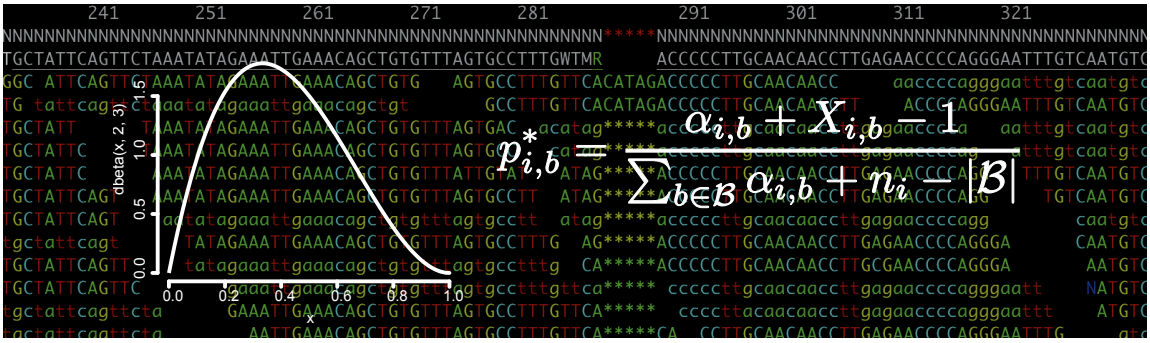
Centro Stefano Franscini

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Statistical Challenges and Biomedical Applications of Deep Sequencing Data

Monte Verità, Ascona, Switzerland, 5–10 June 2011

Organizors: Niko Beerenwinkel, Peter Bühlmann, Darlene Goldstein



May 31, 2011

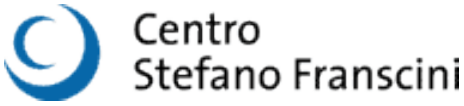
http://www.cbg.ethz.ch/news/ascona2011

Abstract

High-throughput DNA sequencing has opened up new avenues for investigating biological systems. The new sequencing technology triggers many statistical and computational challenges. This interdisciplinary workshop is intended to be a forum for (i) the dissemination of cutting-edge biotechnological and methodological developments and (ii) the identification of challenging data analysis problems.

Il sequenziamento ad alta capacità del DNA (deep sequencing) ha fornito nuove possibilità di studio dei sistemi biologici. Le nuove tecniche di sequenziamento forniscono inedite sfide di tipo statistico e computazionale. Questa conferenza interdisciplinare vuole essere un'occasione per (i) diffondere i più innovativi sviluppi biotecnologici e metodologici e (ii) identificare i problemi rilevanti che emergono nell'analisi dei dati.

Swiss Institute of Bioinformatics



Venue

Monte Verità
Via Collina 84
CH-6612 Ascona
tel. +41 91 785 40 40
http://www.csf.ethz.ch/

About the Centro Stefano Franscini (CSF)

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

Travel directions

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

External hotel

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

Shuttle service from Locarno Station

A free 13-seater shuttle bus to Monte Verità will leave from Locarno railway station on Sunday, June 5. The shuttle meeting point is next to the train platforms in Locarno (see Figure 1). The schedule is as follows:

Sunday, June 5	
Stazione di Locarno → Monte Verità	Note
15.20	stop at hotel Luna
16.00	stop at hotel Luna
16.40	stop at hotel Luna
17.20	direct
18.00	direct
19.05	direct

Moreover, two additional runs will connect hotel Luna to Monte Verità and back:

18.35	hotel Luna → Monte Verità
21.15	Monte Verità → hotel Luna

During the workshop, the shuttle will run according to the following schedule:

Monday, June 6	
09.00	hotel Luna → Monte Verità
21.15	Monte Verità → hotel Luna
Tuesday, June 7	
08.45	hotel Luna → Monte Verità
21.15	Monte Verità → hotel Luna
Wednesday, June 8	
08.45	hotel Luna → Monte Verità
after dinner	ristorante Costa Azzurra → hotel Luna
Thursday, June 9	
08.45	hotel Luna → Monte Verità
21.15	Monte Verità → hotel Luna
Friday, June 10	
08.45	hotel Luna → Monte Verità
after lunch	Monte Verità → stazione Locarno

Excursion and social dinner

The excursion will take place Wednesday, June 8, in the afternoon. We will visit Valle Verzasca. Afterwards, we will go to Ristorante Costa Azzurra for the social dinner.

