Banff International Research Station

for Mathematical Innovation and Discovery

Statistical Genomics in Biomedical Research

BIRS Workshop 10w5076
July 18–23, 2010

ORGANIZERS

Jennifer Bryan, Department of Statistics, University of British Columbia
Sandrine Dudoit, Division of Biostatistics and Department of Statistics, University of California, Berkeley
Jane Fridlyand, Genentech
Darlene R. Goldstein, Institut de mathématiques, École Polytechnique Fédérale de Lausanne, Switzerland
Sunduz Keles, Department of Statistics and of Biostatistics and Medical Informatics, University of Wisconsin, Madison
Katherine S. Pollard, Gladstone Institutes, University of California, San Francisco

MEALS

- **Breakfast*** (Buffet): 07:00–09:30, Sally Borden Building, Monday–Friday
- **Lunch*** (Buffet): 11:30–13:30, Sally Borden Building, Monday–Friday
- **Dinner*** (Buffet): 17:30–19:30, Sally Borden Building, Sunday–Thursday
- **Coffee Break**: As per daily schedule, 2nd floor lounge, Corbett Hall

* **N.B.** *Please remember to scan your meal card at the host/hostess station in the dining room for each meal.*

MEETING ROOMS

All lectures are held in 159 Max Bell Building.
The Max Bell Building is accessible by the walkway on the 2nd floor of Corbett Hall.
LCD projectors, overhead projectors, and blackboards are available for presentations.

N.B. *Please note that the meeting space designated for BIRS is the lower level of the Max Bell Building, Rooms 155–159. Please respect that all other space has been contracted to other Banff Centre guests and that food and beverage in those areas are intended for these guests only.*

SCHEDULE

All lectures are **30-minute** long, followed by a 5-minute question period.

Day 0 **Sunday, July 18**

16:00	Check-in, Front Desk, Professional Development Centre – open 24 hours
17:30–19:30	Dinner
20:00	Informal gathering, 2nd floor lounge, Corbett Hall

Day 1 **Monday, July 19**

07:00–08:45	Breakfast
	Population and Quantitative Genomics (Chair, Katie Pollard)
08:45–09:00	Introduction and welcome to BIRS by BIRS Station Manager, 159 Max Bell Building
09:00–09:35	Jonathan Pritchard, *Expression QTL mapping with RNA-Seq*
09:40–10:15	Jeff Wall, *Estimating human demographic parameters from sequence polymorphism data*
10:15–10:45	Coffee Break
10:45–11:20	Yoav Gilad, *Comparative genomic in primates using next-generation sequencing*
11:25–12:00	John Ngai, *Regulation of olfactory stem cell renewal and differentiation: Insights from transcriptome profiling*
12:00–13:00	Lunch
13:00–13:45	Guided tour of the Banff Centre; meet in the 2nd floor lounge, Corbett Hall
13:45–14:00	Group photo; meet in the 2nd floor lounge, Corbett Hall
	Transcriptional Genomics (Chair, Alisha Holloway)
14:00–14:35	Jason Lieb, *Genome-wide measurement of transcription factor binding dynamics by competition ChIP*
14:40–15:15	Elodie Portales-Casamar, *Deciphering regulatory networks by transcription factor binding site analysis*
15:15–15:45	Coffee Break
15:45–16:20	Sunduz Keles, *MOSAiCS: Model-based One & Two Sample Analysis and Inference for ChIP-Seq data: from multi-reads to background adjustment to peak calling*
16:25–17:00	Ting Wang, *Mapping human DNA methylome with MeDIP-Seq and MRE-Seq*
17:30–19:30	Dinner

Day 2 **Tuesday, July 20**

07:00–09:00	Breakfast
	High-Throughput Sequencing (Chair, Laurent Jacob)
09:00–09:35	James Bullard, *An overview of PacBio data and applications*
09:40–10:15	Margaret Taub, *Detection of single-nucleotide variants with high throughput sequencing*
10:15–10:45	Coffee Break
10:45–11:20	Kasper Hansen, *TBA*
11:25–12:00	Wolfgang Huber, *Differential expression analysis for sequence count data*
12:00–14:00	Lunch
	High-Throughput Biological Assays (Chair, James Bullard)
14:00–14:35	Laurent Jacob, *More power in differentially expressed pathway identification using known gene networks*
14:40–15:15	Jean-Philippe Vert, *Including prior knowledge in shrinkage classifiers for genomic data*
15:15–15:45	Coffee Break
15:45–16:20	Pierre Neuvial, *Targeted maximum likelihood estimation of the relationship between copy number and gene expression in cancer studies*
16:25–17:00	Robert Scharpf, *A multilevel model to address batch effects in copy number estimation for high-throughput SNP arrays*
17:30–19:30	Dinner
19:30–21:30	**Poster Session**

Day 3 **Wednesday, July 21**

07:00–09:00	Breakfast
	High-Throughput Biological Assays (Chair, Darlene Goldstein)
09:00–09:35	Jared Roach, *Pedigree genome sequencing*
09:40–10:15	Mark Segal, *Clustering with exclusion zones: Genomic applications*
10:15–10:45	Coffee Break
10:45–11:20	Ingo Ruczinski, *SNP association studies with case-parent trios*
11:25–12:00	Houston Gilbert, *Statistical applications in the analysis of reverse-phase protein microarray data: Results from a cross-platform evaluation study*
12:00–13:30	Lunch
	Free Afternoon/Hike
17:30–19:30	Dinner

