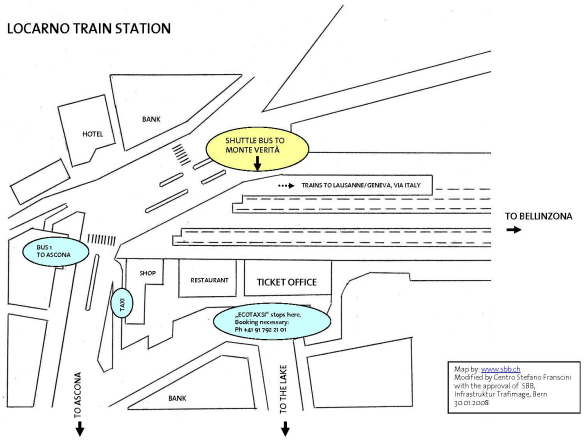
Monte Verità
Via Collina 84
CH-6612 Ascona
tel. +41 91 785 40 40

About the Congressi Stefano Franscini (CSF)

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

Shuttle service from Locarno Station

A free 13-seater shuttle bus to Monte Verità leaves from Locarno railway station Sunday April 30 at the following times: **14.05**; **14.45**; **15.35**; **16.15**; **17.05**; **17.45**; **18.30**. The meeting point is on the right side of the train platforms in Locarno (see image).



Keynote lectures

Identifying differentially variable genes from single-cell RNA-sequencing data and applications in immunity and ageing

John Marioni

Cancer Research UK, Cambridge, UK

Single-cell mRNA sequencing can uncover novel cell-to-cell heterogeneity in gene expression levels within seemingly homogeneous populations of cells. However, the data generated are subject to high levels of technical noise, creating new challenges for identifying genes that show genuine heterogeneous expression within the population of cells under study. In this presentation I will describe BASiCS (Bayesian Analysis of Single-Cell Sequencing data), an integrated Bayesian hierarchical approach that appropriately models noise and identifies both highly variable genes within a population of cells as well as genes that are differentially variable between populations of cells. I will demonstrate how BASiCS can be applied by discussing how aging impacts transcriptional dynamics using single-cell RNA-sequencing of unstimulated and stimulated naive and effector memory CD4+ T cells from young and old mice from two divergent species.

From one to millions of cells: computational approaches for single-cell analysis

Peter Kharchenko

Department of Biomedical Informatics, Harvard Medical School

Over the last five years, our ability to isolate and analyze detailed molecular features of individual cells has expanded greatly. In particular, the number of cells measured by single-cell RNA-seq (scRNA-seq) experiments has gone from dozens to over a million cells, thanks to improved protocols and fluidic handling. Analysis of such data can provide detailed information on the composition of heterogeneous biological samples, and variety of cellular processes that altogether comprise the cellular state. Such inferences, however, require careful statistical treatment, to take into account measurement noise as well as inherent biological stochasticity. I will discuss several approaches we have developed to address such problems, including error modeling techniques, statistical interrogation of heterogeneity using gene sets, and visualization of complex heterogeneity patterns, implemented in PAGODA package. I will discuss how these approaches have been modified to enable fast analysis of very large datasets in PAGODA2, and how the flow of typical scRNA-seq analysis can be adapted to take advantage of potentially extensive repositories of scRNA-seq measurements.

Synthetic gene circuits for in situ cell classification

Kobi Benenson

Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

Classification of biological samples is typically performed by computer algorithms that process sample-derived data in order to assign a sample to a phenotypic class. Recent developments in synthetic biology and biomolecular computing open up a possibility to classify individual live cells in situ using complex multilayered gene circuits. These circuits typically deploy multiple sensors to read out a panel of cytoplasmic biomarkers, and an information-processing module that integrates sensory data and generates an output

whose intensity correlates with the cell phenotype. One potential application of such circuits is specific cell targeting in genetic diseases and cancer.
In the talk I will discuss recent experimental progress in developing foundational technologies for classifier circuits, as well as a computational framework for their automated design. In particular, I will describe novel approaches towards sensing and integration of multiple transcriptional and microRNA inputs. I will also describe novel characterization methods that enable rapid evaluation of input-output relationship of complex circuits in mammalian cells.

Highly multiplexed analysis of the tumor ecosystem by mass cytometry

Bernd Bodenmiller

Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

Cancer is a tissue disease. Cancer cells and normal cells form an evolving ecosystem to support tumour development. The complexity of this system is a key obstacle to heal the disease. The visualization and study of the tumour ecosystem is thus essential to enable an understanding of tumour biology, to define biomarkers, and to identify new therapeutic routs. For this purpose, we developed imaging mass cytometry (IMC) which enables to visualize >50 epitopes simultaneously on tissues with subcellular resolution. Soon >130 markers can be visualized. We applied IMC for the analysis of 100s of breast cancer samples. Our analysis reveals a surprising level of inter- and intra-tumor heterogeneity and identify new diversity within known breast cancer subtypes. Furthermore, we identify cell-cell interaction motifs in the tumor microenvironment correlating with clinical outcomes of the analyzed patients. Our results show that IMC provides high-dimensional analysis of cell type, cell state and cell-to-cell interactions in the tumour ecosystem. We envision that IMC will enable a personalized approach to diagnose disease and to guide treatment.

Identifying novel correlates of protection via single-cell analysis of antigen-specific T-cells

Raphael Gottardo

Fred Hutchinson Cancer Research Center

Antigen (Ag)-specific T-cells are rare among circulating peripheral blood mononuclear cells. Despite their critical role in containing infection, their total frequency in blood does not correlate with clinical protection; therefore, it is likely that some characteristic or quality of a subset of Ag-specific cells confers protective immunity for infectious diseases. Recent technical advances such as polychromatic flow cytometry and single-cell genomics have enabled the high-throughput quantification of genes or proteins at the single-cell level. Although many analytical tools exist for analyzing high-dimensional data, such as from gene expression arrays, very few tools are available for single-cell data, which has its own bioinformatics and statistical challenges. During this talk I will give an overview of the challenges involved in the analysis of single-cell data and show how such technologies combined to novel computational methods can be used to improve the characterization of Ag-specific T-cells to reveal novel correlates of protection for HIV and malaria.

Mapping genetic interactions during intestinal organoid self-organization

Prisca Liberali

FMI, University of Basel, Basel, Switzerland

Collective behaviour is a complex behaviour in which understanding of the individual single cells does not necessarily explain the collective behaviour of the population of cells. The unexplainable behaviour is the self-organization of the system. Recently, model systems have been developed from stem cells that can self-organize into organoid structures in vitro. In particular, intestinal organoids recapitulate most of the spatial and temporal processes of morphogenesis and patterning observed in intestinal tissue in animals and contain all cell types found in the intestine. To identify genes and map

regulatory genetic interactions underlying self-organization during organoid development, a series of small compound screens have been performed in a 3D intestinal stem cell culture model. Extensive image analysis and advanced statistics is applied to quantify multiple features in a large number of single cells in each perturbed cell population. We are then building a statistical trajectory of organoid development in multivariate feature space. This analysis is generating a cross-comparable dataset that is particularly well suited for the inference of regulatory genetic interactions. Modules of genes that functionally interact in regulating some or multiple processes involved in self-organization have been identified. For several genes with strong effects and key roles in the regulatory genetic interaction network we performed live cell imaging of single intestinal stem cell to a fully grown organoid with a custom made light sheet microscope. This approach is establishing a novel paradigm in genetic interaction screening applied to collective cell behavior and symmetry breaking in stem cells.

Symmetry breaking and self-organisation in mouse development

Takashi Hiiragi

EMBL, Heidelberg, Germany

A defining feature of living systems is the capacity to break symmetry and generate well-defined forms and patterns through self-organisation. Our group aims to understand the principle of multi-cellular self-organisation using a well-suited model system: early mouse embryos. Mammalian eggs lack polarity and thus symmetry is broken during early embryogenesis. This symmetry breaking results in the formation of a blastocyst consisting of two major cell types, the inner cell mass and the trophectoderm, each distinct in its position and gene expression. Our recent studies unexpectedly revealed that morphogenesis and gene expression are highly dynamic and stochastically variable during this process. Determining which signal breaks the symmetry and how the blastocyst establishes a reproducible shape and pattern despite the preceding variability remains fundamental open questions in mammalian development. We have recently developed a unique set of experimental frameworks that integrate biology and physics. With this we aim to understand how molecular, cellular and physical signals are dynamically coupled across the scales for self-organisation during early mammalian development.

Droplet-based microfluidic for single cell analysis

Valérie Taly

Paris Descartes

Droplet-based microfluidic has led to the development of highly powerful systems that represent a new paradigm in High-Throughput Screening where individual assays are compartmentalized within microdroplet microreactors. By combining a decrease of assay volume and an increase of throughput, this technology goes beyond the capacities of conventional screening systems. Droplets (in the pL to nL range) are produced as independent microreactors that can be further actuated and analyzed at rates of the order of 1000 droplets per seconds. Added to the flexibility and versatibility of platform designs, such progress in sub-nanoliter droplet manipulation allows for a level of control that was hitherto impossible [1-3].
Microfluidics has recently emerged as a major player in the single cell era that is gradually emerging among biology laboratories, mainly due to the single-cell high-throughput handling solutions it offers. After a presentation of different microfluidic systems and strategies allowing for single cell manipulation and analysis, the presentation will focus on compartment-based microfluidic approaches. Illustrative examples of several technologies will be presented including applications in directed evolution, high throughput screening and omics. Finally recent works with high potential impact for cancer research will be presented [4,5].

References

1. Taly V, Pekin D, Abed AE, Laurent-Puig P. Detecting biomarkers with microdroplet technology. Trends Mol Med. 2012;18(7):405-16. doi:10.1016/j.molmed.2012.05.001.
2. Kelly BT, Baret JC, Taly V, Griffiths AD. Miniaturizing chemistry and biology in microdroplets. Chem Commun (Camb). 2007(18):1773-88. doi:10.1039/b616252e.
3. Taly V, Kelly BT, Griffiths AD. Droplets as Microreactors for High-Throughput Biology. Chembiochem. 2007;8(3):263-72.
4. Perkins G, Lu H, Garlan F, Taly V. Droplet-Based Digital PCR: Application in Cancer Research. Advances in Clinical Chemistry. 2017;85:43-91.
5. Microchip Diagnostics. Series Methods in Molecular Biology. Springer Protocols. 2017; 1547.

Single cell analysis of circulating tumor cell clusters

Nicola Aceto

Department of Biomedicine, University of Basel, Basel, Switzerland

Cancer patients that develop a metastatic disease are currently considered incurable. Mainly, this is due to a limited understanding of the molecular mechanisms that characterize the metastatic process, and the lack of effective metastasis-suppressing agents. The metastatic cascade begins when primary tumor cells enter the circulatory system, and it is followed by their extravasation at distant sites, where they form proliferative metastatic lesions. Cancer cells in the bloodstream are referred to as circulating tumor cells (CTCs), and while technically challenging, their isolation and interrogation holds the key to understanding the principles governing the metastatic spread of cancer. For instance, using a combination of microfluidic technologies, single cell sequencing, molecular and computational biology, we recently understood that CTC-clusters, rather than single migratory CTCs, are highly efficient precursors of metastasis in several cancer types. With the isolation and sequencing of single cells within CTC-clusters, we aim to dissect their cellular and molecular heterogeneity, to shed light on some unique features of these metastatic precursors, and to enable the development of new metastasis-suppressing therapies.

Single Cells, Big Data

Nicholas E. Navin

MD Anderson Cancer Center, Houston, TX, USA

Single cells generate big data sets. Sequencing the genome or exome of a single cell can generate terabytes of data that must be processed to mitigate technical errors that arise during whole-genome amplification. The error profiles of single cell DNA sequencing data are unique compared to standard next-generation sequencing datasets and therefore violate many assumptions that these methods make. Although there has been significant

progress in the last few years in developing computational methods for single cell RNA sequencing data, statistical methods for single cell DNA data are far behind. In this talk I will provide an overview of the experimental and computational methods our group has developed for performing single cell DNA sequencing to measure copy number profiles, point mutations and indels in individual tumor cells. I will also discuss applications of these methods to study metastatic lineages in colon cancer and punctuated evolution in triple-negative breast cancer patients.

Latent variable models for decomposing single-cell expression variation

Oliver Stegle

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, CB10 1SD Hinxton, Cambridge, UK

Technological advances permit assaying the transcriptome and the proteome of single-cells, both in cell suspensions and increasingly in their natural spatio-temporal contexts. Cell-to-cell differences in gene expression can be driven by both observed and unmeasured factors, including technical effects such as batch, or biological processes including the cell cycle, apoptosis or cell differentiation. In this talk, I will describe latent variable models for decomposing the sources of variation in single-cell studies, thereby inferring both biological and confounding sources of variation.
In the first part of this talk, I will describe scalable sparse factor analysis models for single-cell RNA-seq. A particular focus of these methods is the ability to integrate prior annotations on biological gene sets, thereby identifying biological drivers of expression heterogeneity. In the second part I will describe methods based on Gaussian processes to model and test for spatial gene expression heterogeneity.