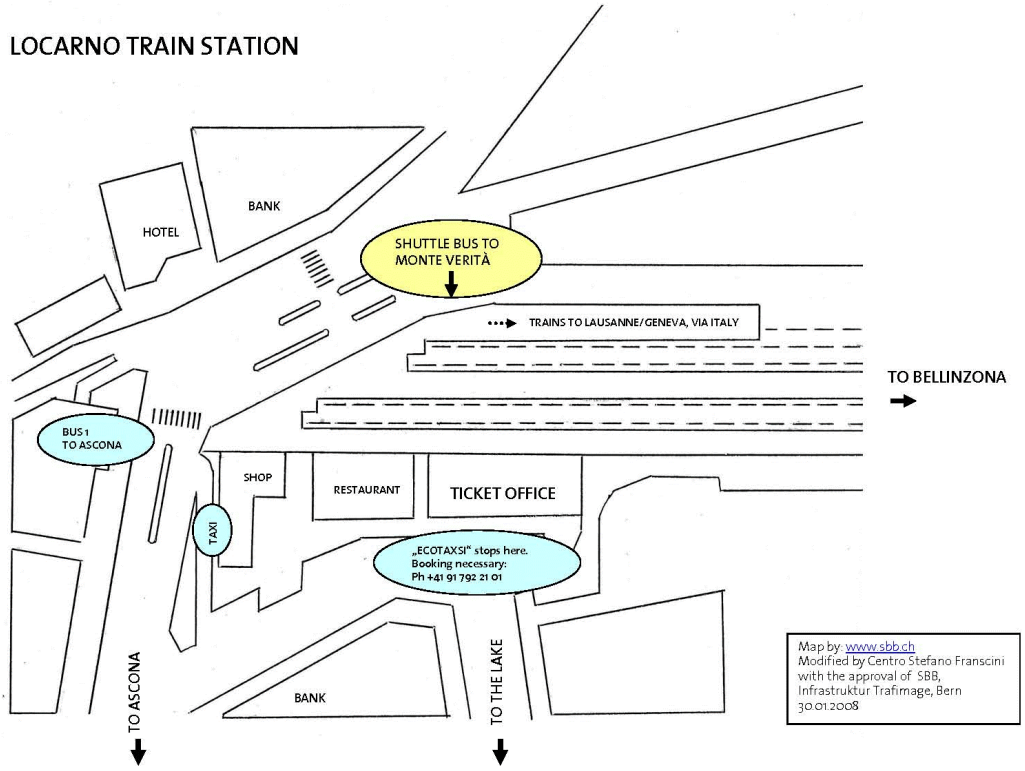


Figure 1: Shuttle bus location

Keynote lectures

Discovering indel and copy number genomic variation from paired-end sequencing

Michael Brudno

Department of Computer Science, University of Toronto

High throughput sequencing (HTS) technologies have enabled the inexpensive sequencing of human genomes, and the discovery of some genomic variants from the resulting short read datasets is well underway. In this talk I will present algorithms for discovery of two types of variants from HTS data: smaller indels (< 50bp) and copy number variants (CNVs). First, I will describe MoDIL: Mixture of Distributions Indel Locator, a novel method for finding insertion/deletion polymorphisms from paired short reads. We explicitly model each genomic locus as a mixture of two haplotypes, and our method takes advantage of the high clone coverage to identify both homozygous and heterozygous variation, even if the individual clone sizes are unreliable. Analysis of a recently sequenced genome demonstrates that MoDIL accurately identifies indels ≥ 20 nucleotides. I will then describe a method to predict CNVs from paired short reads. Our method combines information from paired short reads to identify variable regions and depth-of-coverage to predict the true copy count in the donor genome. Together, the two datasets help overcome both sequencing biases of HTS platforms and spurious read mappings. Our method allows for the detection of CNVs within segmental duplications. We use our method to detect CNVs within the same dataset, and make a total of $\sim$5000 calls that show high concordance with previously known CNVs in this individual.

What ChIP-seq data tell us about gene regulation

Philipp Bucher

EPF Lausanne

The advent of ChIP-Seq amounts to a revolution in transcription regulation research. Thanks to this new technology we are now able to see where specific transcription factors bind to DNA in vivo in a particular cell type. Moreover, ChIP-seq experiments targeted at histone modifications provide a detailed picture of the local chromatin state and are capable of localizing individual nucleosomes at near base-pair resolution. I will present a number of case studies exemplifying how computational analysis of ChIP-Seq data complemented by DNA motif-based and comparative genomics-based approaches can unravel molecular mechanisms of gene regulation. Specifically, I will focus on the mechanism that direct transcription factors to their physiological target sites, and on the consequences of transcription factor-DNA interactions on local chromatin structure in higher eukaryotes.

Inferring the evolution of cancer genomes

Chris Greenman

University of East Anglia, Norwich

Deep sequencing data is resulting in the construction of relatively complete mutation portfolios for cancers. Here we discuss two statistical problems that arise from these constructions; firstly, how can we best use the data to infer copy number, and secondly, how can we use paired end data to infer the rearrangment history of the genome. To answer the first approach we show that a Mixed Markov Poisson Process can be used to segment the data without having to bin the data, resulting in segmentation with highly accurate breakpoints. To answer the second question we show that combining copy number information with rearrangments identified from paired reads results in a graph theoretic construction from which rearrangment history can be inferred. Recent data from deep sequencing of primary samples are used to highlight these methods.

Integrating leaf growth and circadian regulation – a systems approach

Willem Gruissem

ETH Zurich

Leaves originate from meristematic stem cells by cell division and subsequent cell elongation. During this developmental process leaves attain their typical flat structure with adaxial and abaxial surfaces. The continuous ontogeny and morphogenesis of a typical dicot leaf has been divided into specific stages based on anatomical features and developmental potential. The emerging leaf is a metabolic sink, while later in development it becomes a source tissue in which metabolism is regulated during the circadian rhythm. A full understanding of the developmental process and metabolic changes requires quantitative data at various levels. As part of the EU FP6 project AGRON-OMICS we have acquired large multi-scale datasets of Arabidopsis leaf number 6 at four stages of development and at end-of-day/end-of-night from three leaf growth baseline conditions (optimal water, water deficit and 18 hrs light). The data are incorporated into a relational database built with MySQL and linked to external databases for integrated data analysis. I will present first results from the integrated data analysis that reveal interesting developmental correlations between DNA ploidy and protein synthesis as well as unexpected dynamic end-of-day/end-of-night mRNA regulation that is not reflected at protein levels.