Day 4 **Thursday, July 22**

07:00–09:00 Breakfast

From the Bench to the Clinic (Chair, Jane Fridlyand)

09:00–09:35 Adam Olshen, *Something old, something new*

09:40–10:15 Mauro Delorenzi, *Translational studies for predictive and prognostic biomarkers in colon cancer*

10:15–10:45 Coffee Break

10:45–11:20 Pete Haverty, *The mutation spectrum revealed by paired genome sequences from a lung cancer patient*

11:25–12:00 Donald Geman, *Rank statistics in biomedical research*

12:00–14:00 Lunch

Predictive Diagnostics and Designing Clinical Trials (Chair, Adam Olshen)

14:00–14:35 Jane Fridlyand, *Designing proof of concept trials in oncology: Speed, cost and trial success*

14:40–15:15 Ru-Fang Yeh, *Statistical challenges in the development of predictive biomarkers*

15:15–15:45 Coffee Break

15:45–16:20 Venkat Seshan, *Two stage designs for gene-disease association studies*

17:30–19:30 Dinner

Day 5 **Friday, July 23**

07:00–09:30 Breakfast

Free Morning/Hike

11:30–13:30 Lunch

Checkout by 12:00 noon

N.B. *Participants for 5-day workshops are welcome to use the BIRS facilities (2nd floor lounge of Corbett Hall, Max Bell Building meeting rooms, reading room) until 15:00 on Friday, although they are required to checkout of the guest rooms by 12:00 noon.*

Statistical Genomics in Biomedical Research

BIRS Workshop 10w5076
July 18–23, 2010

ABSTRACTS: INVITED LECTURES
(in alphabetic order by speaker surname)

James Bullard
Pacific Biosciences

An overview of PacBio data and applications

The PacBio RS is scheduled for full commercial release later this year. This third generation sequencing platform has enormous potential. The focus of this talk is to describe the types of data available to analysts, the open source software that we are producing, and the repositories where example data can be obtained. I will try to highlight a number of projects where Pacific Biosciences sequencing data would initially be well-suited.

Mauro Delorenzi
Department of Research, Lausanne University Hospital, and Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

Translational studies for predictive and prognostic biomarkers in colon cancer

In colon cancer treatment, chemotherapy is frequently given to patients that are cured by surgery alone and do not benefit from it while other patients to whom chemotherapy is not given will suffer from a distant relapse of the primary tumor. It would be useful, to have better predictive models that would allow better treatment decision making. New profiling technologies might help discovering useful biomarkers.

A translational research project is based on the colon cancer tissue bank of the PETACC 3 randomized clinical trial, which tested improvement of the standard of care 5-FU chemotherapy by the addition of irinotecan. Prognostic models for relapse-free, overall survival and survival after relapse with clinical variables and with a set of molecular markers are being assessed, including TNM stage, tumor site and grade, microsatellite instability, *KRAS* and *BRAF* mutation status and IHC for a panel of proteins. Association of outcome with covariates was observed for relapse with microsatellite instability and SMAD4 protein expression by tumor cells, in addition to well known association with TNM stage; and for survival after relapse with BRAF and tumor site. KRAS mutation status did now show any strong prognostic effect in this population.

Gene expression profiles are being generated from FFPE material with the aim of studying the molecular heterogeneity of the disease and the possibility to define subtypes that are associated with different characteristics of survival or response to therapy. Preliminary results seem to suggest that colon cancer heterogeneity might be rather unstructured-continuous than dominated by a small number of major subclasses as in breast cancer, or that data are too noisy to reliably identify them.

In a second study, we search to identify markers associated with tumor progression in the case of metastasis treated with specific anti-EGFR drugs. In particular we tested the effects of KRAS, BRAF, NRAS and PIK3CA mutations. Data confirm roles for KRAS and BRAF, suggest that in NRAS and PIK3CA mutations might be important too, and that different mutations in the same gene might not be equivalent.

Jane Fridlyand
Genentech
Designing proof of concept trials in oncology: Speed, cost and trial success

Abstract.

Donald Geman
Department of Applied Mathematics and Statistics, Johns Hopkins University
Rank statistics in biomedical research

I will talk about several projects in expression-based biomarker discovery and pathway regulation, mainly focused on cancer. The driving application is translational medicine. I will argue that rank-based statistics can account for combinatorial interactions among genes and gene products; accommodate variations in data normalization and limited sample sizes; and avoid the "black box" representations and decision rules generated by standard methods in computational learning.

Yoav Gilad
Department of Human Genetics, The University of Chicago
Comparative genomic in primates using next-generation sequencing

Progress in evolutionary genomics is tightly coupled with the development of new technologies to collect high-throughput data. The availability of next-generation sequencing technologies has the potential to revolutionize genomic research and enable us to focus on a large number of outstanding questions that previously could not be addressed effectively. In the context of comparatives genomic studies in primates, new sequencing technologies allow us to collect high resolution inter-individual and inter-species variation data from multiple dimensions of the regulatory landscape. We use these data to better understand the contribution of different regulatory mechanisms to overall inter-species differences in gene regulation. These data also allow us to identify individual genes and entire pathways whose regulation evolves under natural selection in primates. These observations have the potential to help us find functional genetic variation in humans. For example, we found that genes previously associated with diseases that affect specific tissues are enriched for genes whose regulation evolves under stabilizing selection in the same tissues.