Inferring early bifurcation events from single-cell RNA-Seq data

Magnus Rattray

University of Manchester

Single-cell RNA-Seq (scRNA-Seq) data can be used to uncover the dynamic processes where cells differentiate into different lineages, through the use of pseudo-time inference methods. We have developed a statistical tool for identifying gene-specific differentiation times given pseudo-time estimates for each cell. We model a branching process in time as a Gaussian process, building on a recent approach for identifying perturbations from two-sample gene expression time-series data (Yang et al. 2016). In the case of single-cell data the inference problem has to be modified to allow inference of branch identity for genes diverging prior to the global cellular branching point identified using the Monocle2 algorithm (Trapnell et al. 2016). We implement our method using the GPflow Gaussian process library which uses Tensorflow to allow automatic differentiation of the variational objective function and exploiting rapid processing by GPUs when available (Matthews et al. 2016). Our method allows for the inference of the branching time for each gene along with an associated Bayesian credible region. We use simulated data to compare our method to a spline-based approach implemented in the BEAM method (Qiu et al. 2017) within Monocle2 and a Bayesian mixture of Factor Analysers approach (Campbell and Yau, 2017). We apply our method for scRNA-Seq and drop-seq datasets to identify early differentiation genes.

Joint work with Alexis Boukouvalas and James Hensman

Contributed talks

Bayesian Inference for Single-cell ClUstering and ImpuTing (BISCUIT)

Ambrose Carr

Memorial Sloan Kettering Cancer Center

Single-cell RNA-seq gives access to gene expression measurements for thousands of cells, allowing discovery and characterization of cell types. However, the data is noise-prone due to experimental errors and cell type-specific biases. Current computational approaches for analyzing single-cell data involve a global normalization step which introduces incorrect biases and spurious noise and does not resolve missing data (dropouts). This can lead to misleading conclusions in downstream analyses. Moreover, a single normalization removes important cell type-specific information. We introduce a data-driven model, BISCUIT, that iteratively normalizes and clusters cells, thereby separating noise from interesting biological signals. BISCUIT is a Bayesian probabilistic model that learns cell-specific parameters to intelligently drive normalization. This approach displays superior performance to global normalization followed by clustering.

We apply BISCUIT to single cell data on tumor-infiltrating cells from breast cancer patients showcasing the strength of this method. BISCUIT identifies both expected and novel biological populations, while common normalization techniques failed to reveal structure, instead amplifying strong biases differentiating patients. BISCUIT enables novel characterization of multiple different subpopulations of T cells, multiple myeloid clusters, and distinct populations of regulatory T cells and dendritic cells which are not detected with other normalization methods. This indicates the appropriateness of BISCUIT for experimental data containing significant diversity in cells. BISCUIT can infer underlying co-expression patterns and cell type-specific expression from data, and therefore it has advantages beyond predicting clusters and imputing data. Specifically, BISCUIT revealed strong differences in co-expression patterns between subpopulations of T-cells. Interestingly, varying co-expression patterns between co-receptor genes in regulatory T cells showed significant variation across patients. These interesting observations could impact immunotherapy and might in the future suggest avenues for tailoring patient-specific immunotherapy. Also, BISCUIT parameters specific to each cell type can be used to explore differences in the tumor ecosystem across patients.

Missing Data and Technical Variability in Single-Cell RNA-Sequencing Experiments

Stephanie Hicks

Dana-Farber Cancer Institute / Harvard T.H. Chan School of Public Health

Recent advances in high-throughput technology permit genome-wide gene expression measurement at the single cell level. Single-cell RNA-Sequencing (scRNA-Seq) is the most widely and has been used in numerous publications. Although systematic errors, including batch effects, have been widely reported as a major challenge in high-throughput technologies, these issues have received minimal attention in published studies based on scRNA-Seq technology. Here, we examine data from published studies and found that systematic errors can explain a substantial percentage of observed cell-to-cell expression variability. Specifically, we show that scRNA-Seq reports more zeros than expected, and that technical variability can lead to cell-to-cell differences that can be confused with novel biological results. Finally, we demonstrate how batch-effects can exacerbate the problem.

Transcriptome-wide splicing quantification in single cells

Yuanhua Huang

School of Informatics, University of Edinburgh

Single cell RNA-seq (scRNA-seq) has revolutionised our understanding of transcriptome variability, with profound implications both fundamental and translational. While scRNA-seq provides a comprehensive measurement of stochasticity in transcription, the limitations of the technology have prevented its application to dissect variability in RNA processing events such as splicing. Here we present BRIE (Bayesian Regression for Isoform Estimation), a Bayesian hierarchical model which resolves these problems by learning an informative prior distribution from multiple single cells. BRIE combines the mixture modelling approach for isoform quantification with a regression approach to learn sequence features which are predictive of splicing events. We validate BRIE on several scRNA-seq data sets, showing that BRIE yields reproducible estimates of exon inclusion ratios in

single cells and provides an effective tool for differential isoform quantification between scRNA-seq data sets. BRIE therefore expands the scope of scRNA-seq experiments to probe the stochasticity of RNA-processing.

Notes: This is a joint work with Guido Sanguinetti. The full manuscript is available on BioRxiv: http://biorxiv.org/content/early/2017/01/05/098517

The two phases in gene expression regulation

Jean (Zhijin) Wu

Brown University

Transcription is a complex process enabled and regulated by many factors. We propose a latent class probabilistic model of sequencing counts that well explains the characteristics of observed single cell transcriptome, including the heterogeneity of the zero inflation, the sparsity and extreme skewness of expression in some genes. Using the analogy of energy levels, we consider the expression level of a gene in a given cell may be in the ground state or an excited state. Unlike existing methods, we do not expect the expression to be zero even if the gene is in ground state. We allow the transcript counts given ground state to be a zero-inflated Poisson random variable, with parameters depending on the cell. Using data from all genes in all cells in an experiment, we estimate cell specific parameters for the ground state, but gene specific parameters for their excited states.

At the specific gene level, we present a robust and fast algorithm to estimate the gene and cell specific states, and differential expression in two forms: the binary state change from ground to excited, and the continuous regulation in the excited state. We report gene expression regulation in both phases, most interestingly, expression compensation when the expression level in excited states are up-regulated to compensate for the reduced proportion of cells in the excited state.

At the whole cell level, we introduce a likelihood ratio based similarity measure that captures the overall concordance of gene expression in both phases between a pair of cells. We demonstrate the use of this model to identify the major regulation phases in different systems.

Single-cell analysis on microfluidic platforms

F. Kurth, P.S. Dittrich

Department of Biosystems Science and Engineering, ETH Zürich, Switzerland

Microfluidics is a promising technology and provides a huge toolbox for analytical and bioanalytical methods in general. With respect to cell analysis, the emergence of microfluidic platforms has paved the way to novel analytical strategies for positioning, treatment and observation of living cells, for creating of chemically defined liquid environments, and for tailoring mechanical or physical conditions. In particular, the development of microfluidic platforms opened new possibilities to manipulate and investigate individual cells and has tremendously increased our knowledge of single-cell behavior [1,2]. In this presentation, our recent approaches to single-cell analysis will be presented. We developed microfluidic platforms produced by so-called multilayer soft lithography to systematically study single cells. In these microdevices, single cells are physically trapped isolated from the environment by round-shaped valves that encapsulate the cell in a volume of a few hundred picoliters [3]. The realization of this tiny analytical chamber is the key requirement for a highly sensitive detection of the target molecules directly in the cell lysate, even when present in very low copy numbers. We integrated the platform with immunological methods such as enzyme-linked immunosorbent assays (ELISA) and competitive ELISA, which allows for the quantification of proteins [4] and other biomolecules [5]. In addition to the chemical analysis of the cell lysate, such devices were employed for fast evaluation of the effect of chemical compounds, e.g. drug candidates, on cells, with the benefit that heterogeneous responses can be easily revealed. The microfluidic platform can be employed for a wide range of molecules, for studies of various cell types including mammalian cells, yeast cells and bacteria as well as for liposomes as artificial cell models. Apart from the analysis of isolated single cells, we adopted microfluidic chambers for mechanobiological single cell studies within small cell populations. We applied fluid flow induced shear stress to induce calcium entry via reputedly mechanosensitive channel classes under varying culture conditions [6]. In combination with pharmacological channel modulation, our results could identify the role of particular channels, which has only been feasible by low throughput technologies before.

Literature

1. L. Armbrecht, P. S. Dittrich, Recent advances in single cell analysis, Anal. Chem.

2. Hümmer, F. Kurth, N. Naredi-Rainer, P. S. Dittrich, Single cells in confined volumes: microchambers and microdroplets, Lab Chip 2016, 16, 447.

3. K. Eyer, P. Kuhn, C. Hanke, P. S. Dittrich, A microchamber array for single cell isolation and analysis of intracellular biomolecules, Lab Chip 2012, 12, 765.

4. K. Eyer, S. Stratz, P. Kuhn, S. K. Küster, P. S. Dittrich, Implementing enzyme-linked immunosorbent assays on a microfluidic chip to quantify intracellular molecules in single cells, Anal. Chem. 2013, 85, 3280.

5. P. Kuhn, K. Eyer, S. Allner, D. Lombardi, P. S. Dittrich, A microfluidic vesicle screening platform: monitoring the lipid membrane permeability of tetracyclines, Anal. Chem. 2011, 83, 8877.

6. F. Kurth, A. Franco-Obregón, M. Casarosa, S. K. Küster, K. Wuertz-Kozak, P. S. Dittrich, Transient receptor potential vanilloid 2-mediated shear-stress responses in C2C12 myoblasts are regulated by serum and extracellular matrix, FASEB J. 2015, 29, 4726

CytoGLMM: Bayesian Hierarchical Linear Modeling for Flow Cytometry Data

Christof Seiler

Department of Statistics, Stanford University

Cytometry by time-of-flight (CyTOF) characterizes around 40 cell markers simultaneously. The numbers of measured cells per sample is usually around 10,000. Despite small donor sample sizes, the Nolan lab predicted surgical recovery with 32 donors and the Blish lab found associations with HIV acquisition with 33 donors. To relate CyTOF data to phenotypic outcomes, current methods cluster cells into subpopulations and then relate cluster features to outcomes (FlowMap-FR, Citrus, flowType-RchyOptimyx, FloReMi, and COMPASS). These approaches assume a discrete set of cell subpopulations. There is growing evidence suggesting that some cells adapt to counteract different threats such as viruses and cancers. To find such salient marker changes, BayesFlow extends cluster-based methods by allowing each cell to belong to a mixture of subpopulations. We propose to skip the clustering step and use a generalized linear model where the response

variable is the outcome and the explanatory variables are the 40 protein expression profiles. We build a hierarchical model of CyTOF data that allows us to estimate population level parameters and marginalize out the donor-specific parameters. Estimating donor-specific parameters is straightforward because the number of cells exceed the number of markers, whereas estimating population-level parameters is challenging because the number of markers usually exceed the number of donors. Therefore it is important to borrow information through partial pooling across donors when estimating the population-level parameters. We check the model through test quantities of the observed data and the posterior predictive distribution. An R implementation using the probabilistic programming language Stan is available in our new R package CytoGLMM. We validate our approach on open access CyTOF data available on FlowRepository and ImmPort.

Sensitive detection of rare disease-associated cell subsets via representation learning

Eirini Arvaniti

ETH Zurich

Rare cell populations play a pivotal role in the initiation and progression of diseases such as cancer. However, the identification of such subpopulations remains a difficult task. This work describes CellCnn, a representation learning approach to detect rare cell subsets associated with disease using high-dimensional single-cell measurements.

Existing approaches [1,2] address the task of detecting phenotype-associated cell populations via small variations of the following pipeline: cell populations are defined via a clustering algorithm, a cluster-based representation of each sample (e.g. in terms of cluster frequencies) is computed and, finally, a supervised learning module is used to associate with the phenotype of interest. Successful application of such approaches may be compromised by the quality of the clustering result, especially for rare hard-to-detect cell types.

To overcome this limitation, CellCnn does not separate the steps of extracting a cell population representation and associating it with disease status. Combining these two tasks requires an approach that (1) is capable of operating on the basis of a set of unordered single cell measurements, (2) specifically learns representations of single cell measurements that are associated with the considered phenotype and (3) takes advantage

of the possibly large number of such observations. We bring together concepts from multiple instance learning and convolutional neural networks to meet these requirements.

In this study, we apply CellCnn in a classification setting to reconstruct cell type-specific signaling responses in samples of peripheral blood mononuclear cells. We additionally apply CellCnn in a regression setting to identify abundant cell populations associated with disease onset after HIV infection, and achieve comparable prediction accuracy to a state of the art analysis performed recently [2], however with computational cost reduced by several orders of magnitude. Finally, we demonstrate the unique ability of CellCnn to identify extremely rare (down to 0.01% frequency) phenotype-associated cell subsets by detecting memory-like NK cells associated with prior CMV infection and leukemic blasts in minimal residual disease-like situations.

1. Aghaeepour, N. et al. Critical assessment of automated flow cytometry data analysis techniques. Nat. Methods 10, 228-238 (2013). 2. Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J. & Nolan, G. P. Automated identification of stratifying signatures in cellular subpopulations. Proc. Natl. Acad. Sci. U. S. A. 111, E2770-7 (2014).