Applications to transcriptomics – characterisation and functions of non-coding RNAs

Wolfgang Huber

EMBL, Heidelberg

There is an abundance of non-coding transcription around the loci of coding genes, and complex patterns of sharing of nucleosome-depleted regions at 5' and 3' ends of transcripts have been observed. The function, if any, of this transcription is not well understood. I will report our findings, in yeast, on the pervasiveness of bidirectional promoters, and of widespread ultrasensitivity effects in gene expression regulation mediated

by non-coding transcription on the opposite strand (antisense) of coding genes.

eQTL studies, where genetic variations in a population are associated with, and thus causally linked to, variations in gene expression, are a powerful tool to prioritize those genetic variations with potential downstream phenotypic effects. Until now, such studies have mostly focused on single nucleotide polymorphisms (SNPs). I will report our findings on the roles of copy number variations (CNVs). CNVs have a substantial impact on gene expression, and thus likely on downstream phenotypes. They are in many causes plausible candidates for being the causal variant; in addition, when they are unlinked to a "tagging SNP", CNV genotyping closes an important gap left by SNP genotyping.

The HeLa cell is a popular model system in cell biology for investigating processes related to cell cycle and metabolism. More recently, genomic tools such as genome-wide RNAi libraries and NGS are applied to this model. I will present a 30x coverage genome sequence dataset of a HeLa line and will report on some of our findings on the extent of genetic differences between this model system's genome and the human reference genome, and discuss the rationale for using this system in genomics.

The regulation of gene expression levels in mammals

John Marioni

EMBL-EBI, Hinxton (Cambridge)

Understanding the mechanisms that regulate gene expression levels is critical for developing insights into numerous biological processes. In this talk, I will discuss how data generated using RNA-sequencing, in conjunction with information about variation in DNA sequence both within and between species, can be used to understand better the regulation of gene expression. My talk will focus primarily on describing how we have used RNA-sequencing to compare the divergence of liver gene expression levels among closely related mouse strains with the difference in expression observed between alleles of their first generation inter-strain hybrid offspring (F1). I will describe how our analysis, which employs a hierarchical Bayesian model, enables a comparison of gene expression levels in the F1 crosses to the corresponding parental expression levels, allowing the identification of genes whose expression levels are consistent with: imprinting, a pattern of cis regulatory evolution, a pattern of trans regulatory evolution, and a pattern of compensatory evolution in cis and in trans. After outlining the statistical model, I will describe how we can use it to obtain insights into the regulation and evolution of gene expression levels.

This work is performed in collaboration with the groups of Duncan Odom (CRUK CRI), Paul Flicek (EBI), and Alvis Brazma (EBI).

Graphical models for cancer signalling

Sachi Mukherjee

Department of Statistics and Centre for Complexity Science, University of Warwick, U.K.

Signalling networks play a key role in the control of diverse cellular processes; their aberrant functioning is heavily implicated in cancer. Genomic aberrations in cancers are thought to "re-wire" the normal connectivity of signalling networks, with important biological and therapeutic implications. Yet we remain limited in our understanding of cancer-specific signalling. Advances in proteomics have begun to enable quantitative studies of signalling at the network level. I will discuss statistical approaches by which to elucidate signalling networks in cancer, focusing on the use of sparse graphical models to interpret high-throughput proteomic data and thereby generate biologically-testable hypotheses regarding signalling phenotypes that are specific to biological context.

Characterization of somatic mutations in cancer

Ben Raphael

Department of Computer Science and Center for Computational Molecular Biology, Brown University

Cancer is driven in part by somatic mutations that accumulate in the genome during an individuals lifetime. Next-generation DNA sequencing technologies now enable the measurement of these mutations across many cancer samples. I will describe computational techniques to address two challenges in interpreting data from these large-scale sequencing studies: (1) Deriving reliable measurements of structural aberrations in cancer genomes from the short DNA sequences produced by sequencing machines; (2) Distinguishing functional driver mutations responsible for cancer from random passenger mutations. For detection of structural aberrations, I will describe algorithms to analyze

paired-end, mate pair, and strobe sequencing data from different sequencing technologies. For driver mutation prediction, I will describe two approaches for identification of *driver pathways*, groups of genes containing driver mutations, in a large cohort of cancer samples. In the first approach, we use prior information about interactions between genes to identify pathways (or networks) of genes that are frequently mutated across samples. In the second approach, we optimize a measure derived from the statistical properties of mutations on driver pathways. I will describe applications our algorithms to data from The Cancer Genome Atlas (TCGA) and other large cancer sequencing projects.

Integrative analyses of epigenome sequencing data

Mark D. Robinson

Senior Postdoctoral Fellow Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Melbourne Epigenetics Laboratory, Garvan Institute of Medical Research, Sydney Australia

The complement of expressed genes in a cell determines its physiological state; approximately 200 cell types result from alternative readings of the common DNA instructions. Cancer is a disease where both genetic (e.g. point mutations, copy number variations) and epigenetic aberrations (e.g. DNA methylation, histone modifications) force changes in gene expression, such as activation of oncogenes and silencing of tumour suppressors. So, aberrations in expression levels (or the structure of expressed transcripts) need to be put in context with the changes in the genome and epigenome. High-throughput sequencing (HTS) technologies are revolutionizing the study of cancer biology and large-scale data collection efforts are well underway. In this talk, I will discuss our recent efforts to integrate genomic, epigenomic and transcriptomic profiles, comparing normal and cancer cells.

Issues on the joint analysis of two genomes: the viral infection of the cell

Amalio Telenti

Institute of Microbiology, University of Lausanne

A paradigm of infectious diseases is the co-existence of two genomes. There is however limited experience in the understanding of the reciprocal influences of the pathogen and its host. Specifically to viral infection, the virus acts as a parasite that co-opts cellular factors, re-programs the cell transcriptional profile, and escapes from the triggering of a cellular antiviral state. We have approached this model by investigating the 24 hours life cycle of HIV, the virus of AIDS, in the infected cell. During this time HIV converts from foreign RNA to DNA, enters the nucleus of the cell, integrates into the host genome, initiates transcription and translation using the cellular machinery, and exits the cell. These molecular milestones serve to structure the cellular response as measured through the use of deep sequencing. The emerging pattern from this model is that of an early cellular shutdown, followed by the identification of a set of genes that are upregulated in consonance with the key viral processes. The timed events are organized in biologically plausible scenarios demonstrating the recruitment of cellular factors by the virus. Current analyses aim at testing the hypothesis that key recruited factors, if knocked down through functional assays will result in the arrest of the viral replication.