Houston Gilbert
Genentech
Statistical applications in the analysis of reverse-phase protein microarray data: Results from a cross-platform evaluation study

Reverse-phase protein microarrays (RPPMA) allow for the simultaneous detection of a single protein in complex analyte mixtures, such as those obtained from cell tissue culture or clinical sample protein lysate. To gain a better understanding of the RPPMA arena, we evaluated three fee-for-service providers of this technology. Practical, statistical and biological results from the evaluation study have informed our own strategies for moving forward with RPPMA technology in research and development programs. The

evaluation study has also highlighted areas for each of the companies to improve upon their own platforms.

Joint work with Maureen Wong, Zachary Boyd, Jenny Wu, Sree Ranjani Ramani, Yibing Yan, Mark Lackner, Lisa Belmont, and Lino Gonzalez.

Pete Haverty
Genentech
The mutation spectrum revealed by paired genome sequences from a lung cancer patient

Although previous studies have identified important common somatic mutations in lung cancers, they have primarily focused on a limited set of genes and have thus provided a constrained view of the mutational spectrum. Here we present the complete sequences of a primary lung tumour (60 coverage) and adjacent normal tissue (46). Comparing the two genomes, we identify a wide variety of somatic variations, including ¿50,000 high-confidence single nucleotide variants. We validated 530 somatic single nucleotide variants in this tumour, including one in the KRAS proto-oncogene and 391 others in coding regions, as well as 43 large-scale structural variations. These constitute a large set of new somatic mutations and yield an estimated 17.7 per megabase genome-wide somatic mutation rate. Notably, we observe a distinct pattern of selection against mutations within expressed genes compared to non-expressed genes and in promoter regions up to 5?kilobases upstream of all protein-coding genes. Furthermore, we observe a higher rate of amino acid-changing mutations in kinase genes. We present a comprehensive view of somatic alterations in a single lung tumour, and provide the first evidence, to our knowledge, of distinct selective pressures present within the tumour environment.

Wolfgang Huber
European Molecular Biology Laboratory, Heidelberg, Germany
Differential expression analysis for sequence count data

High-throughput nucleotide sequencing provides quantitative readouts in assays for RNA expression (RNA-Seq), protein-DNA binding (ChIP-Seq) or cell counting (barcode sequencing). Statistical inference of differential signal in such data requires estimation of their variability throughout the dynamic range. When the number of replicates is small, error modeling is needed to achieve statistical power. We propose an error model that uses the negative binomial distribution, with variance and mean linked by local regression, to model the null distribution of the count data. The method controls type-I error and provides good detection power. A free open-source R software package, DESeq, is available from the Bioconductor project

Laurent Jacob
Department of Statistics, UC Berkeley
More power in differentially expressed pathway identification using known gene networks

We cast the problem of identifying sets of genes which are differentially expressed between two clinical groups as a multivariate two-sample test. Under the assumption that the shift of expression is coherent with a known network structure, we show that integrating this structure in the test statistic leads to more powerful tests. We also study systematic testing of all the sub-networks of a large network for de novo pathway identification. We illustrate the behaviour of our approach on synthetic data, and on a breast cancer hormone therapy resistance expression dataset.

Sunduz Keles
Department of Statistics and of Biostatistics and Medical Informatics, University of Wisconsin, Madison
MOSAiCS: Model-based One & Two Sample Analysis and Inference for ChIP-Seq data: from multi-reads to background adjustment to peak calling

Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) has revolutionized experiments for genome-wide profiling of DNA-binding proteins, histone modifications, and nucleosome occupancy. As the cost of sequencing is decreasing, many researchers are switching from microarray-based technologies (ChIP-chip) to ChIP-Seq for genome-wide study of transcriptional regulation. Despite its increasing and well-deserved popularity, there is little work that investigates and accounts for sources of biases in the ChIP-Seq technology. These biases typically arise from both the standard pre-processing protocol and the underlying DNA sequence of the generated data. In this talk, I discuss various statistical aspects of ChIP-Seq data analysis including handling of multi-reads and developing background models that adjust for apparent sources of biases due to ChIP-Seq experimental protocol.

In particular, we focus on data from a naked DNA sequencing experiment, which sequences non-cross-linked DNA after deproteinizing and shearing, to understand factors affecting background distribution of data generated in a ChIP-Seq experiment. We propose a background model that accounts for the observed sources of biases such as mappability and GC content. Consequently, we develop a flexible mixture modeling approach named MOSAiCS for detecting peaks in ChIP-Seq data. This model incorporates the background component derived from naked DNA experiments and introduces a flexible model for the actual signal component. We illustrate that this model fits actual ChIP-Seq data very well. An important practical advantage of this framework is that one-sample analysis of ChIP-Seq data with MOSAiCS performs as well as the two-sample ChIP-Seq data analysis that utilizes sequenced naked DNA as control. We further develop an extension of this model for two-sample ChIP-Seq data analysis with Input DNA control and discuss the utilization of multi-reads and its downstream effects.