Mapping genetic effects on interactions with single-cell states

Davis McCarthy

EMBL-EBI, Hinxton, UK

Technological advances have made it possible to sequence transcriptomes at single-cell resolution (scRNA-seq), at high-throughput, yielding insights into development and the etiology of different tissues and cell states. Simultaneously, it has become possible to assay multiple genomic layers in large cohorts of individuals. This now allows genotyping and scRNA-seq to be carried out for large numbers of genetically distinct individuals. Resources such as the Human Induced Pluripotent Stem Cells Initiative (HipSci; www.hipsci.org) provide cell lines from hundreds of human donors that can be used to investigate genetic effects on heterogeneity in single cells, in cell types inaccessible in large numbers of human individuals.

The particular characteristics of scRNA-seq data demand that approaches developed for bulk RNA-seq need to be adapted and extended to map genetic effects on expression heterogeneity in single cells. The large number of repeated measurements (cells from the same individual) provides challenges and opportunities for QTL mapping. We go beyond

mapping eQTLs to study variance QTLs (genetic variants that associate with variance phenotypes for a gene) and investigate interactions between expression, genetic variation and cell state, such as differentiated cell type or expression activity in particular pathways of interest.
We demonstrate the application of such methods to a dataset of more than 10,000 cells from over 60 individual human donors. In this study, HipSci iPSC lines are differentiated to definitive endoderm, with cells sampled at four time-points across the three-day differentiation experiment. We can map eQTLs and varQTLs in these data. Further more, using clock-time, pseudotemporal ordering of cells from expression data and independent measurements of cell-surface markers we can define multiple cell states and map genetic effects on interactions between these states and gene expression, yielding insights into genetic effects on interactions in single-cell states in early development.

Pooled CRISPR screening with single-cell transcriptome readout

Andre Rendeiro

CeMM Research Centre for Molecular Medicine of the Austrian Academy of Sciences

CRISPR-based genetic screens are accelerating biological discovery, but current methods have inherent limitations. Widely used pooled screens are restricted to simple readouts including cell proliferation and sortable marker proteins. Arrayed screens allow for comprehensive molecular readouts such as transcriptome profiling, but at much lower throughput. Here we combine pooled CRISPR screening with single-cell RNA sequencing into a broadly applicable workflow, directly linking guide RNA expression to transcriptome responses in thousands of individual cells. Our method for CRISPR droplet sequencing (CROP-seq) enables pooled CRISPR screens with single-cell transcriptome resolution, which will facilitate high-throughput functional dissection of complex regulatory mechanisms and heterogeneous cell populations.

Quantifying developmental plasticity by integrated single-cell RNA-seq and ex vivo culture

Lars Velten

EMBL

Single cell RNA-seq has emerged as a powerful tool to map cellular trajectories and branch points during development. However, due to the snapshot nature of RNA-Seq, no direct statements can be made about the reversibility of cell fate decisions and the ability of cells to transit between trajectories: The progeny of a cell that appears advanced towards a particular developmental end point might still functionally contribute to multiple lineages. To quantitatively assess cellular plasticity on scRNAseq-based developmental maps of human blood formation, we have combined high-dimensional surface marker indexing by FACS with both single cell RNA-Seq and single cell ex vivo cultivation. Our data reveal that while the position on the "map" is tightly correlated to the predominant cell type generated ex vivo, the probability of switching developmental fate gradually declines as development progresses. Our work suggests that blood formation, and possibly development in general, occurs in a Waddington landscape permissive of stochastic transitions between lineages downstream of bifurcation points, and it provides the tools required for quantitatively integrating single cell functional and transcriptomic data.

Single cell-based detection of diverse classes of genomic DNA rearrangements

Jan Korbel

European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany, and European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, UK.

Our laboratory has recently applied Strand-seq, a single cell based technology, to detect different classes of genomic structural variation (SVs), including copy-number unbalanced SVs and copy-neutral rearrangements. This unique technology sequences only DNA template strands present in single cells to preserve the structure and identity of individual homologues. Homologue resolution allows diverse classes of SV to be identified

including copy-neutral genomic rearrangements such as inversions and highly complex SVs that are extremely challenging, or even impossible, to identify with alternative sequencing approaches. While this places Strand-seq as a forefront method for detection of diverse classes of structural variation in genomes and in single cells, tools for integrative computational analysis of Strand-seq data in single cells are under active development in our group. We are currently pursuing extensive benchmarking and validation of these tools to verify that simple and complex DNA rearrangements can be predicted sensitively and with accuracy. Our presentation will, first and foremost, focus on novel genetic variant classes that we are able to discover in single cells using novel computational approaches operating on Strand-seq data. We have performed extensive benchmarking of our integrative Strand-seq computational analysis approaches using a set of "gold standard" lymphoblastoid cell lines enabling us to compare Strand-seq based rearrangement calls against various "3rd-generation" genomic techniques (i.e., Pacific Biosciences and Moleculo single molecule sequencing data, Bionano optical maps, and 10X Genomics sub-chromosome scale haplotypes). We have also applied this to study events formed by chromothripsis, a catastrophic DNA rearrangement process that can generate hundreds of SV breakpoints in a single cell division cycle. Modelling this event using our group's in vitro "complex alterations after selection and transformation (CAST)" approach we have identified extensive SV heterogeneity in chromothriptic cell lines, illustrating the new potential for studying complex genomic heterogeneity at the cell population level.

Leveraging single-cell technology for haplotyping

Tobias Marschall

Saarland University / Max Planck Institute for Informatics

Joint work with David Porubsky, Shilpa Garg, and Ashley Sanders

Humans and many other species are diploid. Every individual inherits two versions of each autosomal chromosome, called haplotypes, one from its mother and one from its father. Moving from (sequences of) genotypes to haplotypes is known as phasing or haplotyping. The knowledge of haplotypes is critical for addressing a variety of important questions in fundamental and clinical research [doi:10.1038/nrg2950].

While phasing is commonly approached by statistical inference on genotype data of large cohorts, such methods have the inherent limitation that they cannot reliably phase vari-

ants of low allele frequency, which can be of particular clinical interest.

Strand-Seq is a protocol that selectively sequences only template strands after one round of DNA replication in the presence of BrdU [doi:10.1038/nmeth.2206]. Single-cell sequencing of the daughter cells leads to sequencing reads whose directionality are informative of the haplotype of origin [doi:10.1101/gr.209841.116]. This technique is extremely powerful as it naturally provides chromosome-length haplotype information, i.e. it is able to phase over difficult regions such as centromeres, homozygosity regions or segmental duplications.

As Strand-Seq bypasses whole genome amplification, the sequencing yield from each single cell is low and many cells of the same individual need to be sequenced to obtain dense, whole-genome haplotypes. Here, we explore a hybrid strategy of combining Strand-Seq single-cell technology with long-read PacBio bulk sequencing.

We show how this data integration task can be cast as the weighted Minimum Error Correction (wMEC) problem. Solving this NP-hard problem leads to Maximum Likelihood haplotypes. We show how problem instances encountered in practice can be solved optimally using a fixed-parameter tractable algorithm [doi:10.1089/cmb.2014.0157].

This novel hybrid approach allows us to generate dense and extremely accurate chromosome-length haplotypes using as few as 10 Strand-seq cells combined with only 10-fold coverage PacBio data, paving the way to make haplotype-level genomics a reality.

Reconstructing tumour mutation histories from single-cell sequencing data

Katharina Jahn

CBG, BSSE, ETH Zurich

The mutational heterogeneity observed within tumours is a key obstacle to the development of efficient cancer therapies. A thorough understanding of subclonal tumour composition and the underlying mutational history is essential to open up the design of treatments tailored to individual patients. Recent advances in next-generation sequencing offer the possibility to analyse the evolutionary history of tumours at an unprecedented resolution, by sequencing single cells. This development poses a number of statistical challenges such as elevated noise rates due to allelic drop out, missing data and contamination with doublet samples.

We present SCITE our probabilistic approach for reconstructing tumour mutation histories from single-cell sequencing data with a focus on two recent extensions, the explicit modelling of doublet samples and a rigorous statistical test to identify the presence of parallel mutations and mutational loss.

Towards a statistical theory of ontogenetics

Geoffrey Schiebinger

Broad Institute

The past decade has witnessed rapid improvement of single cell DNA sequencing technology, with growth rates that rival the computer boom of the 1990's. These techniques are beginning to shed light on the variation that exists at the single cell level in the human body: while each cell in the body has the same basic information encoded in its DNA, different cell types use it differently by expressing different genes differently over time. For example, neurons express neutransmitters, B cells express antibody genes, and beta cells in the pancreas express insulin.

This project develops a mathematical theory of ontogeny – the process by which multiple differentiated cell types develop from a progenitor cell, such as a stem cell or fertilized egg. Just as phylogenetics reveals the relationship between species over evolutionary time, ontogenetics traces cell fates through the course of development. While phylogenetics is mature and is well studied within mathematics and statistics, there is no comprehensive statistical treatment of ontogenetics. As the data begin to accumulate, an urgent need arises for methods to map out developmental trajectories.

We address this need by developing mathematically rigorous methods to reconstruct ontogenetic trees from gene expression profiles. Our methods are based on mathematical tools from optimization and statistics, including optimal transport and mixing times of markov chains. We analyze our methods in the context of a time-varying nonparametric mixture model.

By carefully examining the trajectories upstream of branch points, we uncover some of the genetic mechanisms responsible for differentiation. A better understanding of how normal development proceeds could help us understand what goes wrong in disease states such as

cancer. Moreover, a deep understanding of the process by which stem cells differentiate to heal wounds or regenerate tissues could ultimately lead to new therapies and have a profound impact on human health.

Whole organism lineage tracing with genome editing

James Gagnon

Harvard University

Multicellular organisms develop from single cells by way of a lineage. However, current lineage tracing approaches scale poorly to whole, complex organisms. I have developed a lineage tracing method – GESTALT – that uses genome editing to progressively introduce and accumulate mutations in a DNA barcode. Lineage relationships can be reconstructed from the patterns of mutations shared between cells. In zebrafish, I generated thousands of lineage-informative alleles – "barcodes" – in single animals through CRISPR/Cas9 editing of a genetic target array. These barcodes are fixed in the genome of all daughter cells and describe relationships between embryonic progenitor cells and their fate in larval and adult tissues. Most cells in adult organs derive from relatively few embryonic progenitors. This phenomenon of clonal restriction occurs after embryogenesis, perhaps during mysterious mechanisms of tissue homeostasis. I will finally discuss my ongoing work connecting lineage and cell identity using single-cell RNAseq with the goals of understanding cell fate acquisition and stem cell dynamics.

Sparse Gamma/Poisson PCA to unravel the genomic diversity of single-cell expression data

Laurent Modolo

CNRS - Lyon

The developpment of high throughput single-cell technologies now allows the investigation of the genome-wide diversity of transcription. This diversity has shown two faces:

the expression dynamics (gene to gene variability) can be quantified more accurately, thanks to the measurement of lowly-expressed genes. Second, the cell-to-cell variability is high, with a low proportion of cells expressing the same gene at the same time/level. Those emerging patterns appear to be very challenging from the statistical point of view, especially to represent and to provide a summarized view of single-cell expression data. PCA is one of the most powerful framework to provide a suitable representation of high dimensional datasets, by searching for new axis catching the most variability in the data. Unfortunately, classical PCA is based on Euclidian distances and projections that work poorly in presence of over-dispersed counts that show zero-inflation. We propose a probabilistic PCA for single-cell expression data, that relies on a sparse Gamma-Poisson model. This hierarchical model is inferred using a variational EM algorithm, and we revisit the selection of the number of axis using an integrated likelihood criterion. We show how this probabilistic framework induces a geometry that is suitable for single-cell data, and produces a compression of the data that is very powerful for clustering purposes. Our method is competed to other standard representation methods like tSNE, and we illustrate its performance on a project that is based on transcriptomic data of CD8+ T cells. Understanding the mechanisms of an adaptive immune response is of great interest for the creation of new vaccines. We show that our method allows a better understanding of the transcriptomic diversity of T cells, which constitutes a new challenge to better characterize the short and long-term response to vaccination.

Posters

ProSolo - Discovering and Typing Genetic Variants in Single Cells

Alexander Schönhuth

Centrum Wiskunde & Informatica, Amsterdam

Calling and genotyping genetic variants from single cell sequencing experiments poses a variety of novel statistical challenges. The major source of biases is obvious: the initial amplification step, required for generating sufficient amounts of DNA for sequencing. Dealing with these biases in the discovery and genotyping of genetic variants is statistically involved and leads to computational issues that have not yet been comprehensively addressed. Exemplary questions requiring new answers are: (1) How to robustly genotype in the presence of differentially amplified alleles, (2) how to resolve ambigously mapped reads, (3) how to identify amplification errors and (4) how to integrate several single cells and bulk sequencing experiments into a comprehensive differential analysis of genetic variants.