Talks

Statistical methods for comparative RNA-seq studies

Simon Anders, Alejandro Reyes, Wolfgang Huber

EMBL Heidelberg

RNA-Seq has been used with much success so far to create inventories of the transcriptome of samples. The next step is now to use comparative assays to study the regulation of isoform expression in a quantitative way. Several commonly used approaches make improper assumptions on the noise properties of RNA-Seq count data, namely, they typically ignore overdispersion. We show that, especially in data sets from very deep sequencing, this leads to serious loss of type-I error control. We discuss empirical observations about the typical character of noise in RNA-Seq data and propose methods to estimate dispersion and to test for differential expression as well as for alternative isoform regulation using generalized linear models. We present an R package to perform such analyses in a convenient yet flexible manner.

Robust subnetworks: computing confidence values for functional modules

*Daniela Beisser, Thomas Dandekar, Gunnar Klau†, Marcus Dittrich, Tobias Müller**

Department of Bioinformatics, University of Wuerzburg, Wuerzburg, Germany; † Centrum Wiskunde & Informatica (CWI), Science Park Amsterdam, Netherlands; *Corresponding author
Tobias.Mueller@biozentrum.uni-wuerzburg.de

High-throughput data provides a wealth of information (on genomic, transcriptomic and

proteomic level) that is widely used in integrated network analysis. Several heuristic approaches exist that allow to identify functional modules, pathways or gene signatures, containing differentially expressed genes in the context of protein-protein interaction networks. Recently, an exact approach was introduced by Dittrich and co-workers that resolves the subnetwork-finding problem to optimality using integer linear programming. The objective of the presented study is to assess the accuracy and variability of the identified optimal functional modules. Therefore, we propose a novel concept of a consensus module based on jackknife resampling. This allows to compute support values for the nodes and edges of the resulting optimal functional module.

Since our objective is not only to find a module which obtains a good accuracy but also yields results that are robust to minor changes in the underlying data, we assess the robustness and variability of the obtained solutions in an extensive simulation study. Furthermore, the presented resampling procedure is applied to two biological microarray data set on diffuse large B-cell lymphomas (DLBCL) and acute lymphoblastic leukemia (ALL). Robust parts of the resulting functional module are identified by the assigned support values of the nodes and edges. These include for the DLBCL data a well-known NF?B signature of up-regulated genes in the ABC subtype which has a poor prognosis.

The algorithm including integration of data, scoring of nodes and methods for network search and visualization are implemented in the open-source R package BioNet available from `http://bionet.bioapps.biozentrum.uni-wuerzburg.de` and the Bioconductor project.

Association testing in sequencing studies: a novel framework

Brooke L. Fridley, Abra Brisbin

Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA.

As the majority of variants discovered via sequencing are expected to be rare, methods are needed that are able to detect the effect of multiple rare variants on the phenotype. Most current analysis methods for rare variants have focused primarily on collapsing approaches for variants in some region/gene of interest and impose strong assumptions for analysis. We propose a method to overcome these limitations for testing for association, referred to as D-MAF test, which: does not assume that all rare variant effects are deleterious; does not require a definition of rare; allows incorporation of weights based on MAF or predicted function; and can incorporate both rare and common variants. In addition to

these features, the testing procedure can be extended to the analysis of pooled sequencing data, for which many collapsing methods are not applicable. Preliminary simulations show the D-MAF test to be robust to genetic model with good power, as compared to other existing rare variant testing methods.

A test for low-frequency mutations in ultra-deep sequencing data of heterogeneous tumors

Moritz Gerstung[a,b], *Holger Moch*[c], *Peter Wild*[d], *Christian Beisel*[a] *and Niko Beerenwinkel*[a,b]

[a] ETH Zurich, Department of Biosystems Science and Engineering, Mattenstrasse 26, 4058 Basel; [b] SIB–Swiss Institute of Bioinformatics; [c] University Hospital Zurich, Institute of Surgical Pathology, Schmelzbergstrasse 12, 8091 Zurich; [d] ETH Zurich, Institute of Cell Biology, Schafmattenstrasse 18, 8093 Zurich

Recent systematic sequencing studies underline that cancer is driven by thousands of alterations in the genome and it is thought carcinogenesis is a somatic evolutionary process that is driven by the competition of malignant subclones in the tumor. Massively parallel sequencing technologies allow not only for assessing entire genomes, but also for resolving rare genetic variants in heterogeneous samples. We derive new statistically rigorous algorithms to distinguish genetic variants occurring at low frequencies from ubiquitous sequencing and PCR errors in deep sequencing data with ultra-high coverage. We present a likelihood ratio statistic for testing the presence of low-frequency single nucleotide variants that can be tuned by a prior that is learned across the genome. We demonstrate the power of the test on simulation studies and data from synthetic control experiments. When applied to data from multiple sites of a renal cell carcinoma we identify several low-frequency variants in a relatively short region of 15kb. Our work presents a first glimpse at the heterogeneity of a tumor cell population, uncovering parallel somatic evolution of the primary tumor and the metastasis in line with stochastic population genetics models.