Mark R. Lackner
Genentech
Identification of predictive biomarkers for a selective PI3K inhibitor

The class I phosphatidylinositol 3kinase (PI3K) is activated in a wide variety of human malignancies and inhibitors targeting the PI3K pathway hold great promise in the treatment and management of cancer. Successful development of such inhibitors will be enhanced by identification of responsive patients through the use of predictive biomarkers. We used cell lines and xenografted tumors to evaluate a collection of putative biomarkers predictive of response to the selective inhibitor GDC-0941 in breast cancer. We found high activity of the PI3K inhibitor GDC-0941 in luminal and HER2 amplified models, and that PIK3CA mutations and HER2 amplification are highly specific biomarkers of response to this agent. We found that a number of models that do not harbor these alterations also showed sensitivity, suggesting a need for additional diagnostic markers. Gene expression studies identified a collection of genes whose expression was associated with in vitro sensitivity to GDC-0941, and expression of a subset of these genes was found to be intimately linked to signaling through the pathway. Clinical implications and strategies for biomarker validation will also be considered.

Jason D. Lieb
Department of Biology and Carolina Center for the Genome Sciences, University of North Carolina, Chapel Hill
Genome-wide measurement of transcription factor binding dynamics by competition ChIP

Chromatin immunoprecipitation (ChIP) is an incredibly powerful tool for interrogating the function of individual DNA binding proteins, and the regulatory architecture of the genome. However, the assay

is inherently blind to kinetics. To measure transcription factor binding dynamics, we have built a system to perform genome-wide competition chip with a sequence specific transcription factor, RAP1. RAP1 turnover is slow at the promoters of the highly transcribed ribosomal protein genes and faster at promoters of infrequently transcribed genes and at telomeres. Relative to static RAP1 occupancy, Rap1 binding dynamics correlate more strongly with many genomic features, including transcription, nucleosome occupancy, nucleosome acetylation, and nucleosome turnover. The data suggests that competition between RAP1 and nucleosomes regulates the transcriptional outcome of Rap1 binding, with longer RAP1 residency leads to more transcription. To test this hypothesis we performed a competition chip experiment under oxidative stress conditions and observe the predicted changes in Rap1 binding dynamics, even at sites with Rap1 occupancy that appears unchanged by conventional ChIP.

Joint work with Colin R. Lickwar, Florian Mueller, Sean Hanlon, and James G. McNally.

Pierre Neuvial
Department of Statistics, UC Berkeley

Targeted maximum likelihood estimation of the relationship between copy number and gene expression in cancer studies

Looking for genes whose DNA copy number is "associated" with their expression level in a cancer study can help pinpoint candidates implied in the disease and improve our understanding of its molecular bases. DNA methylation is an important player to account for in this setting, as it can down-regulate gene expression.

We have developed a method based on Targeted Maximum Likelihood to quantify the relationship between copy number and expression, accounting for DNA methylation. I will explain the method and some of its statistical properties. I will also show preliminary results on a simulation study and on a real data set from the Cancer Genome Atlas (TCGA).

Joint work with Antoine Chambaz.

John Ngai
Department of Plant and Microbial Biology, UC Berkeley

Regulation of olfactory stem cell renewal and differentiation: Insights from transcriptome profiling

Tissue regeneration is a complex process that requires the coordination of stem cell proliferation and differentiation to maintain or repair the structure. The olfactory epithelium (OE) is a sensory neuroepithelium whose constituent cell types – including the olfactory sensory neurons – are continuously replaced during the lifetime of the animal. Following severe injury that results in the loss of mature cell types, the OE is rebuilt by the proliferation and differentiation of adult tissue stem cells. The regenerative capacity and limited number of cell types make the OE an excellent model for investigating stem cell regulation in vivo. Previous studies have identified the horizontal basal cell (HBC) as the multipotent neural stem cell of the OE; the molecules and pathways regulating this adult tissue stem cell are unknown, however. Using whole genome expression profiling of FACS-purified HBCs, we characterized the mRNA and miRNA transcriptomes of HBCs under conditions of quiescence and proliferation/differentiation. Through these studies we identified groups of genes associated with different phases of the HBC life cycle. In addition, we found that p63, a member of the p53 tumor suppressor gene family, is highly enriched in quiescent HBCs. p63 is a key regulator of stem cell self-renewal and differentiation in all stratified epithelia investigated to date. Conditional inactivation of the p63 gene in HBCs results in the appearance of mature cells but loss of HBCs following regeneration. These results demonstrate a critical role of p63 in olfactory stem cell renewal and differentiation, and provide an entrée toward elucidating the downstream targets and

interaction partners of this transcription factor. Our studies provide the first molecular insights into the genetic network regulating stem cell dynamics in the OE and reveal an unexpected parallel between stem cell regulation in this sensory neuroepithelium and other epithelial tissues.

Joint work with Russell Fletcher, Melanie Prasol, Jose Estrada, Yoon Gi Choi, and Karen Vranizan.

Adam B. Olshen
Department of Epidemiology and Biostatistics and HDFC Cancer Center, UCSF
Something old, something new

I will discuss two projects involving high throughput data. One is more mature and concerns distinguishing primary tumors from metastases utilizing copy number data. The methodology will be demonstrated on a lung cancer data set. The second is a work-in-progress involving methylation sequencing data. I will address integrating multiple types of such data and methods for estimating copy number from it.