Here, we present PROSOLO, a latent variable framework that provides sound answers to these questions. One key observation made use of by PROSOLO is that all read-based observations relating to a putative genetic variant are conditionally independent given the (unknown) allele frequency after amplification. This crucial insight exposes the inherent algorithmic problem to be of a runtime depending linearly – and not exponentially – on the number of the reads aligned to the region of interest. Overcoming this crucial computational bottleneck allows for a substantially refined analysis of single cell genetic variants and provides answers to the questions listed above. In addition, PROSOLO also quantifies the inherent uncertainties and allows to determine sequential motifs leading to amplification errors.

We applied PROSOLO to a large collection of single human blood cells and demonstrate the ability to call cell-specific single nucleotide variants for those cells at favorable performance rates. We further demonstrate how to identify amplification induced nucleotide errors and how to integrate them into variant calling pipelines. Finally, our analyses yield interesting insights into hematopoiesis.

**Joint with David Laehnemann, Alice McHardy, Helmholtz Center for Infection Research; Johannes Köster, Dana Farber, HarvardU.

Parameter estimation of nonlinear mixed effects models by two-stage approaches

Hans-Michael Kaltenbach

ETH Zurich

Until recently, modeling of cellular signaling largely relied on cell population data and corresponding ODE-based models described the average behavior of a population. Advances in live-cell imaging technology with fluorescent probes now allow simultaneous long-term recording of single-cell data for hundreds of cells at densely sampled time-points.
Nonlinear mixed effects models (NLMEs) provide a statistical approach to extend ODE-based models by maintaining a single deterministic model for all cells, with random effects (partially) explaining cell-to-cell heterogeneity of the observed responses by cell-specific values of parameters and initial conditions.
However, parameter estimation in NLMEs is notoriously difficult and existing approaches are tailored toward situations with few individuals and few observations per individual, typical for pharmacokinetic/pharmacodynamic (PK/PD) studies. For single-cell data, traditional approaches such as first-order conditional linearization are problematic due to the severe nonlinearity of the mean model (given as the solution of a system of ODEs), while current approaches such as stochastic approximation EM (SAEM) algorithms have heavy computational demands.
We show that for typical single-cell microscopy data, two-stage approaches that first independently estimate parameters for each individual cell and then combine these estimates and their uncertainties to yield the final NLME parameters become attractive alternatives to handle NLME models. Using a recent systems biology model of cell signaling, we demonstrate that the resulting estimates are comparable to SAEM in precision and accuracy, but computations are typically several times faster. Moreover, two-stage methods are conceptually very simple, easy to implement, can easily be adapted to incorporate, e.g., robust estimation procedures at the individual cell and the population parameter level, and their computationally more demanding first stage is easily parallelizable.

PCA for zero-inflated negative binomial data

Jean-Philippe Vert

ENS Paris

Single-cell gene expression data are characterized by large number of zero counts due to drop-out, and over-dispersion of non-zero counts. We propose a framework to fit a linear model, including or not latent factors, to such count data modeled by a zero-inflated negative binomial distribution. This allows to perform for example principal component analysis (PCA) while correcting for batch effects or sequencing depth on single-cell gene expression data. I will show how the model allows to better differentiate biological fluctuations from technical fluctuations, which are often confounded when a more naive approach is used, such as performing a standard PCA on the log counts. This is joint work with Davide Risso, Svetlana Gribkova, Fanny Perraudeau and Sandrine Dudoit.

Disentangling the different sources of variation in multi-omics single-cell sequencing

Ricard Argelaguet

European Bioinformatics Institute

Single-cell RNA sequencing is becoming a well-established routine that is revolutionising our understanding of cellular phenotypes. Interestingly, other data modalities are also starting to be assayed at the single-cell level, including epigenetics, proteomics and metabolomics, raising the question of how to jointly analyse these set of complex high-dimensional data sets using a statistically rigorous framework. Here we present single-cell Group Factor Analysis (scGFA), a generalisation of traditional factor analysis that integrates multi-omics data sets and is suited to the analysis of noisy single-cell sequencing data. scGFA calculates a low dimensional representation of the data which hopefully captures an inherent structure that might be masked by the noisy high-dimensional representation. Furthermore, it disentangles the different sources of variation and it calculates whether they are unique to a single omics or shared by multiple data sets, thereby revealing hidden sources of covariation. We applied scGFA to a recent data set of 61 embryonic

stem cells generated by a technology called scMT-seq, a recent method which uses single-cell genome-wide bisulfite sequencing and RNA sequencing to perform a parallel profiling of the DNA methylation and the gene expression in single cells. Our results show the existence of several axes of variation related to known biological processes and suggest the existence of three subpopulations that are associated with different pluripotency potential and genome-wide methylation rate.

Bead based compensation to correct for channel crosstalk in mass cytometry

Helena Crowell

University of Zurich

By addressing the limit of measurable fluorescent parameters due to instrumentation and spectral overlap, mass cytometry (CyTOF) combines heavy metal spectrometry to allow examination of up to 100 parameters at the single cell level. While spectral overlap is significantly less pronounced in CyTOF than flow cytometry, spillover due to detection sensitivity, isotopic impurities, and oxide formation can impede data interpretability. We designed CATALYST (Cytometry dATa anALYSis Tools) to provide tools for preprocessing and analysis of cytometry data, including compensation and in particular, an improved implementation of the single-cell deconvolution (SCD) algorithm for debarcoding and doublet-removal (Zunder et al. 2015, Nature Protocols 10, 316-333).
The CATALYST R package is available on Github and will be submitted to Bioconductor for review shortly. Currently, CATALYST provides a user-friendly R implementation of the SCD algorithm, and a function for estimating a compensation matrix from a priori identified single positive populations, which may be preceded by estimation of an optimal trim value that minimizes the sum of population- and channel-wise squared medians upon compensation. The matrix returned by this work flow may be directly applied to the measurement data or exported, e.g. to FlowJo or Cytobank.
We have demonstrated that spill estimates are to a great extent panel-specific, thereby eliminating the need for single-stain controls in each measurement. Moreover, removal of spillover artefacts will considerably effect downstream data interpretation, for example,

increase correlation between channels using the same antibody and decrease correlation between cross talking channels.

Additionally, we forsee the CATALYST R package as a collector for future developments in CyTOF data processing, such as statistical methods for differential discovery.

Studying the genotypic and phenotypic evolution of tumours using single cell RNAseq

Edith Ross

Cancer Research UK Cambridge Institute, University of Cambridge

Tumour evolution leads to genetic intra-tumour heterogeneity, which poses major challenges to cancer therapy. While tumour heterogeneity has been documented in several cases, many details of the underlying evolutionary processes are still unknown.

Recent advances in single-cell sequencing technologies have triggered the development of phylogenetic methods for single cell data that take into account the noise that is inherent to this type of data and promise to reveal tumour heterogeneity at a much higher resolution. This includes our own method oncoNEM (oncological Nested Effects Models), which is based on the nested structure of mutations observed between cells and jointly infers the tree structure, the number of clones and their composition.

So far these methods have only been applied to data derived from DNA sequencing. Here, we present the results of inferring tumour phylogenies from single cell RNAseq. Using RNA instead of DNA offers the opportunity to combine the phylogenetic insights with the gene expression of cells. Since the selective forces that shape evolution act on the phenotype not the genotype, this is an important step towards understanding tumour evolution. After inferring the phylogenetic trees with oncoNEM, we correlated the gene expression patterns found in the sequencing data with the structure of the tree to identify phenotypic similarities and differences between the clones. In this talk we will present the results of two case studies.

Evolutionary history of circulating tumor cells

Ewa Szczurek

Institute of Informatics, Faculty of Informatics, Mathematics and Mechanics, University of Warsaw

In this work, we focus on the phylogeny of circulating tumor cells (CTCs). From available single cell sequencing data of CTCs, we infer phylogenetic models for, and apply coalescent theory to speculate about the principles of the evolutionary process that generated their genomic sequences.

Beyond comparisons of means: understanding changes in gene expression at the single-cell level

Catalina Vallejos

Alan Turing Institute and UCL

Traditional differential expression tools are limited to detecting changes in overall expression, and fail to uncover the rich information provided by single-cell level data sets. We present a Bayesian hierarchical model that builds upon BASiCS to study changes that lie beyond comparisons of means, incorporating built-in normalization and quantifying technical artifacts by borrowing information from spike-in genes. Using a probabilistic approach, we highlight genes undergoing changes in cell-to-cell heterogeneity but whose overall expression remains unchanged. Control experiments validate our method's performance and a case study suggests that novel biological insights can be revealed. Our method is implemented in R and available at https://github.com/catavallejos/BASiCS.

Single-cell data reveals widespread recurrence of mutational hits in the life histories of tumors

Jack Kuipers

CBG, BSSE, ETH Zurich

Intra-tumor heterogeneity poses substantial challenges for cancer treatment. A tumor's composition can be deduced by reconstructing its mutational history. Central to current approaches is the infinite sites assumption that every genomic position can only mutate once over the lifetime of a tumor. The validity of this assumption has never been quantitatively assessed. We developed a rigorous statistical framework to test the infinite sites assumption with single-cell sequencing data. Our framework accounts for the high noise and contamination present in such data. We found strong evidence for the same genomic position being mutationally affected multiple times in individual tumors for 8 out of 9 single-cell sequencing datasets from a variety of human cancers. Six cases involved the loss of earlier mutations, five of which occurred at sites unaffected by large scale genomic deletions. Two cases exhibited parallel mutation, including the dataset with the strongest evidence of recurrence, indicating convergent evolution at the base pair level. Our results refute the general validity of the infinite sites assumption and indicate that more complex models are needed to adequately quantify intra-tumor heterogeneity for more effective cancer treatment.

Single-cell mutation calling via phylogenetic tree inference

Jochen Singer

CBG, BSSE, ETH Zurich

Understanding the evolution and dynamics of cancer is a crucial aspect towards the development of appropriate cancer therapies. This is a challenging task because cancers evolve as heterogeneous tumor populations with an unknown number of subclones of varying frequencies. Typically insights are gained through bulk sequencing. However, since here mutations cannot be directly assigned to subclones the subclonal information needs to be deconvolved. Here, the deconvolution is based on mutation frequencies, which is challenging for the identification of nested and similar size subclones. In con-

trast, single cell sequencing information provides a direct assignment of mutations to single cells. Here, a major challenge is the elevated error rates, allelic dropout and uneven coverage compared to traditional bulk sequencing data. To robustly account for these sources of noise we first identify sites which are likely to show a mutation in at least one cell. This is achieved by efficiently computing the probabilities of all possible mutation combinations across cells. Then the phylogeny of the tumor is computed with a stochastic search over the possible tree space via a Markov-Chain-Monte-Carlo scheme. In addition to offering a maximum likelihood phylogeny and a mutation to cell assignment, we provide a confidence to the mutation calls by sampling from the posterior. In contrast to existing methods, by using evolutionary information of the tumor tissue our approach enables us to reliably call mutations for a single cell even in the absence of sequencing information, as we demonstrate on several data sets.

Employing a Mixture Nested Effects Model to account for the variance of a mixed population of single cells.

Martin Pirkl

CBG, BSSE, ETH Zürich

New technologies allow for the elaborate measurement of different traits of single cells. These data promise the opportunity to elucidate causal intra-cellular mechanisms in unprecedented detail. Insights like those help not only to learn how cells generally function, but why and at what point they cease to function properly or in the worst case are deregulated. That de-regulation can lead to life threatening diseases like cancer. The battle against those diseases benefits from our understanding of cellular function on the single cell level. We follow the assumption, that all cells harbor the same underlying signaling pathways. However, the data, which is usually produced only shows a snapshot of the pathway at different times for each single cell, thus leading to a high variance for the observed multi trait phenotypes. We employ a mixture model to estimate the snapshot of the underlying graph for each cluster of cells from the whole population and combine the family of inferred networks to one consensus signaling pathway with detailed logical structures. That information is further be inspected to forward our understanding of cellular function.

Combining population and single-cell RNA-Seq to investigate determinants of successful HIV expression reactivation

Monica Golumbeanu

CBG, BSSE, ETH Zürich

HIV establishes latency in a minority of cells that are infected. These cells represent one reservoir from where the virus can reinitiate new rounds of infection and are currently considered as a major obstacle to HIV cure. The switch between HIV latency and HIV production from infected cells is at the center of HIV eradication strategies, including the "shock and kill" strategy that uses pharmacological and immunological agents to purge this reservoir. To date, it has been shown that the infected cells are not equal and that HIV gene expression depends on multiple parameters. These include the viral integration site location within the host cell genome and the host cell protein composition, which is affected by the cellular activation state and by environmental conditions. To investigate heterogeneity in the transcriptional reprogramming during HIV latency and reactivation, primary human CD4+ T cells were infected with an HIV-based vector and allowed to return to a resting, latent state for about 4 weeks. Subsequently, latently infected cells were exposed to either SAHA, a histone deacetylase inhibitor, or to anti-CD3/anti-CD8 antibodies mimicking antigen-mediated T-cell receptor stimulation. RNA-Seq from bulk as well as Fluidigm-isolated single cells was performed for each condition. We explored transcriptional heterogeneity of HIV and host cell gene expression using various statistical models designed for bulk and single-cell data analysis in order to identify subpopulations of cells based on their differential expression profiles. This type of analysis aims to identify transcriptional programs leading to successful reactivation of HIV expression, and to facilitate screening of future reactivating agents.