Analysis of shotgun bisulfite sequencing of cancer samples

Kasper Daniel Hansen, Winston Timp*, Héctor Corrada Bravo*, Sarven Sabunciyan*, Benjamin Langmead*, Oliver G. McDonald, Bo Wen, Hao Wu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A. Irizarry, Andrew P. Feinberg*
** equal contributions*

Johns Hopkins University, Baltimore

A unified framework for the statistical analysis of rare variant sequence data, with applications to autism and study design

Iuliana Ionita-Laza[1], *Ruth Ottman*[2]

[1] Department of Biostatistics, Columbia University, New York, NY 10032; [2] G.H. Sergievsky Center and Departments of Epidemiology and Neurology, Columbia University, Epidemiology Division, NYSPI, New York, NY 10032

Rapid advances in sequencing technologies set the stage for large-scale medical sequencing efforts to assess the importance of rare variants in complex diseases. The low frequency and expected large number of such variants pose great difficulties for the statistical analysis of these data. We introduce here a novel concept for analysis of rare variants in biologically- related individuals: the effective number of rare variants in a set of relatives. This frame- work allows for unified approaches to the analysis of rare variants that can include both biologically-related cases, and unrelated cases and controls. We illustrate the approach with an application to autism, and discuss important implications for study design of large-scale sequencing studies to be performed in the near future.

Measuring microRNA expression by next-generation sequencing

Edoardo Missiaglia, Simona Rossi, Pratyaksha Wirapati, Mauro Delorenzi

Bioinformatics Core Facility, Swiss Institute of Bioinformatics, University of Lausanne, Switzerland

The use of massively parallel sequencing in microRNA expression profile has promised to provide more accurate measurements combined with a larger dynamic range, enabling also the detection of new microRNA molecules as well as isomeRs. However, this technique required a long and complex procedure which could introduce some biases in the measurement. Furthermore, studies designed to evaluate the reliability of this approach have shown controversial results.

Our group had the opportunity to analyze a large collection of 145 samples which were tested for the expression of small RNA sequences (between 15-30nts) using next-generation sequence. A 3' adapter barcodes were used in the RNA libraries preparation, so to analyze multiple samples per channel. After sequence trimming and demutiplexing, reads were aligned to the human genome (hg19) and other RNA databases (miRBase, Rfam, EST).

Unsupervised analysis of microRNAs reads showed a strong association between their expression and barcode sequences. This bias was observed also among other small RNA families. More accurate analysis of the sequence reviled that some barcodes were found to ligate RNA fragments with specific nucleotides sequences, particularly in proximity to the 3' end of the tags, which is the part that interact with the barcode during the ligation process. This bias was not observed when analyzing samples indexed using 5' adapter barcodes.

Overall our analysis has shown a poor performance of a protocol in which a 3' adapter barcode is use for sequence mutiplexing.

Statistical challenges in testing clonal relatedness of tumors using their genomic profiles

Irina Ostrovnaya

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering, New York

Multiple tumors in the same patient can either arise independently or originate from the

same cell and thus be clonal. This distinction has clinical implications for a patient and is important for understanding the biology of the disease. Many investigators have examined this issue in recent years by comparing the molecular profiles of tumors using studies of candidate markers or using genome-wide techniques, though these methods have not yet achieved clinical applicability. I will describe how the diagnosis can be accomplished by evaluating loss of heterozygosity in selected markers or copy number profiles from CGH or SNP arrays, and by comparing the tumors on the basis of these somatic molecular profiles. Statistical methodology includes specifying a similarity measure between tumors that takes into account the frequencies of mutations in a particular cancer. The reference distribution for the similarity measure can be obtained using permutation procedures. The assumptions behind our model will have to be re-evaluated as deep sequencing data for pairs of tumors from the same patient become available. I will discuss issues and assumptions that are likely to be important in quantifying similarity between deep-sequencing profiles of paired tumors for the purpose of developing valid statistical methods for evaluating clonal relatedness. This is the joint work with Colin Begg and Venkatraman Seshan.

Normalization and analysis of ChIP-seq of histone modifications that cover long stretches and are prevalent in the genome

Hubert Rehrauer

Functional Genomics Center Zurich

Published ChIP-seq analysis strategies are mainly tailored to immunoprecipitations of bound transcription factors where the ChIP-enriched regions are short ($<$100 bases) and occur typically ony a few thousand times in the genome. However, ChIP-seq experiments that target histone modifications have different properties: The ChIP-enriched regions can be long (103-106 bases) and highly prevalent ($>$104 different regions). As a result, the assumptions of many normalization schemes are violated as well as the assumption that the coverage of the enriched regions has a peak-like shape. Here we propose a simple normalization scheme that identifies a set of not-enriched genomic regions and uses those as reference for normalization. For the identification of enriched regions we make use of the genome annotation and look in a multi-stage approach first in potentially relevant regions like promoters and transcript regions. With this restricted search we increase the statistical power to detect enriched regions compared to a naïve genome-wide scan. Based

on the identified enriched regions we build a model with prior probabilities for width and height of enriched regions and do a model-based scan of the remaining genome. Based on two data sets measuring H4K5 and H4K12 acetylation respectively we show that our algorithm outperforms existing algorithms with respect to reproducibility in replicates and agreement of promoter acetylation and gene expression.

HIV haplotype inference using a constraint-based Dirichlet process mixture model

Volker Roth

Department of Computer Science, University of Basel

The recent introduction of the Deep Sequencing technology has led to a drastic increase in sequenced data, but the limited length of these sequenced reads and non-negligible sequencing errors have also opened new statistical challenges in analyzing the data. This work aims at analyzing deep sequencing data obtained in a single run from genetically diverse samples, in particular from intra-host HIV populations: HIV exists in a single infected patient as quasispecies, a population of evolutionarily related haplotypes. We present a constraint-based Dirichlet process mixture model for identifying (an a priori unknown number of) these haplotypes in a sample and thereby quantifying the observed genetic diversity, which defines an important step in administering personalized medication.