Elodie Portales-Casamar
Centre for Molecular Medicine and Therapeutics, University of British Columbia
Deciphering regulatory networks by transcription factor binding site analysis

Regulation of gene expression can happen at multiple levels, including, but not limited to, chromatin modifications, initiation of transcription at gene promoters, alternative splicing and stability of RNA, protein modifications. The binding of transcription factors (TFs) to DNA sequences near or within genes is one of the primary mechanisms directing gene transcription. Understanding the interplay between TFs and their target genes is key to deciphering cellular regulatory networks that generate diverse types of cells and tissues within an organism. Many computational approaches to TF binding site analysis have been adopted. Sets of known binding sequences are compared to construct TF binding models. Such necessary information is collected and disseminated through community-driven resources like PAZAR and JASPAR, but the compiled data remains too sparse to cover the full spectrum of DNA-binding proteins. Genome-wide chromatin immunoprecipitation techniques (e.g. ChIP-Seq) are now providing larger data collections that allow for more accurate models and increase the quality of genome annotation. Such methods enable researchers to decipher entire regulatory networks in specific cellular contexts as, for instance, the upregulation of detoxification systems by the Nrf2 TF in cells exposed to stress. The presentation will provide an overview of TF binding site prediction methods using examples drawn from online resources and tools as well as a direct application of these methods for the Nrf2 ChIP-Seq data analysis.

Jonathan Pritchard
Department of Human Genetics, The University of Chicago
Expression QTL mapping with RNA-Seq

An important challenge of the post-genomic era is to make sense of how genome sequences control gene regulation. In this talk I will discuss work that we are doing using expression- and splicing-QTL (quantitative trait loci) in human lymphoblastoid cells as a model system for understanding how genetic variation can modify gene regulation. I will describe the application of next-generation sequencing to measuring gene expression and splicing patterns and our work on trying to understand the mechanisms of action of eQTL SNPs.

Jared Roach
Institute for Systems Biology
Pedigree genome sequencing

Full-genome sequences of a pedigree with p individuals can be represented as a series of genotype vectors. Consider a single chromosome with n positions. A genotype, $g_{i,p}$, is an observation of two alleles at a position i for individual p. For example, $g_{23145140,3}$ may be $\{A, C\}$. A genotype vector, $V_i = \{g_{i,1}, g_{i,2}, g_{i,3}, \ldots, g_{i,p}\}$, is an ordered list of genotypes for all individuals in the pedigree. The series of genotype vectors for a chromosome is thus $\{V_1, V_2, V_3, \ldots, V_n\}$. Binary inheritance vectors represent the not-directly-observed flow of alleles through the pedigree, and are parallel in structure to genotype vectors. These series of vectors can be regarded as emissions from Hidden Markov Models (HMMs) and illuminate underlying genetic features. We analyzed the whole-genome sequences of a family of four. HMMs enabled the precise identification of recombination sites and 70% of the sequencing errors. These analyses permit matching inheritance states and inheritance modes, and thus disease-gene identification.

Ingo Ruczinski
Department of Biostatistics, Johns Hopkins University
SNP association studies with case-parent trios

While most SNP association studies are population based, family based designs also have some very attractive features. Case-parent trio designs in particular allow for the assessment of de-novo copy number variants, parent-of-origin effects, and transmission distortion. We discuss and demonstrate those via a genome-wide and a candidate gene association study that employ case-parent trios. We also extend the logic regression methodology, originally developed for cohort and case-control studies, to detect SNP-SNP and SNP-environment interactions in studies of trios with affected probands, and derive an efficient algorithm to simulate case-parent trios where genetic risk is determined via epistatic interactions.

Robert Scharpf
Department of Biostatistics, Johns Hopkins University
A multilevel model to address batch effects in copy number estimation for high-throughput SNP arrays

Submicroscopic changes in chromosomal DNA copy number dosage are common and have been implicated in many heritable diseases and cancers. Recent high-throughput technologies have a resolution that permits the detection of segmental changes in DNA copy number that span thousands of basepairs across the genome. Genome-wide association studies may simultaneously screen for copy number-phenotype and SNP-phenotype associations as part of the analytic strategy. However, genome-wide array analyses are particularly susceptible to batch effects as the logistics of preparing DNA and processing thousands of arrays often involves multiple laboratories and technicians, or changes over calendar time to the reagents and laboratory equipment. Failure to adjust for batch effects can lead to incorrect inference and requires inefficient post-hoc quality control procedures that exclude regions that are associated with batch. Our work extends previous model-based approaches for copy number estimation by explicitly modeling batch effects and using shrinkage to improve locus-specific estimates of copy number uncertainty. Key features of this approach include the use of diallelic genotype calls from experimental data to estimate batch- and locus-specific parameters of background and signal without the requirement of training data.