Exploiting heterogeneity in single-cell transcriptomic analyses: how to move past comparisons of averages

Keegan Korthauer

Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health

The ability to quantify cellular heterogeneity is a major advantage of single-cell technologies. Specifically, it is now possible to elucidate gene expression dynamics that were invisible using bulk RNA-seq, such as the presence of distinct expression states. However, statistical methods often treat cellular heterogeneity as a nuisance. I will present a novel method to characterize differences in expression in the presence of distinct expression states within and among biological conditions. I will demonstrate that this framework can detect differential expression patterns under a wide range of settings. Compared to existing approaches, this method has higher power to detect subtle differences in gene expression distributions that are more complex than a mean shift, and can characterize those differences. The R package scDD implements the approach, and is available on Bioconductor.

Unraveling Cortical Development Using Population and Single-cell RNA-Seq Data

Zahra Karimaddini

Department of Biosystems Science and Engineering, ETH Zurich, Switzerland

The brain, as part of the central nervous system, is the most complex organ in the mammalian body and the mechanisms that regulate its development are poorly understood. During brain development, neural stem cells (NSCs) generate thousands of different neuronal subtypes that are organized in precise and functionally distinct layers of the cerebral cortex. This process is a prerequisite for normal brain functions and any deviation from the standard developmental path can lead to debilitating brain disorders. To unravel these mechanisms, we study changes in the expression of transcription factors and signaling components in NSCs and progenitor populations (NeuroStemX, SystemsX.ch). To this end, we use population and single-cell RNA sequencing of each population at daily intervals during mouse cortical development obtaining a data set containing more than 100

population samples and more than 1000 single cells. This enabled us to identify a set of novel genes that characterizes NSCs and progenitor cells at the population and single cell level at distinct stages of brain development. Using machine learning methods, we identified a continuous differentiation path and, from this, determined different transcriptional states. Remarkably, we can show that the single cells follow a similar differentiation path to that predicted from transcriptional analysis at the population level. In addition, the single cells can be divided into subpopulations that emerge over time. To summarize, our gene expression data of different cell types at the population and single-cell level for daily intervals during neurogenesis combined with the appropriate data analyses give an unprecedented insight into the complex process of stem cell patterning and fate decision making in early brain development.

Unsupervised identification of multifurcations in high-dimensional single cell data with TreeTop

Will Macnair

ETH IMSB

Branching processes, including hematopoeisis and clonal evolution of cancer, are an important area of study. Recent advances in single-cell technology such as mass cytometry and single-cell RNAseq permit branching processes to be measured with a high level of dimensionality, via simultaneous measurement of 10s / 1000s of species. However, approaches to date have been restricted to points where at most three branches meet, and are limited to datasets which are well-represented by trees.

By fitting an ensemble of trees to the data and seeking points which partition these trees consistently, our algorithm TreeTop identifies points of bifurcation and their corresponding branches. TreeTop then allows analysis of transitions undergone during such branching processes, in terms of evolution of markers and transitional gated celltypes.

TreeTop represesents a cell population by an ensemble of trees, which connect random points evenly distributed through the dataset to represent subpopulations. Connections in each tree represent the range of possible transitions between these subpopulations. The ensemble of trees is visualized via a force-directed graph layout algorithm. To identify points of bifurcation within this ensemble of trees, we define a point of bifurcation as where three or more branches of a process meet. Across an ensemble of trees, cutting

at such points induces consistent partitions of nodes into branches which are consistent across the ensemble of trees. TreeTop utilizes a score evaluating the consistency of such partitions to identify such points of bifurcation and their corresponding branches.

We have successfully applied TreeTop to datasets representing varying topologies: simply branching (T cell development in the thymus), circular (cell cycle) and multifurcating (healthy human bone marrow). The points of bifurcation identified fit with biological expectations. TreeTop also identifies multiple layers of bifurcation within synthetic data sampled from a hierarchy of bifurcations. TreeTop permits points of bifurcation to be explored in less well-characterized datasets.

Accurate modeling and correction of GC content and gene length bias of single-cell RNA-seq

Hubert Rehrauer

ETH Zurich and University of Zurich

There are very diverse single cell library preparation protocols in practical use that provide either 3'-end or full-length transcript coverage. Next to the obvious differences implied by these two strategies, we observe GC content and gene length effects that are specific to the individual protocols. Additionally we observe also cell-to-cell variations within individual protocols, and for the joint analysis of heterogeneous datasets a library bias correction (LBC) is needed. While several approaches that deal with GC content and gene length dependent biases in RNA-seq based transcript abundances exist, they are not satisfactorily because they do not consider the potential interdependence and interaction of both effects. Additionally, they are prone to introduce artifacts for genes that have partially or entirely zero counts. Thirdly, the biases may affect a large fraction of genes such that the assumed global normalization schemes are invalid. With our work, we present a novel method for library bias correction with a true two-dimensional function that simultaneously depends on GC content and gene length. Our approach is unique by modeling not the absolute biases but the sample-specific deviations from the dataset wide bias. The computed correction factors are precise even in the case of wide-spread zero counts. The presented method is useful for correcting data sets with subsets of deviating cells as well as for the joined analysis of different data sets generated with different biases.

scEM: An EM algorithm for Differential Expression Analysis of Single-Cell RNA-seq

Agus Salim

La Trobe University and Walter and Eliza Hall Institute of Medical Research

We developed an EM algorithm to perform differential expression analysis of single-cell RNA-seq data, specifically UMI-based data. For a cell type, the unobserved number of molecules for a particular gene follows zero-inflated negative binomial (ZINB) distribution with mean parameter proportional to the size factor parameter that measures differences in starting materials across cells. Dropout phenomenon in which only a fraction of molecules is observed due to low capture efficiently is modelled using logistic regression whose parameters are estimated using external spike-ins. Simulation studies demonstrated that our approach has better sensitivity (at similar specificity) when compared to SCDE, MAST and BPSC. The log fold-change estimates are also robust when capture efficiency in endogenous RNA is higher than the external spike-ins, as long as the ratio of the two capture efficiencies is fixed across cells. We demonstrated the potential applicability of our method by applying it to analyse differential expression of mouse blood cells at different stage of development.

DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation

Fabian Müller

Max Planck Institute for Informatics, Saarbrücken, Germany

Although virtually all cells in an organism share the same genome, regulatory mechanisms give rise to hundreds of different, highly specialized cell types. These mechanisms are governed by epigenetic patterns, such as DNA methylation, which determine DNA packaging, spatial organization, interactions with regulatory enzymes as well as RNA expression and which ultimately reflect the state of each individual cell. Using low-input and single-cell whole genome bisulfite sequencing, we generated genome-wide DNA methyla-

tion maps of blood stem and progenitor cells [1]. These maps enabled us to characterize cell-type heterogeneity, and aggregating methylation levels of small pools of cells and single cells across putative regulatory regions, we dissected the DNA methylation dynamics of human hematopoiesis. We observed lineage-specific DNA methylation patterns between myeloid and lymphoid progenitors and associate these patterns to regulatory elements, gene expression and chromatin accessibility. Using statistical learning, we were able to accurately infer cell types from DNA methylation signatures and the resulting models could be used for a data-driven reconstruction of the human hematopoietic system. Our observations illustrate the power of DNA methylation analysis for the in vivo dissection of differentiation landscapes as a complementary approach to lineage tracing and in vitro differentiation assays. The generated methylome maps and analysis methods provide a comprehensive framework for studying epigenetic regulation of cell differentiation and blood-linked diseases.

[1] Farlik, M., Halbritter, F., Müller, F., et al. (2016). DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. Cell Stem Cell, 19(6), 808-822. http://doi.org/10.1016/j.stem.2016.10.019

Statistical methods for differential discovery in multi-dimensional flow and mass cytometry (CyTOF) data

Lukas Weber

Institute of Molecular Life Sciences, University of Zurich

Recent technological advances in multi-dimensional flow cytometry and mass cytometry (CyTOF) enable the measurement of up to 50 protein marker expression levels per cell, allowing cell populations to be characterized in unprecedented detail. Due to the high dimensionality, significant efforts are underway to develop automated data analysis methods to replace traditional "manual gating" or visual inspection of projected data. Differential discovery experiments aim to detect biological features that vary consistently between biological samples in different conditions, such as diseased and healthy. For example, in cytometry, it may be of interest to detect differentially abundant cell populations, or differential expression of functional markers within specific cell populations, especially rare populations. We have tested or adapted multiple new as well as existing methods for performing differential discovery analyses in large-scale, multi-dimensional,

small-sample cytometry data. Our methods make use of a range of statistical techniques, including empirical Bayes moderation of variances and functional data analysis. In particular, we take into account information contained in biological replicate samples, which are crucial for statistical inference. Our methods are benchmarked against existing approaches, such as Citrus and CellCnn, using simulated and experimental data; and will be available as an R/Bioconductor package.

Imputation of single-cell DNA methylation profiles by transferring information across cells

ChantriolnI Andreas Kapourani

University of Edinburgh, School of Informatics

New technologies enabling the measurement of DNA methylation at the single cell level are promising to revolutionise our understanding of epigenetic control of gene expression. Yet, intrinsic limitations of the technology result in very sparse coverage of CpG sites, effectively limiting the analysis repertoire to a semi-quantitative level. Here we propose a Bayesian hierarchical method to quantify spatially-varying methylation profiles across genomic regions from single-cell Bisulphite sequencing data (scBS-seq). The method clusters individual cells based on the methylation profiles, enabling the discovery of epigenetic diversities and commonalities among individual cells. The clustering also acts as an effective regularisation method for imputation of methylation on unassayed CpG sites, enabling transfer of information between individual cells. We show that the resulting imputation is highly accurate both on simulated and real data sets.
Joint work with Guido Sanguinetti

Latent Factor Analysis of scRNAseq

Daniel Wells

University of Oxford

A common challenge in the analysis of single cell RNA sequencing data is clustering of cells into biologically appropriate classes. We show that SDA (Sparse Decomposition of Arrays) is able to simultaneously identify cell types along with the genes which define them. SDA decomposes a digital gene expression matrix into components which have associated cell scores and gene loadings. In this framework a cell type can be represented as a component where the cell scores indicate which cells are members and the gene loadings indicate which genes are active. SDA uses a bayesian framework with a sparsity prior on the gene loadings which facilitates interpretation of the cell classes in biological terms. We compare this method to other commonly applied techniques such as PCA (principal components analysis) and ZIFA (zero inflated factor analysis) using both real datasets and simulated data.

On the probability of differential distribution in single-cell RNA-seq

Michael Newton

University of Wisconsin, Madison

Accounting for the mixture of unlabeled cell populations underlying a sample of single-cell RNA-Seq profiles leads to natural structural constraints on the form of empirical Bayesian inference for changes in gene-level distributions of gene expression. We describe a computationally and statistically efficient model-based approach that delivers posterior probabilities of differential distribution through a novel discrete mixture defined jointly over expression changes per cellular type and over subtype cell proportions. We illustrate the computations in both synthetic data and on data from embryonic stem cell experiments. This is joint work with Xiuyu Wang and Christina Kendziorski.

Evaluation of methods for differential expression analysis of single-cell RNA-seq data on a collection of consistently processed public data sets

Charlotte Soneson

University of Zurich

As single-cell RNA-seq becomes increasingly widely used, the amount of publicly available data grows rapidly. This provides a useful resource for computational method development as well as for reanalysis and extension of published results. In public data repositories, typically both the raw data and a processed data set, used by the data generators for their analysis, are available. However, the procedure to obtain the processed data set is tailored to the specific application, and can be widely different between data sets. We present conquer, an open collection of consistently processed, analysis-ready single-cell RNA-seq data sets. For each data set, we provide count and TPM estimates for both genes and transcripts, as well as quality control and exploratory analysis reports to assist users in determining whether a particular data set is suitable for their purpose. To illustrate the usefulness of the repository, we use some of the data sets to perform an extensive evaluation of multiple methods for differential gene expression analysis. Several methods developed specifically for single-cell RNA-seq data were compared to existing methods developed for bulk RNA-seq differential expression analysis. Considerable differences were found between the behaviour of different methods, but also for the same method applied to different randomly selected subsets of a given data set. We further investigate the characteristics of the significant genes called with different methods, and show that gene filtering can have a substantial effect on the performance.

Discovering gene and drug connections via Image-based morphological profiling

Juan C. Caicedo

Broad Institute of MIT and Harvard

Microscopy images are now used as a source of quantitative information in a variety of biological experiments. Images acquired for studying phenotypes in cell biology capture high resolution morphological variations of cells. This morphological information can

be extracted by transforming pixels into single cell measurements that encode the phenotypic changes of the experiment. The collection of single cell measurements ("profiles") can be used to reveal similarities and differences between cell populations under different experimental conditions.