Some study designs to improve power in association tests for rare variants

Ingo Ruczinski

Johns Hopkins University, Baltimore

The assumption that common complex diseases are attributable in part to allelic variants that are reasonably common in a population is often termed the "common disease, common variant" hypothesis, and is the underlying rationale for genome-wide association

studies (GWAs). While GWAs have been successful identifying hundreds of such genetic variants associated with many complex diseases, the individual variants typically only represent a small increment in risk for any particular disease, and together, can usually explain only a small proportion of the familial clustering (heritability) observed. Thus, the paradigm has shifted somewhat towards whole exome and whole genome sequencing approaches to assess the effects of rare variants (with possibly larger effect sizes), which are poorly tagged by standard genotyping arrays. In this talk, we focus on family and population based study design considerations, and show how family records can be leveraged to improve power even in population based studies.

Causal gene ranking

Daniel J. Stekhoven

Seminar for Statistics, ETH Zurich

Discovering structure in high-dimensional, observational data, as for example in microarray gene expression experiments, is an elaborate and crucial task. We introduce a method to stably infer the causal influence of predictor variables on a response. Combining the estimation of Markov equivalence classes of directed acyclic graphs using the PC-algorithm and causal intervention calculus for the effects of the variables on the response, and putting these two parts in a stability selection environment, we are not only able to rank variables according to their stable, causal-type influence, but also to assign the per-comparison error rate to each of them. Our causal inference method takes the cumulative nature of effects through a cascaded pathway into account. Furthermore, assigning ranks using stability makes the approach less prone to sampling variability and allows to choose the amount of regularization. We apply our method to real data from observational gene expression experiments of Arabidopsis thaliana with floral development as response of main interest. High-ranked genes were validated by growing the corresponding knock-out mutants and collecting key response elements of floral development.

The effects of low-Level choices on detecting genetic variants with high-throughput sequencing

Margaret Taub

Johns Hopkins University, Baltimore

High-throughput sequencing is fast becoming a heavily used technology for detecting genetic variation in disease association studies, with the hope that detecting rare variants (which is challenging to do with genotyping microarrays) will help illuminate the genetic causes of common diseases. Several successes have been achieved in identifying causative variants for Mendelian disorders using sequencing data, and many large consortium- based projects are moving forward with sequencing large numbers of samples with the aim of revealing the genetic causes of common, complex diseases. In this work, I present some of the challenges of accurately and completely determining genetic variants using sequencing data, comparing results from different sequencing strategies, including whole-genome resequencing, exome sequencing, and targeted resequencing, using data provided by my collaborators and publicly available data. I will focus in particular on the impact of low-level processing choices (e.g., error correction and alignment methods) on the set of called variants produced from a sequencing data set. Time permitting, I will discuss implications of these challenges for the detection and testing of variants associated with complex diseases.

The effect of prenatal famine exposure on DNA methylation

Erik Van Zwet

Leiden University Medical Centre, Netherland

A Next Generation Sequencing (NGS) experiment produces an amount of data that is two orders of magnitude larger than a typical microarray experiment. Moreover, NGS data consist of overdispersed counts which are best described by the class of Generalized Linear Mixed Models. GLMMs are computationally more demanding than the linear models that are typically used for microarrays. This is a real problem as we are working with such large data sets. We are currently involved with a project aimed to identify the effect of prenatal famine exposure on DNA methylation. It is becoming increasingly

clear that DNA methylation, i.e. the epigenetic information layer, is where nature meets nurture. The best-characterized epigenetic marks are the methylation of cytosines in cytosine-guanine (CpG) dinucleotides. We have data on roughly 3 million CpGs in 24 same sex sibling pairs. In every family, one sibling has been exposed in utero to the famine. A particular challenge is to account for the family structure in an appropriate way. We found that Generalized Estimating Equations allow us to meet that challenge in a computationally efficient way. We do hesitate to use the standard asymptotics for the computation of p values. By permuting the sibling pairs and applying the scheme of Westfall and Young, we can get correct p values and at the same time adjust for multiple testing.

Posters

Short hairpin RNAs modeling

Cristina Della Beffa

Helmholtz Centre for Infection Research, Braunschweig, Germany

Subject: A non-coding RNA molecule with a hairpin shape, characterized by a silencing function on gene expression, is called short hairpin RNA (shRNA). This has been widely used as a novel effective tool for functional genomics studies, displaying a great potential in treating human diseases.

Aim: A project aiming at detecting genes that support liver regeneration after damage, recently started and a deep analysis of data sets of shRNA read counts to determine the cause of variations between experiments in different conditions was made with the purpose of defining the optimal statistics to select significantly regulated shRNAs, attached to genes that are in turn regulated, enhancing or repressing the proliferation of liver cells. The liver has a peculiar property: its cells regenerate fastly, after damage. Finding the genes that help liver reconstruction, can lead to the development of new drugs for the treatment of patients with chronic liver damages.

Methods: Three data sets of shRNAs read counts (generated with Solexa-Illumina deep sequencing platform) with three or four replicates in each condition were analysed, after some suitable transformations of each data set, to avoid computational problems with the reads containing zeros, respectively: adding a pseudo-count (Laplace correction); substituting them with the mean value of the remaining non-zero elements in the same condition; treating them as missing values. Three statistical methods have been used to detect regulated shRNAs and then the intersection of the results obtained with these methods was considered.