Mark Segal
Division of Biostatistics, UCSF
Clustering with exclusion zones: Genomic applications

Methods for formally evaluating the clustering of events in space or time, notably the scan statistic, have been richly developed and widely applied. In order to utilize the scan statistic and related approaches it is necessary to know the extent of the spatial or temporal domains wherein the events arise. Implicit in their usage is that these domains have no "holes" – hereafter *exclusion zones* – regions in which events *a*

priori cannot occur. However, in many contexts, this requirement is not met. When the exclusion zones are known it is straightforward to correct the scan statistic for their occurrence by simply adjusting the extent of the domain. Here, we tackle the more ambitious objective of formally evaluating clustering in the presence of *unknown* exclusion zones. By examining the behavior of *clumps* over the grid of putative cluster counts and lengths, we show that the existence of exclusion zones manifests as a characteristic signature. We exploit this patterning to develop an algorithm for estimating total exclusion zone extent, the parameter needed to correct scan statistic based inference. Performance of the algorithm is assessed via simulation study. We showcase applications to genomic settings for differing marker (event) types – *binding sites*, *housekeeping genes*, and *microRNAs* – wherein exclusion zones can arise through a variety of mechanisms. In several instances, dramatic changes to unadjusted inference that does not accommodate exclusions, are evidenced.

Joint work with Yuanyuan Xiao.

Venkat Seshan
Memorial Sloan-Kettering Cancer Center
Two stage designs for gene-disease association studies

Gene-disease association studies based on case-control designs may often be used to identify candidate SNPs (markers) conferring disease risk. If a large number of markers are studied, genotyping all markers on all samples is inefficient in resource utilization. Here, we propose an alternative two-stage method to identify disease-susceptibility markers. In the first stage all markers are evaluated on a fraction of the available subjects. The most promising markers are then evaluated on the remaining individuals in Stage 2. This approach can be cost effective since markers unlikely to be associated with the disease can be eliminated in the first stage.

This work was done in collaboration with Jaya Satagopan & Colin Begg.

Margaret Taub
Department of Biostatistics, Johns Hopkins University
Detection of single-nucleotide variants with high throughput sequencing

In this talk I'll discuss single-nucleotide variant detection with high throughput sequencing, including current practices and pitfalls. I will present results on one targeted re-sequencing dataset and some of the publicly available data from the 1000 genomes project with an aim toward understanding the impacts of technical and sequence-specific properties on accurate variant detection.

Jean-Philippe Vert
Mines ParisTech
Including prior knowledge in shrinkage classifiers for genomic data

Estimating predictive models from high-dimensional and structured genomic data, such as gene expression of CGH data, measured on a small number of samples is one of the most challenging statistical problems raised by current needs in post-genomics. Popular tools in statistics and machine learning to address this issue are shrinkage estimators, which minimize an empirical risk regularized by a penalty term, and which include for example support vector machines or the LASSO. In this talk I will discuss new penalty functions for shrinkage estimators, including generalizations of the LASSO which lead to particular sparsity patterns, and which can be seen as a way to include problem-specific prior information in the estimator. I will illustrate the approach by several examples such as the classification of gene expression

data using gene networks as prior knowledge, or the classification and detection of frequent breakpoints in CGH profiles.

Jeff Wall
Division of Biostatistics, UCSF
Estimating human demographic parameters from sequence polymorphism data

Population genetic data sets have the potential to inform us about our species' demographic history, but most existing methods are not suitable for genomic-scale data. We present a composite-likelihood framework for estimating demographic parameters such as split times and migration rates, and apply this method to analyze polymorphism data from different sub-Saharan African populations. We find evidence for population structure that likely predates the exodus of modern humans out of Africa, and discuss the relevance of this finding with regard to current theories of human evolution.

Ting Wang
Department of Genetics and Center for Genome Sciences, Washington University in St. Louis
Mapping human DNA methylome with MeDIP-Seq and MRE-Seq

We present two complementary approaches to detect methylated and unmethylated genomic DNA. The first, methyl DNA immunoprecipitation and sequencing (MeDIP-Seq), uses antibody-based immunoprecipitation of 5-methylcytosine and sequencing to map the methylated fraction of the genome. In the second method, unmethylated CpG sites are identified by sequencing size-selected fragments from parallel DNA digestions with the methyl-sensitive restriction enzymes (MRE-Seq). We generated a genome-wide, high-resolution methylome map of human brain tissue, and a second map of human ES cell H1. These maps on average interrogate close to 90% of all CpGs (25 million of 28 million total) and 98% of CpG islands in the human genome, at the modest expense of relatively small amount specimen and a few lanes of Illumina flowcell.

We investigated the role of DNA methylation in gene bodies with these methylome maps. From high-resolution coverage of CpG islands, the majority of methylated CpG islands were revealed to be in intragenic and intergenic regions, while less than 3% of CpG islands in 5 promoters were methylated. The CpG islands in all three locations overlapped with RNA markers of transcription initiation, and unmethylated CpG islands also overlapped significantly with trimethylation of H3K4, a histone mark enriched at active promoters. The general and CpG-island-specific patterns of methylation are conserved in mouse tissues. These and other results support a major role for intragenic methylation in regulating cell context-specific alternative promoters in gene bodies.

Ru-Fang Yeh
Genentech
Statistical challenges in the development of predictive biomarkers

It has been increasingly important to incorporate diagnostics in the drug development process to improve response to treatment and help reduce adverse drug reaction. In this talk, I'll discuss statistical issues arising during the development of predictive biomarkers that aim to identify patients who will benefit from a treatment. Examples will be used to highlight statistical challenges in biomarker discovery and clinical applications, including threshold selection for continuous biomarkers and implmentation of complex predictors.