In our lab, we aim to make morphological information computable and reproducible through the development of open source software tools (such as CellProfiler and Cytominer) that can create rich quantitative profiles efficiently. In collaboration with biology labs, we use morphological profiling to address fundamental research questions such as identifying the functional impact of variants of a gene associated with cancer, identifying drugs to target diseases with known genetic basis, and uncovering novel functions for unannotated genes. These are just a few examples of the wide range of biological applications that can be approached using image-based morphological profiling We expect image-based profiling and analysis to be powerful tools that complement well-established -omics methods to address these challenging questions.

A comparison of bioinformatics tools to analyze single-cell transcriptomes

Elisabetta Mereu

CNAG-CRG, Universitat Pompeu Fabra (UPF), Barcelona, Spain

Single cell RNA-seq has become a powerful method to explore cell-to-cell variability in transcriptome profiles with unprecedented resolution. However, single cell transcriptomics data contain many sources of technical noise (e.g. dropout events or batch effects), often hiding the real structure and challenging downstream analysis. To address this, many tools have been proposed to characterize cell heterogeneity using co-expressed gene sets and to identify differentially expressed transcripts. To evaluate the performance of such tools , we generated high quality single cell transcriptome datasets and analyzed cell heterogeneity using Pagoda, CleanCount and Seurat. To identify differentially expressed genes we applied statistical methods implementing scDE and scDD tools. We investigated the robustness of the results by comparing gene clusters and differentially expressed genes by their distribution of overlap sizes between methods. The stability in cell clustering has been assessed using bootstrap resampling, based on metrics such as Jaccard similarity coefficients. We assessed the strengths of the methods in terms of their sensitivity to detect subpopulations. We also critically discussed pipeline characteristics,

such as input formats, handling of covariates, applicability and/or specific requirements and assumptions. Finally, we suggest an analysis strategy, combining these tools in order to optimize results and maximize information from single cell transcriptional profiles.

Quantifying punctuated equilibrium via stochastic modeling across cancer types

Simona Cristea

Dana-Farber Cancer Institute & Harvard School of Public Health

Understanding the evolutionary dynamics of cancers is an essential step towards clinical success. Recently, various studies have reported experimental and theoretical evidence of cancer progressing via relatively few instances of simultaneous genomic alterations (such as point mutations, copy number alterations or chromosomal rearrangements). This model of tumor evolution has been termed "punctuated", as opposed to the "gradual", or classical model, in which tumors sequentially accumulate genomic alterations across large periods of time. Here, we investigate these two paths of cancer progression via stochastic modeling, across various cancer types. We simulate tumor evolution by a multi-type branching process with various mutation rates and fitness distributions. After learning some of the model parameters from bulk sequencing data, we use single cell data to fit the two types of models. Based on these fits, we devise a new measure for punctuated equilibrium and propose various biological mechanisms that can explain the higher likelihood of punctuated evolution in various cancer types.

Hierarchical population model for multivariate single-cell data

Carolin Loos

Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, 85764 Neuherberg, Germany

Joint work with: Katharina Moeller, Fabian Fröhlich, Tim Hucho, Jan Hasenauer

A comprehensive understanding of biological systems requires the investigation of heterogeneity and its underlying sources and mechanisms. To elucidate cellular heterogeneity, mechanistic population models need to be calibrated to single-cell data, which are collected by e.g. flow cytometry or microscopy. However, the simulation and analysis of these models is challenging and therefore hinders parameter estimation and model selection. Especially the study of cell populations that comprise multiple subpopulations remains an unsolved problem.

We present a data-driven modeling framework for heterogeneous cell populations, which combines mixture modeling with approaches for the approximation of distributions to account for several levels of heterogeneity. This method facilitates the detection of causal differences between subpopulations and between cells of the same subpopulation. Its computational efficiency allows the comparison of hundreds of competing hypotheses and thus enables a detailed study of biological processes.

We apply our method to artifical data and experimental data of NGF-induced Erk signaling in neuronal cell populations. Our approach elucidates the influence of the extracellular matrix on pain sensitization and is able to detect different levels and sources of variability. Our results suggest that the presented method will enable a better understanding of cellular heterogeneity.

Stochastic Profiling for mRNA-Seq Data

Lisa Amrhein

Helmholtz Zentrum München, Institute of Computational Biology

Acute Myeloid Leukemia (AML) is a type of blood cancer affecting the myeloid lineage. The incidence of AML increases with age, and it is the most frequent type of acute leukemia among adults. Although approximately 70% of patients achieve complete remission, very small numbers of leukemic cells remain and cannot be detected with current diagnostic techniques. Nearly everybody with AML will relapse in the end if no further postremission or consolidation therapy is given, and this relapse is almost always lethal.

AML patients frequently carry a mixture of different cancer cell types, so-called subclones, which evolve over time, so that the mixture at relapse is different from the one at diagnosis. Understanding clonal evolution and identifying rare subclones, especially for those mutations causing relapse, is still an open challenge.

We aim to parameterize transcriptional heterogenity from mRNA-Seq counts taken from small groups of cells (e.g. 10-cells). To that end, we will extend our Stochastic Profiling Method previously proposed for microrray data. This technique infers single-cell regulatory states by mathematically deconvolving n-cell measurements. This averaging-and-deconvolution approach allows us to quantify single-cell regulatory heterogeneities while avoiding the technical measurement noise of single-cell techniques.

Integrative analysis of single-cell expression data reveals distinct regulatory states in bidirectional promoters

Fatemeh Behjati

Max Planck Institute for Informatics

Bidirectional promoters (BPs) are prevalent in eukaryotic genomes. However, it is poorly understood how the cell integrates different epigenomic information, such as transcription factor (TF) binding and chromatin marks, to determine directionality of gene expression at BPs. Single cell sequencing technologies are revolutionizing genetics and this project focuses on the integration of single-cell RNA data with bulk ChIP-seq and other epigenetics data for which single cell technologies are not yet established. We utilized novel human single cell RNA-seq data to reveal clusters of BP genes exhibiting various states of directionality across individual cells. For instance, we found a cluster with highly expressed genes at both genes of a BP for almost all single cells. However, some BP genes are expressed in an alternating manner, where the expression of one gene always dominates the other one, and vice versa, depending on the subpopulation of cells. We integrated other levels of genomic and epigenomic information to shed light on this previously unrecognized complexity in BP gene regulation. We found unique TF motifs and binding patterns associated with these expression states. Further, we stratified different Histone Modifications (HM) on these clusters. Despite the fact that the clusters are derived from the single cell data, the bulk HM and TF profiles were consistent with those states. It is an interesting research direction to try to deconvolute bulk-seq data with single-cell expression data, although many statistical challenges are as of yet unexplored.

---

scmap: a tool for fast and accurate mapping of cells to a reference database using scRNA-seq data

Vladimir Kiselev

Sanger Institute

In recent years, technological developments have allowed researchers to collect up to 10^6 cells. Moreover, large scale projects such as the NIH BRAIN Initiative and the Human Cell Atlas aim to generate even larger datasets. However, analyzing and integrating of large single-cell RNA-seq datasets remains challenging. Here, we present scmap - a tool for fast and accurate mapping of cells to a reference database for scRNA-seq data. We demonstrate that scmap can be used for integrating publicly available different datasets collected from different platforms and labs. Moreover, we show that scmap can be used for one of the central tasks of a cell atlas: projecting new samples (e.g. from a disease model) onto an existing reference.

Bayesian Unidimensional Scaling for latent ordering and uncertainty estimation

Lan Huong Nguyen

Stanford University

Detecting patterns in high-dimensional multivariate datasets is non-trivial. Clustering and dimensionality reduction techniques often help in discerning inherent structures. In biological datasets such as microbial community composition or gene expression data, observations can be generated from a continuous process, often unknown. Estimating data points' 'natural ordering' and their corresponding uncertainties can help researchers draw insights about the mechanisms involved.

We introduce a Bayesian Unidimensional Scaling (BUDS) technique which extracts dominant sources of variation in high dimensional datasets and produces their visual data summaries, facilitating the exploration of a hidden continuum. The method maps multivariate data points to latent one-dimensional coordinates along their underlying trajectory, and provides estimated uncertainty bounds. By statistically modeling dissimilarities and applying a DiSTATIS method to their posterior samples, we are able to incorporate visu-

alizations of uncertainties in the estimated data trajectory across different regions using confidence contours for individual data points. We also illustrate the estimated overall data density across different areas by including density clouds. One-dimensional coordinates recovered by BUDS help researchers discover sample attributes or covariates that are factors driving the main variability in a dataset. We demonstrated usefulness and accuracy of BUDS on a set of published microbiome 16S and single cell RNA-seq data. Our method effectively recovers and visualizes natural orderings present in datasets. Automatic visualization tools for data exploration and analysis are available at: `https://github.com/nlhuong/visTrajectory`.

Supervised classification and post-stratification with scRNAseq

Andrew McDavid

University of Rochester

To reveal latent structure in single cell RNA sequencing (scRNAseq) experiments many unsupervised clustering methods have been developed. In some cases, however, a subset of cells may have known population classifications, and these can serve to train a predictive model. In other cases, a historical experiment, perhaps of bulk data, provides labels and expression measurements to train a model which is hoped to generalize onto a new, unlabeled scRNAseq experiment. Here I consider various basis expansions tailored for scRNAseq data, and argue that these enhance the out-of-experiment predictive performance compared to use of the raw expression values. I also review the accuracy and calibration of several procedures for supervised classification, and ways in which classification uncertainty can be propagated onto downstream analysis.

Integrated single cell data analysis for understanding mechanisms of neuronal diversity

Jean Yang

University of Sydney

Technological advances such as large scale single cell transcriptome profiling has exploded in recent years and enabled unprecedented insight into the behavior of individual cells. Identifying genes with high levels of expression using data from single cell RNA sequencing can be useful to characterize very active genes and cells in which this occurs. In particular single cell RNA-Seq allows for cell-specific characterization of high gene expression, as well as gene coexpression. In this talk, I will describe a versatile modeling framework to identify transcriptional states motivated by a neuronal single cell project.

Neuronal cell systems exhibit extraordinary levels of complexity. Thus it is of great interest to explore the ways in which this neuronal diversity is generated and manifested to encompass achieve such complexity. One such mechanism is patterns of gene transcription across neurons. We will describe how a gamma-normal mixture model is used to identify active gene expression across cells; we then use these to characterise markers for olfactory sensory neutron cell maturity and to build cell-specific coactivation networks. We found that combined analysis of multiple datasets results in more known maturity markers being identified, as well as pointing towards some novel genes that may be involved in neuronal maturation. We also observed that the cell-specific coactivation networks of mature neurons tended to have a higher centralization network measure than immature neurons. Finally, we will describe an approach to evaluate evidence of gene transcriptional mosaics as a mechanism for achieving diversity of neuronal cells.

Single-Cell Phenotype Classification Using Deep Convolutional Neural Networks

Beate Sick

ZHAW

Deep learning methods are currently outperforming traditional state-of-the-art computer vision algorithms in diverse applications and recently even surpassed human performance

in object recognition. A short introduction in deep learning and convolutional neural network models will be given. Then we demonstrate the potential of deep learning methods to high-content screening-based phenotype classification. We trained a deep learning classifier in the form of convolutional neural networks with approximately 40,000 publicly available single-cell images from samples treated with compounds from four classes known to lead to different phenotypes. The input data consisted of multichannel images. The construction of appropriate feature definitions was part of the training and carried out by the convolutional network, without the need for expert knowledge or handcrafted features. We compare our results against the recent state-of-the-art pipeline achieved by experts in the field and demonstrate a reduced classification error rate when using a deep learning approach.

Targeted Drop-seq: a hypothesis-driven approach to single-cell transcriptomics

Andreas Gschwind

EMBL Heidelberg

Gene expression and regulation are central questions in many fields of biology, ranging from applied medical to fundamental evolutionary biology. Transcriptomic technologies such as RNA-seq became the standard in the post-genomic era to analyze gene expression and how genes are regulated. Recent advances in single-cell RNA sequencing (scRNA-seq) enable to measure cell-to-cell differences in gene expression, which remain elusive in conventional bulk sequencing approaches. Droplet-based technologies such as Drop-seq emerge as powerful tools for whole transcriptome sequencing on single-cell level. They enable the analysis of an exceptionally high number of single cells at low cost, however the high number of assessed transcripts can decrease the power to precisely measure expression changes of individual genes. We develop an adapted version of the Drop-seq protocol for targeted transcriptomics, where the expression of a pre-defined set of transcripts is measured exclusively. By this, we expect an increase in sensitivity with which changes in expression of the targeted transcripts can be measured compared to conventional Drop-seq at the same sequencing depth. This approach is suitable for a wide range of hypothesis-driven applications, where a group of targets can be defined. Such sets might contain genes in a specific genomic region or genes involved in a biological process of interest. One promising application could be in conjunction with single-cell pertur-

bation methods such as CRISPR, where multiple genomic elements can be targeted by specific sgRNAs in a pooled approach. Targeted Drop-seq could then be used to monitor the resulting expression changes of genes of interest in individual cells, thus enabling to efficiently test a large number of regulatory hypotheses.