Identification of context specific TF interactions with GEMULA: Gene Expression Modeling Using LAsso

G. Geeven[1], *R. E. van Kesteren*[2], *A. B. Smit*[2] *and M. C. M. de Gunst*[1]

[1] Department of Mathematics, Faculty of Sciences, VU University, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands; [2] Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands

Dynamic gene regulatory networks, in which edges between nodes correspond to interactions between transcriptional regulators and their target genes, describe the coordinated spatiotemporal expression of genes. Especially in higher organisms, context specific combinatorial regulation by transcription factors (TFs) is believed to determine cellular states and fates. TF-target gene interactions can be studied using high-throughput techniques such as ChIP-chip or ChIP-Seq. These experiments are time and cost intensive, and further limited by for instance availability of high affinity TF anti-bodies. Hence, there is a practical need for methods that can predict TF-TF and TF-target gene interactions in silico, i.e. from gene expression and DNA sequence data alone. Regression models, in which predictor variables represent in silico predicted TF-DNA interactions, can be used to identify associations between TF binding and observed variation in gene expression, but there are challenges from a statistical point of view. We propose GEMULA, a novel approach based on linear models to predict TF-gene expression associations and TF-TF interactions from experimental data. GEMULA is based on linear models, fast and considers a wide range of biologically plausible models that describe gene expression data as a function of predicted TF binding to gene promoters. We show that models inferred with GEMULA are able to explain roughly 70% of observed variation in gene expression in the yeast heat shock response. The functional relevance of the inferred TF-TF interactions in these models are validated by different sources of independent experimental evidence. We also have applied GEMULA to an in vitro model of neuronal outgrowth. Our findings confirm existing knowledge on gene regulatory interactions underlying neuronal outgrowth, but importantly also generate new insights into the temporal dynamics of this gene regulatory network that can now be addressed experimentally.

Statistical approach to absolute protein quantification

Sarah Gerster, Peter Bühlmann

Seminar für Statistik, ETHZ

A major goal in proteomics is the comprehensive and accurate description of a proteome. Proteomics provides additional insights into biological systems that cannot be provided by genomic or transcriptomic approaches. In particular, proteomics holds great promise for the identification of biomarkers capable of accurately predicting disease already at a very early stage. The method of choice for the analysis of complex protein mixtures is shotgun proteomics. Proteins are identified and quantified based on experimentally measured peptides. While several probabilistic models exist for the identification of proteins, label-free quantification is often done in a deterministic way. We propose a statistical approach to protein quantification with three main advantages. (i) Peptide intensities are modeled as random quantities, allowing to account for the uncertainty of these measurements. (ii) Our Markovian-type model for bipartite graphs ensures transparent propagation of the uncertainties and reproducible results. (iii) The problem of peptides mapping to several protein sequences (often neglected in other models) is addressed automatically according to our statistical model. The performance of our model is shown on two synthetic control datasets and compared to the results of two common approaches for protein quantification.

Study of PROX1 as a modulator of the Wnt pathway in colon cancer cells using ChIP-seq technology

Nawal Houhou

SIB, Swiss Institut of Bioinformatics, Lausanne

Abnormal Wnt pathway signaling are the cause of various human diseases including cancer. In particular Wnt pathway mutations cause around 90% of the colorectal cancers (CRC). The Wnt/beta-CATENIN pathway signaling regulates the beta-CATENIN protein level. The beta-CATENIN protein interacts with TCF4 to contol some target genes transcription. In the other hand, PROX1 has an important role in tumor progression in colorectal cancer and is a potential target for the development of colon cancer therapy.

PROX1, TCF4 and betaCAT being one of the core components of CRC, we investigate the possible link between these proteins. Our strategy is to identify their binding regions which will lead to the corresponding target genes as well as consensus motifs. We use the Chromatin Immuno-Precipitation sequencing (ChIP-seq), a method which determines the location of DNA binding sites by direct sequencing of DNA fragments. Finally, a ChIP-seq analysis and a gene expression profiling show that PROX1 interact with betaCATENIN and TCF4.

Next generation sequencing and its application in epigenetics research – a bioinformatics perspective

Markus Jaritz

Research Institute of Molecular Pathology, Vienna

The results of next generation sequencing techniques such as Illuminas Genome Analyzer platform pose a challenge to both bioinformaticians and experimental biologists. In the context of epigenetics research, we face the analysis of genome wide ChIP-Seq data of transcription factor binding sites and histone marks, as well as RNA-Seq expression information. We will discuss a selection of high throughput data analysis approaches that thrive to answer the most pressing biological questions, such as discovery of binding site motifs, chromatin state dependency of binding sites and genome wide mRNA expression measurements.

GemSIM – Generic, error model-based SIMulator of next-generation sequencing

Kerensa McElroy[a], *Fabio Luciani*[b], *Torsten Thomas*[a]

[a] Centre for Marine Bio-Innovation and School of Biotechnology and Biomolecular Sciences, UNSW, Sydney, NSW Australia, 2052; [b] Inflammatory Diseases Research Unit, School of Medical Sciences, UNSW, Sydney, NSW Australia 2052.

Next-generation sequencing (NGS) has unprecedented potential for assessing genetic di-

versity, however extracting true variants from errors is challenging due to high NGS error rates, multiple sequencing platforms with varied error profiles, and an every increasing variety of downstream analysis choices. While simulation can facilitate analysis, existing simulators are limited by simplistic error-models, unrealistic quality score information, or restricted platform applicability. GemSIM, or General Error-Model based SIMulator, is a next-generation sequencing simulator capable of generating single or paired-end reads for any sequencing technology compatible with the generic formats SAM and FASTQ (including Illumina and Roche-454). By creating and using empirically derived error models and quality score distributions, GemSIM realistically emulates individual sequencing runs and/or technologies. GemSIM draws reads from either a single genome or a haplotype set, facilitating simulation of either individual or population level sequencing projects. Here, we demonstrate GemSIMs value for next-generation sequencing projects, by simulating reads from a set of known, related bacterial haplotypes and optimising a parameter for the popular SNP-calling program VarScan. Reads simulated using error models derived from Illumina paired-end reads required different SNP calling parameters to those simulated using Roche-454 derived models, demonstrating the need for simulation when designing and analysing NGS projects. Project page: `http://sourceforge.net/projects/gemsim/`.