Statistical Genomics in Biomedical Research

BIRS Workshop 10w5076
July 18–23, 2010

ABSTRACTS: POSTERS
(in alphabetic order by presenter surname)

Presenter
Affiliation
Title

Abstract.

Mauro Delorenzi
Department of Research, Lausanne University Hospital, and Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

Molecular classes in CRC: Characterization of MSI by expression profiling in the translational study of the PETACC 3 - EORTC 40993 -SAKK 60-00 trial

Microsatellite instability (MSI) is the hallmark of a deficient mismatch repair system (MMR) in about 15% of colorectal cancers (CRC). Other studies and our own confirm MSI as an independent prognostic marker associated with a better outcome in stage II and III CRC. As MSI can be caused by different mechanisms, and dMMR leads to secondary oncogenic alterations, heterogeneity in the clinical and molecular features of MSI CRC is likely but not well understood. In this transcriptome expression study, we explore which genes differentiates MSI from Microsatellite stable (MSS) tumors and look for evidence of additional subclasses.

RNA extracted from formalin-fixed paraffin-embedded (FFPE) tissue blocks was used for expression profiling on the ALMAC platform SColorectal Cancer DSA Y T. Classifiers were constructed using AdaBoost and DLDA algorithms and assessed with area under the curve (AUC) by cross-validation. Survival analysis was based on Cox regression.

240 chips passed data quality control (42 MSI, 17.5%). Unsupervised methods allowed only weak separation of MSI and MSS specimen, despite 494 genes with significantly different expression (≤1% FDR), due to high variation in both groups. Gene expression differences were in agreement with results from reanalysis of three public datasets. Classifiers discriminated MSI and MSS with AUC of 0.96 [95% CI: 0.95-0.97], using 40-80 selected genes. Prominent discriminatory genes include various pathways: Wnt (f.ex. Axin2), MAPK (DUSP4); inflammation-immunity (REGs, STAT1), differentiation (TNNC2, mucins), metallothioneins. Association of these genes with RFS is heterogeneous.

Efficient discrimination of MSS and MSI in gene expression profiles can be obtained, with good quality FFPE material, using a multi-gene classifier. Inside both classes there is high residual heterogeneity. More samples will be profiled in order to further define molecular subgroups and to search for prognostic genes.

The ability to obtain reliable profiles form FFPE material implies that relevant information can be obtained from archival material stored in many biobanks.

Kevin Eng
Department of Statistics and of Biostatistics and Medical Informatics, University of Wisconsin, Madison
The neutral expectation in gene expression and tests of evolutionary hypotheses

For gene expression traits, the comparison of DNA sequence divergence and gene expression divergence has implications for the biological function, evolutionary constraints, and regulatory mechanism of the transcript. The comparison of divergences reduces to testing the apparent covariance versus the covariance expected under a neutral model.

We introduce tests of this neutral expectation with power against a broad set of evolutionary alternatives rendering them useful for the high-throughput screening of traits for evidence of selection. Because our simulation study finds the power of per-gene tests lacking at current sample sizes, we investigate the benefits of using a functional category-based analysis. We demonstrate the analysis on two different gene expression data structures.

Genevieve Erwin
UCSF Graduate Program in Bioinformatics, Gladstone Institute of Cardiovascular Disease
Identification of novel genomic regions associated with cardiovascular disease

Non-coding DNA accounts for 98% of our genome, and clinical and molecular data indicate that regulatory sequences in these non-coding regions strongly influence disease phenotypes. Functional non-coding regions regulate the timing and amplitude of gene expression, contributing to disease and diversity, but are poorly characterized compared to proteins. I am developing new methods to identify functional non-coding genomic regions and pinpoint candidate CVD-associated polymorphisms in these sequences. My approach leverages statistical models, as well as emerging genome-wide data sets of epigenetic markers and sequence variants in humans and other species.

Alisha Holloway
J. David Gladstone Institutes
Which genomic loci contribute to transcriptional divergence?

Differences in the gene regulatory network are hypothesized to contribute significantly to phenotypic divergence between and within species. Motivated by this idea, several studies have screened for rapidly evolving loci in the non-coding part of genomes in an effort to uncover the molecular basis for lineage specific traits. Conserved non-coding sequences with a burst of changes in a single lineage are promising candidates, because clusters of nearby substitutions are a hallmark of selection and have the potential to modify evolutionarily conserved regulatory elements. However, sequence divergence does not directly measure change in regulatory potential. Here we propose an alignment-free method to more directly relate sequence changes to regulatory changes by measuring divergence in transcription factor binding site (TFBS) profiles between pairs of evolutionarily related sequences.

Our approach utilizes probabilistic motifs of experimentally characterized TFBS, such as those in the JASPAR or TRANSFAC databases. For each sequence, we calculate the number of hits for a TFBS motif, which we denote n1 and n2 for a pair of sequences. We then assess the difference in TFBS hits k = n1 n2. When k is large we hypothesize the sequence differences are likely to affect transcription. To quantify large, we employ a model of correlated Bernoulli trials and derive the distribution of k.

We estimate two free parameters (p and z) from genome-wide data. This model enables fast calculation of p-values for the TFBS turnover statistic k, thereby providing a scalable tool to screen loci for regulatory divergence.

We apply this method to conserved regions near all human-mouse one-to- one orthologs. We identify functional categories of genes with the highest levels of TFBS divergence and compare these to the gene categories identified using sequence divergence alone.