3D-organoid segmentation and drug-response testing with a deep neural network

Jan Sauer

German Cancer Research Center (DKFZ)

Early detection and treatment of colorectal cancer is vital to the long-term survival of a patient. The goal of this study is explore the response of pharmaceuticals available to treat colorectal cancer in a patient specific way with the goal to make a treatment recommendation. Recently, organoids have been found to be useful tools to model organs and study drug-response phenotypes on a tissue level. 3D high-content microscopy screening of organoids grown from both human and mouse colon tissue, after treatment with various FDA approved compounds, promises to show how new drugs and drug combinations may affect the growth and survival of cancerous cells. The analysis of these 3D culture images requires the development of novel software capable of accurately segmenting the individual cells of these spheroids from the background. The problem is to detect partially overlapping objects of a specific shape but with largely varying diameter in the images. To achieve this, a deep neural network (DNN) is being developed for the segmentation and the subsequent feature extraction of the images. The design of the DNN allows the detection of edges of the organoids on its first layers. On the higher levels of the network, the edge map is used to detect the organoids and to determine their size. To train this network a small number of annotated images are required. From these training samples, a large number of small image blobs are extracted that are used to train the edge detector. After segmentation of organoids and single cells, phenotypic features are calculated and averaged to an experiment specific feature vector. With the currently available data, we will show that we can distinguish the phenotypic effect of different drugs and have a basis for treatment recommendation.

Cell lineage tracing in zebrafish using CRISPR/Cas9 induced genetic marks

Maria Florescu

Hubrecht Institute, Utrecht University

Understanding how a zygote grows into a complex, multicellular organism is a central question in developmental biology. We present scartrace, a strategy for whole-organism lineage tracing based on Cas9 induced genetic marks, named scars, in several GFP integrations of a zebrafish line (Junker et al, BioRxiv 2016). Scarring is a dynamic process, meaning that scars are progressively introduced over multiple rounds of cell division. Using scartrace, we investigate morphogenesis in the early zebrafish embryo from scars of adult organs. We assign organ tissue to the corresponding germ layers and estimate the number of clones generating a certain tissue from bulk scarred tissue. Recently we achieved single cell scar detection, which greatly improves lineage-tracing resolution.

Measuring batch effects in single-cell RNA sequencing data

Maren Büttner

Institute of Computational Biology, Helmholtz Zentrum München

Quantification of the complete transcriptome of single cells using RNA-seq is a versatile tool for exploring heterogeneous cell populations, but suffers substantially from technical noise and batch effects. This observation has motivated the development of several batch correction and data normalisation techniques to separate technical and biological variation in the data. Measuring batch effect amounts to quantifying the difference of distributions in different batches to decide whether they are equivalent, or not. Traditionally, this decision has been based on visual inspection of dimension-reduced representations of the data, as obtained, for example, by principal component analysis. A generalization to this approach involves testing the significance of a batch effect in each component of principal component analysis, which has high statistical power, but still it is computationally inefficient and hard to interpret. We present kBET, a batch effect test based on k-nearest neighbours that tests nonparametrically for inequivalence of distributions. The batch label is known for each sample; hence a particular spatial subset (i.e. a subset of samples

that potentially forms a cluster in the high-dimensional space) is hypothesised to be composed of the same batch labels (with equal fractions) as the complete data set. Being easy to interpret and computationally efficient, kBET overcomes the above-mentioned limitations while keeping high statistical power. We demonstrate its performance on various experimental and simulated data sets and demonstrate substantial contribution of the batch effect in single-cell RNAseq data that is intractable by inspecting projected data (as in principal component analysis for example). Thus, kBET enables an unbiased comparison of the efficiency of recently developed batch effect correction techniques.

Exploring influenza A virus infection using single-cell RNA-sequencing

Lam-Ha Ly

Max-Planck-Institute for Molecular Genetics

In virology influenza A infected cells show a large cell-to-cell variability in the number of released progeny virions, the virus yield. To study the cellular heterogeneity of infected cells we performed single-cell RNA-sequencing (scRNA-seq) to profile both, the cellular and the viral transcriptome using the Smartseq2-protocol. Here, we treated Madin-Darby Canine Kidney (MDCK) cell lines with influenza A viruses and classified cells having a high and low virus yield. Additionally, as a reference, we performed bulk RNA-sequencing to profile infected and non-infected cells. This approach enables the investigation of host-pathogen cell interactions in order to explain differences in the viral replication and integrity. First computational analysis includes a comparison between bulk and single-cell samples and differential expression analysis of high and low virus yield cells. Preliminary results will be shown.

Single cell mRNA sequencing to reveal the in vivo hierarchy of miRNA targets

Andrzej Rzepiela

ScopeM ETH

By promoting target mRNA decay, miRNAs can down-regulate mRNA targets, reduce the cell-to-cell variability of protein expression, and provide a channel through which mRNA targets can act as "competing RNAs" to influence each other's expression. The relative importance of these regulatory levels for cellular function is strongly debated, owing to the regulatory network size and the lack of in vivo measurements of miRNA-target interaction parameters. Combining single cell analysis with mathematical modeling we derived a method for estimating Michaelis-Menten constants of endogenous miRNA targets, comprehensively and in the context of live cells. Applying the approach to two distinct miRNAs and using data from hundreds of single cell mRNA-Seq measurements, we tested the possibility to uncover a hierarchy of targets, and to find the targets that are very sensitive to the miRNA presence. We show that the sensitivity of a target to the miRNA is determined by the interplay between the rates of target-miRNA association and dissociation as well as the intrinsic and miRNA-induced target decay rates. We discuss the characteristic of the most sensitive targets that we discovered with the method. Our approach enables elucidation of complex behaviors resulting from the interactions of a large number of targets with a common regulator.

Single Cell DNA Sequencing Reveals a Late-Dissemination Model in Metastatic Colorectal Cancer

Alexander Davis

The University of Texas MD Anderson Cancer Center

Metastasis is a complex process and has been difficult to study in human patients. A major technical obstacle has been the extensive intratumor heterogeneity at the primary and metastatic tumor sites. To address this problem, we developed a highly-multiplexed single cell DNA sequencing approach to trace the metastatic lineages of two colorectal cancer (CRC) patients with matched liver metastases. Single cell copy number and muta-

tional profiling was performed on 444 cells, in addition to bulk exome and deep-targeted sequencing. In the first patient we observed monoclonal seeding, in which a single clone had evolved a large number of mutations prior to migrating to the liver to establish the metastatic tumor. In the second patient we observed polyclonal seeding, in which two independent clones seeded the metastatic tumor after having diverged at different time points from the primary tumor lineage. The single cell data also revealed an unexpected independent tumor lineage that did not metastasize, and early progenitor clones with the first hit in APC that subsequently gave rise to both the primary and metastatic tumors. Collectively, these data reveal a late-dissemination model of metastasis in two CRC patients, and provide an unprecedented view of metastasis at single cell genomic resolution.

Modelling zeros for differential expression and feature selection in scRNASeq

Tallulah Andrews

Wellcome Trust Sanger Institute

Single cell RNA sequencing (scRNASeq) detects far fewer genes in each cell than traditional RNASeq. As a result typical RNASeq expression datasets contain a majority of zero values (dropouts), even after filtering out poor quality cells and undetected genes. Existing RNASeq analysis methods are ill-suited for dealing with these dropouts. We demonstrate how modelling zeros in scRNASeq data, both with and without unique molecular identifiers (UMIs), can improve the identification of biologically important genes (feature selection), and accuracy of differential expression (DE) testing. We introduce two novel methods for differential expression testing which perform favourably against 11 published methods for both UMI-tagged and full-transcript single-cell RNASeq data. In addition, we use the relationship between gene expression and the frequency of dropouts to identify features corresponding to differentially expressed genes without a priori knowledge of cell populations. These features are robust to batch effects and they can be used to determine consistent cell types across multiple datasets from the same biological system. We apply this method to a meta-analysis of single-cell data of mammalian development and identify novel markers for distinct embryonic stages.

Investigating Microenvironment-to-cell Signaling in 3D Spheroids through Imaging Mass Cytometry

Vito Zanotelli

IMLS UZH/ Life Science Graduate School Zurich, Systems Biology Program

Joint work with: Georgi F, Schulz D, Schapiro D, Andriasyan V, Yakimovich A, Catena R, Jackson H, Bodenmiller B

Question Every cell senses its local 3D environment and adapts its phenotype, accordingly. This process is critical in tissue development and homeostasis. Deregulated it can lead to diseases such as cancer. However, how the interactions between cells and their environment shape cellular phenotypes in tumors and contribute to tumor heterogeneity is largely unknown. Here we set out to develop a high throughput setup to quantify the influences of these interactions on the phenotypic heterogeneity in an in vitro 3D spheroid cell culture system.
Methods We developed a workflow based on metal label barcoding. It allows an efficient coupling of 3D spheroid assays with an imaging mass cytometry (IMC) readout. This enables the simultaneous quantification of more than 40 phenotypic and functional markers at subcellular resolution in 3D microtissue slices. Firstly, this system is suitable to perform large scale studies on spatial relationships of complex phenotypes. Secondly, it can be used to follow perturbations mediated by small molecule inhibitors.
Results We first demonstrate the feasibility of the barcoding approach. Based on data of unperturbed breast cancer spheroids we show how IMC can capture phenotypic heterogeneity and coordination in the microenvironment. Finally, we explore how such data can be integrated using mathematical modeling to gain quantitative insights.
Conclusions We present the development of a broadly applicable, scalable screening approach efficiently combining high throughput 3D tissue culture with IMC. The approach can be applied to more complex cell culture settings, including 3D co-cultures, organoids as well as advanced perturbations, such as stimulation time courses. Combined with inhibitor screens, the technology is a solid basis to investigate spatial coordination in 3D tissue models.

Systematic analysis of cell phenotypes and cellular social networks in tissues using multiplexed image cytometry analysis toolbox (miCAT)

Denis Schapiro

University of Zurich, Institute of Molecular Life Sciences Systems Biology PhD Program, ETH and University of Zurich

Single-cell, spatially resolved 'omics analysis of tissues is poised to transform biomedical research and clinical practice. We developed an open-source computational multiplexed image cytometry analysis toolbox (miCAT) to enable the interactive, quantitative and comprehensive exploration of single cell phenotypes, cell-to-cell interactions, microenvironments and morphological structures within intact tissues. We highlight the unique abilities of miCAT by analysis of highly multiplexed mass cytometry images of human breast cancer tissues.

FACSanadu: An open source tool for rapid visualization and quantification of flow cytometry data

Thomas Bürglin

Dept Biomedicine, University of Basel

1) European Molecular Biology Laboratory European Bioinformatics Institute Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD UK
2) Department of Biomedicine University of Basel Mattenstrasse 28 CH 4058 Basel Switzerland
* presenting author

Motivation: Flow cytometry is a fundamental technique in cell biology, yet few open source packages are available to analyze these data. Here we describe FACSanadu, an interactive package for rapid visualization and measurement of FCS data. It is the first open source package that can read data of the COPAS Biosorter. Availability and Implementation: FACSanadu is implemented in Java and uses the Qt framework for display. Binary distributions are made for all major operating systems (Windows, Macintosh, Linux). The source code and documentation is available as free software at www.facsanadu.org.

Network-based regularized optimization for cancer survival data analysis

Susana Vinga

IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Learning statistical models from oncological data has now become a major challenge due to the significant increase of molecular information available. The inherent high-dimensionality of these 'omics datasets, where the number of features largely exceeds the number of observations, leads to ill-posed inverse problems and, consequently, to models that often lack interpretability and are prone to overfitting. In order to tackle this problem, regularized optimization has emerged as a promising approach, allowing to introduce constraints on the structure of the solutions. These include the application of sparsity-inducing norms on the loss function, for which the l1-norm leading to the Lasso (least absolute shrinkage and selection operator) method is probably the best-known example, along with the elastic net, based on the l1 and l2-norms. More recently, several network-based regularizers have been proposed to analyze survival data when the features have a graph relationship, such as the Net-Cox, which promotes parameter smoothness across the network, and DegreeCox, where the degree of each node is taken into account. Preliminary testing on breast and ovarian cancer survival data from the The Cancer Genome Atlas (TCGA) illustrates some of the advantages of these methods. The comparison criteria include the concordance c-index for Cox models, and the p-value of the log-rank tests for the separation of high vs. low-risk patients. Preliminary results show an improvement of the c-index when network information is included, whereas the separation seems to decrease. However, the models estimated by cross-validation are different in terms of number of selected features, which may hamper the correct comparison of the results. Extensions to other norms are being conducted and may lead to improved survival models in terms of interpretability and overfitting control, a key aspect to improve and support clinical decision systems and prognostic assessment of oncological patients.