Changepoint models with dependence within segments for detecting epigenetic patterns

Andrea Riebler

University of Zurich

Over the last years, high-throughput sequencing has become very popular. Since it overcomes many limitations of microarrays, e.g. non-specific background noise, it is expected to become the new standard for many applications including the analysis of epigenetic patterns. We would like to propose a novel Bayesian multiple changepoint model for detecting epigenetic patterns along the genome. Multiple changepoint models are designed to detect multiple changepoints in ordered data, whereby the number of changepoints is unknown. They have clear advantages, compared to other segmentation methods, such as circular binary segmentation or hidden Markov models. However, usually independence of data within segments is assumed, which is certainly questionable in data on epigenetic patterns. We will illustrate a new approach of Wyse et al.

(2010, http://adsabs.harvard.edu/abs/2010arXiv1011.5038W) who proposed the use of Gaussian Markov random field (GMRF) models within segments, so that spatial dependence along the genome could be easily incorporated. The applicability of this new technique will be investigated for MeDIP-seq data to detect DNA methylation along the genome.

Edge detection in graphical models

Maya Shevlyakova

EPF Lausanne

Studying partial correlations between variables is based on the concentration (inverse covariance) matrix. In genomic applications we would like to uncover such associations based on limited data for large sets of variables. We use the Kullback-Leibler divergence to examine the difference between a situation with no partial correlations and one with a few non-zero partial correlations. We get exact values for the divergence for one, two and three elements. We show that the divergence does not depend on the number of variables and grows with the sample size. In an asymptotic study in which partial correlation is of order $1/\sqrt{n}$ for the sample size of n, it is impossible to detect it. Multiple partial correlations are easier to detect.

Participants

Khalid Abnaof University of Bonn
Christian Ahrens University of Zurich
Simon Anders EMBL Heidelberg
Claudia Angelini IAC-CNR
Viktoria Bastic Schmid Nestlé Research Center
Niko Beerenwinkel ETH Zurich and Swiss Institute of Bioinformatics, Basel
Daniela Beisser University of Würzburg
David Bernasconi University of Lausanne
Michael Brudno University of Toronto
Peter Bühlmann ETH Zurich
Andreas Buness Novartis
Peter Butzhammer University of Regensburg
Benilton Carvalho University of Cambridge
Diego Colombo ETH Zurich
Jerome Dauvillier Merck Serono
Italia De Feis CNR
Cristina Della Beffa Helmholtz Centre for Infection Research, Braunschweig, Germany
Mauro Delorenzi Swiss Institute of Experimental Cancer Research and Swiss Institute of Bioinformatics, Lausanne
Julia Di Iulio University of Lausanne
Heide Fier University of Bonn
Bernd Fischer EMBL Heidelberg
Martina Fischer German Cancer Research Centre, Heidelberg
Brooke Fridley Mayo Clinic College of Medicine
Geert Geeven University of Amsterdam
Steven Geinitz University of Zurich
Sarah Gerster ETH Zurich
Moritz Gerstung ETH Zurich
Darlene Goldstein EPF Lausanne

Chris Greenman University of East Anglia, Norwich
Wilhelm Gruissem ETH Zurich
Kasper Hansen Johns Hopkins university, Baltimore
Nawal Houhou University of Lausanne, CHUV
Wolfgang Huber EMBL Heidelberg
Iuliana Ionita-Laza Columbia University
Markus Jaritz Research Institute of Molecular Pathology, Vienna
Rebecka Joernsten Chalmers University
Vindi Jurinovic LMU Munich
Ivo Kwee IOSI Laboratory of Experimental Oncology, Bellinzona
Alix Leboucq EPF Lausanne
John Marioni EMBL-EBI
Kerensa McElroy University New South Wales
Katharina Meyer University of Regensburg
George Michailidis University of Michigan
Tom Michoel University of Freiburg
Eugenia Migliavacca University of Lausanne
Edoardo Missiaglia Lausanne
Pejman Mohammadi ETH Zurich
Sach Mukherjee University of Warwick
Marc Noguera-Julian IrsiCaixa – Retrovirology lab
Jan Oosting LUMC, Leiden
Irina Ostrovnaya Memorial Sloan-Kettering Cancer Center
Gregoire Pau EMBL Heidelberg
Susana Pérez IrsiCaixa
Wolfgang Raffelsberger IGBMC
Ben Raphael Brown University
Hubert Rehrauer University of Zurich
Bernhard Renard Robert Koch-Institut, Berlin
Nora Rieber DKFZ Heidelberg
Andrea Riebler University of Zurich
Mark Robinson Walter and Eliza Hall Institute
Simona Rossi SIB, University of Lausanne
Volker Roth University of Basel
Ingo Ruczinski Johns Hopkins University
Thomas Sakoparnig ETH Zurich
Alex Sanchez University of Barcelona
Frédéric Schütz Swiss Institute of Bioinformatics

Maya Shevlyakova EPF Lausanne
Katarzyna Sikora EPF Lausanne
Sandeep Singhal Université Libre de Bruxelles
Daniel Stekhoven ETH Zurich
Margaret Taub Johns Hopkins University
Amalio Telenti University Hospital CHUV, Lausanne
Achim Tresch LMU Munich
Bie Verbist University of Gent
Osvaldo Zagordi ETH Zurich
Nadine Zangger University Hospital CHUV, Lausanne
Stefan Zoller ETH Zurich
Simon de Bernard AltraBio
Erik van Zwet Leiden University Medical Centre
Peter von Rohr Nebion Spin-off ETH Zurich