Finding subtle differences in single cell RNA-seq data

Camille Stephan-Otto Attolini

IRB Barcelona

Cell populations characterized by different functional programs usually present radical differences in the expression of relatively large sets of genes. In the context of single cell RNAseq this is easily detected despite the highly variable nature of the count data.
Colon cancer stem cells have until now been considered to be a homogeneous population defined mainly by the expression of the LGR5 gene. We analyzed a large dataset of LGR5 positive (LGR5+) cells and their one-division progeny (LGR5-) in order to determine the existence of heterogeneity within the LGR5+ cells and the relation to the populations seen in LGR5- cells. We found that applying existing methodologies for SC-RNAseq lead to no apparent heterogeneity among the stem cell population. We hypothesized that the subtle differences expected among these cells were masked by the high abundance of zero counts and the large differences between the expressions of highly and lowly expressed genes.
Our methodology consists of quality filtering of cells and genes, a transformation of the count matrix, simple dimensional reduction methods and unsupervised clustering. Our results suggest the existence of two populations characterized by genes associated with proliferation and with the gene Mex3a. The expressions of these two gene sets are inversely correlated as expected from experimental observations. We investigated the robustness of these clusters and found that a minimum number of cells is necessary to observe the different populations. We were also able to detect progenitors of known cell types present in the differentiated colon tissue among the LGR5- cells.
Our investigation suggests that specific methodologies may be necessary depending on the nature of the data when dealing with SC-RNAseq data.

TRACE: Reconstructing trajectories of cell cycle evolution using single-cell mass cytometry data

Maria Anna Rapsomaniki[1], *Xiaokang Lun*[2], *Johanna Wagner*[2], *Bernd Bodenmiller*[2] *and Maria Rodriguez Martinez*[1]

[1] *IBM Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland*
[2] *Institute of Molecular Life Sciences, University of Zurich, 8057 Zurich, Switzerland*

As single-cell experimental approaches become increasingly popular, cell-to-cell heterogeneity has emerged as a key determinant factor contributing to variability in gene expression and signaling responses. Mass cytometry (CyTOF) is a new proteomic technology that enables the simultaneous quantification of dozens of proteins in thousands of individual cells. In the context of cancer research, recent applications of CyTOF include the characterization of inter- and intra-tumor heterogeneity and the identification of novel cell subpopulations. However, as already demonstrated for single-cell RNA-seq, the resulting measurements are largely influenced by confounding factors, such as the cell cycle and cell volume. We present here TRACE, a novel computational approach to quantify this source of variability. TRACE first exploits a hybrid machine learning approach to classify single cells into discrete cell cycle phases according to measurements of established markers. Next, a metric embedding optimization technique creates a one-dimensional continuous marker that tracks biological pseudotime and individual cells are subsequently ordered according to this pseudotime marker. The resulting cell cycle trajectories across perturbation time points allow us to separate cell cycle effects from experimentally induced responses, enabling the direct comparison of signaling responses through cell cycle progression. Additionally we show that volume biases can be corrected using housekeeping gene measurements. Our approach, implemented in a simple and intuitive Graphical User Interface, was used to analyze data from various cell lines subject to different stimulations. In each case, TRACE was able to separate confounding effects from signaling responses, enabling the unbiased analysis of biological processes.

Seeing is Believing: Biology-Driven Visualizations for Single Cell High Dimensional Datasets

Yann Abraham

Janssen Pharmaceutica

Joint work with: Bertran Gerrits, Marie-Gabrielle Ludwig, Frederik Stevenaert, Greet Vanhoof, Anish Suri, Pieter Peeters, Caroline Gubser Keller

Visualization and interpretation of high dimensional data is a common challenge, and several solutions have been developed that address specific issues such as the identification of groups of points sharing similar characteristics. Most methods rely on projections to a plane, and any information on the initial dimensions is lost. When applied to biological dataset, this makes the interpretation of the resulting visualization difficult, whether one is trying to infer the nature of the different groups of points or to determine if the differences between samples are biologically relevant. To overcome this challenge we applied to CyTOF datasets two visualizations that had been previously developed to visualize trends in high dimensional data. First, we used Radviz to project all cells in a two-dimensional space in a way that maintains the relation between each point and every channel that has been measured. Using standard visualization techniques several conditions can be compared at the sample level to identify trends in the data. Once groups of relevant cells have been identified we used Fan plots to visualize the contribution of each channel and to estimate the variability of each channel within the group. Fan plots can also be used to compare a given group of cells across several conditions. Both methods can be used together or in addition to other methods, including analytical ones. Compared to other methods such as t-SNE and SPADE, Radviz and Fan plots make the interpretation of complex datasets easier. They provide an important interface between analytical methods and the biology.

Identification of Expanded Clonotypes based on TCR sequencing of Single T-Cells

An De Bondt

Janssen Research & Development, Beerse 2340, Belgium

Our aim is to characterize the TCR repertoire as a method to identify onset and to intercept and cure disease by developing targeted strategies. To this end we developed a platform to identify expanded T-cell clones. T-cells are a major component of the immune system, and the T-cell Receptor (TCR) plays a pivotal role in the identification of antigens. Via recombination of the TCR locus, each T-cell expresses a different TCR which is specific to a given antigen. All recombined TCR sequences across all T cells within an individual, the 'T-cell repertoire', provides the specificity required for a proper immune response. The interaction between a TCR and a given antigen triggers activation and expansion of the T-cell, accompanied by a switch in gene expression that ultimately modulate the T-cell function. We have established and validated protocols for TCR sequencing on cell lines as well as on single T-cells from healthy donors. After ex-vivo stimulation of peripheral blood mononuclear cells (PBMCs), we sequenced the most variable region of both chains of the receptor. The analysis of the TCR sequencing data, allowed us to identify the oligoclonal expansion of T-cells. For this analysis, we evaluated tools like MiXCR and TraCeR and further developed visualisations and standardized reporting. In conclusion, we have established a platform for analyzing the immune repertoire which can be readily applied to diseases where immune infiltration is at play such as oncology, hepatitis and diabetes.

Linear mixed effects models for complex experimental designs in single cell mass cytometry

Marjolein Crabbe and Marianne Tuefferd

Janssen Research and Development BE, Translational Biomarkers, Infectious Diseases
Johnson & Johnson Pharmaceutical Research & Development

Joint work with: W. Talloen, Y. Abraham, Y. Zhang, G. Vanhoof, F. Stevenaert, K. Spittaels, J. Bollekens, J. Aerssens

Single cell profiling technologies open new perspectives in clinical research, aiming at improved insights into disease heterogeneity and detailed patient response profiles to treatments. Integrating the increasing complexity of modern clinical study designs in the data analysis strategy of single cell assessments poses, however, significant challenges. Here we propose an analysis approach for single cell patient-derived data that considers the clinical study setup, based on linear mixed effects modeling.

To illustrate this approach, a CyTOF (Cytometry by Time-Of-Flight) dataset generated from patient samples collected in a cross-sectional study that recruited multiple resolver types of chronic HBV patients is considered. Ex vivo stimulations of individual patient samples (paired stimulated versus unstimulated samples) and the potential interaction of the stimulation with the distinction in clinical resolver profiles in this study, contribute to the complexity of the experimental design. In addition to 19 markers used in the CyTOF analysis to annotate different cell types, each individual cell is characterized by 18 functional channels. Hence, earlier described methods for multivariate statistical analysis of such high-dimensional datasets are challenging in this complex experimental design setting.

The proposed exploratory statistical analysis approach of the single cell CyTOF data generated in the context of more complex experimental designs assesses the advantage of mixed models in their ability to identify functional differences within specific cell populations between the different resolver groups. Applying a univariate approach, the paired design structure embedded in the experimental study design can be exploited and the potential interaction between cohort type and stimulation accounted for, enabling a more refined analysis. Moreover, mixed models provide a flexible framework to analyze a variety of extensions including repeated sampling in a longitudinal study design.

Participants

Yann Abraham Janssen Pharmaceutica
Nicola Aceto University of Basel **Lisa Amrhein** Helmholtz Zentrum München, Institute of Computational Biology
Tallulah Andrews Wellcome Trust Sanger Institute
Ricard Argelaguet European Bioinformatics Institute
Eirini Arvaniti ETH Zurich
Niko Beerenwinkel CBG, BSSE, ETH Zurich
Fatemeh Behjati Max Planck Institute for Informatics
Kobi Benenson ETH Zurich
Nicolas Bennett Seminar for Statistics, ETH Zurich
Bernd Bodenmiller University of Zurich
Thomas Bürglin Dept Biomedicine, University of Basel
Peter Bühlmann ETH Zürich
Maren Büttner Institute of Computational Biology, Helmholtz Zentrum München
Juan C. Caicedo Broad Institute of MIT and Harvard
Ambrose Carr Memorial Sloan Kettering Cancer Center
Luciano Cascione Bioinformatics Core Unit at IOR
Francesc Castro-Giner University of Basel
Marjolein Crabbe Janssen Research and Development BE
Simona Cristea Dana-Farber Cancer Institute & Harvard School of Public Health
Helena Crowell University of Zurich
Alexander Davis The University of Texas MD Anderson Cancer Center
An De Bondt Janssen Research & Development
Maria Florescu Hubrecht Institute, Utrecht University
James Gagnon Harvard University
Johann Gagnon-Bartsch University of Michigan, Statistics
Javier Gayan F. Hoffmann-La Roche
Monica Golumbeanu CBG, BSSE, ETH Zürich
Raphael Gottardo Fred Hutchinson Cancer Research Center

Andreas Gschwind EMBL Heidelberg
Stephanie Hicks Dana-Farber Cancer Institute / Harvard T.H. Chan School of Public Health
Takashi Hiiragi EMBL Heidelberg
Susan Holmes Statistics Dept, Stanford
Yuanhua Huang School of Informatics, University of Edinburgh
Wolfgang Huber EMBL Heidelberg
Steffen Jaensch Johnson & Johnson
Katharina Jahn CBG, BSSE, ETH Zurich
Hans-Michael Kaltenbach ETH Zurich
Chantriolnt Andreas Kapourani University of Edinburgh, School of Informatics
Zahra Karimaddini BSSE, ETH Zurich, Switzerland
Peter Kharchenko Harvard Medical School
Vladislav Kim European Molecular Biology Laboratory
Vladimir Kiselev Sanger Institute
Jan Korbel European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany, and European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, UK.
Keegan Korthauer Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health
Jack Kuipers CBG, BSSE, ETH Zurich
Ivo Kwee Bioinformatics Core Unit at IOR
Griet Laenen Open Analytics
Lisa Lamberti University of Oxford
Prisca Liberali FMI Basel
Carolin Loos Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology
Junyan Lu European Molecular Biology Laboratory
Lam-Ha Ly Max-Planck-Institute for Molecular Genetics
Will Macnair ETH IMSB
John Marioni Cancer Research UK, Cambridge
Tobias Marschall Saarland University / Max Planck Institute for Informatics
Davis McCarthy EMBL-EBI, Hinxton, UK
Andrew McDavid University of Rochester
Elisabetta Mereu CNAG-CRG, Universitat Pompeu Fabra (UPF), Barcelona, Spain
Eugenia Migliavacca NIHS
Laurent Modolo CNRS - Lyon
Fabian Müller Max Planck Institute for Informatics, Saarbrücken, Germany

Nick Navin MD Anderson
Michael Newton University of Wisconsin, Madison
Lan Huong Nguyen Stanford University (prof. Susan Holmes Lab)
Martin Pirkl CBG, BSSE, ETH Zürich
Michael Prummer NEXUS, ETHZ
Magnus Rattray University of Manchester
Maria Anna Rapsomaniki IBM Research Laboratory
Hubert Rehrauer ETH Zurich and University of Zurich
Andre Rendeiro CeMM Research Centre for Molecular Medicine of the Austrian Academy of Sciences
Edith Ross Cancer Research UK Cambridge Institute, University of Cambridge
Andrzej Rzepiela ScopeM ETH
Agus Salim La Trobe University and Walter and Eliza Hall Institute of Medical Research
Jan Sauer German Cancer Research Center (DKFZ)
Denis Schapiro University of Zurich, Institute of Molecular Life Sciences Systems Biology PhD Program, ETH and University of Zurich
Geoffrey Schiebinger Broad Institute
Alexander Schönhuth Centrum Wiskunde & Informatica, Amsterdam
Sven Schuierer Novartis Institutes of Biomedical Research
Christof Seiler Department of Statistics, Stanford University
Beate Sick ZHAW
Jochen Singer CBG, BSSE, ETH Zurich
Charlotte Soneson University of Zurich
Michael Stadler Friedrich Miescher Insitute for Biomedical Research
Oliver Stegle EMBL-EBI Hinxton
Daniel Stekhoven ETH Zürich
Camille Stephan-Otto Attolini IRB Barcelona
Barbara Szczerba University of Basel
Ewa Szczurek Institute of Informatics, Faculty of Informatics, Mathematics and Mechanics, University of Warsaw
Valérie Taly Paris Descartes
Marianne Tuefferd Translational Biomarkers, Infectious Diseases, Johnson & Johnson Pharmaceutical Research & Development
Catalina Vallejos Alan Turing Institute and UCL
Lars Velten EMBL
Britta Velten EMBL Heidelberg
Jean-Philippe Vert ENS Paris
Susana Vinga IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Sijian Wang Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison USA
Lukas Weber Institute of Molecular Life Sciences, University of Zurich
Daniel Wells University of Oxford
Jean (Zhijin) Wu Brown University
Jean Yang University of Sydney
Vito Zanotelli IMLS UZH/ Life Science Graduate School Zurich, Systems Biology Program