Wolfgang Huber
European Molecular Biology Laboratory, Heidelberg, Germany
Clustering phenotype populations by genome-wide RNAi and multiparametric imaging

Genetic screens for phenotypic similarity have made key contributions to associating genes with biological processes. With RNA interference (RNAi), highly parallel phenotyping of loss-of-function effects in cells has become feasible. One of the current challenges however is the computational categorization of visual phenotypes and the prediction of biological function and processes. In this study, we describe a combined computational and experimental approach to discover novel gene functions and explore functional relationships. We performed a genome-wide RNAi screen in human cells and used quantitative descriptors derived from high-throughput imaging to generate multiparametric phenotypic profiles. We show that profiles predicted functions of genes by phenotypic similarity. Specifically, we examined several candidates including the largely uncharacterized gene DONSON, which shared phenotype similarity with known factors of DNA damage response (DDR) and genomic integrity. Experimental evidence supports that DONSON is a novel centrosomal protein required for DDR signalling and genomic integrity. Multiparametric phenotyping by automated imaging and computational annotation is a powerful method for functional discovery and mapping the landscape of phenotypic responses to cellular perturbations.

F. Fuchs, G. Pau, D. Kranz, O. Sklyar, C. Budjan, S. Steinbrink, T. Horn, A. Pedal, W. Huber, and M. Boutros (2010). Clustering phenotype populations by genome-wide RNAi multiparametric imaging. *Mol. Syst. Biol.*, **6**, 370.

G. Pau, F. Fuchs, O. Sklyar, M. Boutros, and W. Huber (2010). EBImage–an R package for image processing with applications to cellular phenotypes. *Bioinformatics*, **26**, 979–981.

Jason D. Lieb
Department of Biology and Carolina Center for the Genome Sciences, University of North Carolina, Chapel Hill
ZINBA: A unified modeling framework for the exploration and analysis of diverse chip-seq signal patterns

Despite Next Generation Sequencings increasing popularity, much remains to be characterized about experimental, sequencing, and computational biases that may artificially influence local read density. These biases, in addition to the diversity in signal patterns across various NGS datasets, pose problems to current signal processing methods which may not be sufficiently general to address these issues. We present a novel method applicable to a wide range of experimental NGS datasets and signal patterns, including ChIP-seq, FAIRE-seq, and Histone Modification data. Our mixture regression-based framework rigorously identifies, assesses and quantifies sets of factors that are relevant in explaining enriched and background signal in parallel. In addition, adjacent regions significantly enriched for signal are merged, allowing us to identify both broad and short regions of activity. We demonstrate how these factors play different roles across different data types, and show how incorporating these factors into our modeling framework can lead to greater performance in the determination of biologically relevant loci. We present a significant shift away from earlier methods of peak calling to a more flexible and unified modeling framework, applicable to many types of experimental situations and one that provides novel insight into NGS data.

Joint work with Niam Rashid, Paul G. Giresi, Joseph Ibrahim, and Wei Sun.

Jason D. Lieb
Department of Biology and Carolina Center for the Genome Sciences, University of North Carolina, Chapel Hill

Regulatory elements that define breast cancer progression and subtypes

Identification of open chromatin regions has been one of the most accurate and robust methods to identify functional promoters, enhancers, silencers, insulators, and locus control regions in mammalian cells. Here we have used FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) to survey the set of regulatory elements active during cellular transformation and between clinical tumor samples.

We were able to infer the function of active regulatory elements specific to tumor subtypes, regulatory factor binding sites, and gene expression profiles. FAIRE regions at sites distant from proximal promoter elements distinguished cancer subtypes and were highly associated with changes in transcriptional status. These sites reflected the activity of regulatory pathways specific to each subtype, including hormone receptor status. Thus, FAIRE is useful for characterizing human cancer and furthering our understanding of tumor biology.

In addition, we characterized the set of regulatory elements activated throughout the process of cellular transformation. In this model of oncogenesis, transformation is initiated by transient activation of Src, which induces a switch within untransformed MCF10A cells resulting in complete transformation within 36 hours. This switch entails the activation of the inflammatory response and other transcriptional regulatory pathways, including NF-kB and the Jak/Stat. We found that the majority of chromatin responses during cellular transformation were initiated within 4 hours. The set of regulatory elements were organized into distinct regions associated with differentially expressed genes, which we call COREs (Clusters of Open Regulatory Elements).

Joint work with Paul G. Giresi, Heather A. Hirsch, Charles M. Perou, and Kevin Struhl.

Pall Melsted
Department of Human Genetics, University of Chicago

De novo assembly and evolutionary analyses of liver-expressed genes in 16 mammal species

Changes in gene regulation have been proposed as critical in the evolution of phenotypic diversity of primates. However, the lack of high-quality reference genomes for most species and the limited number of independently derived transcripts for non-human primates has made it difficult to study gene regulation across multiple primates. To overcome these problems, we used massively-parallel sequencing to interrogate mRNA samples extracted from the livers of 16 species (12 primates including human, and 4 non-primate out groups; 4 samples per species). Of the 12 primate species, 8 do not have currently available reference genome sequences (vervet, galago, slow loris, and 5 lemur species), which means that we had to assemble the transcriptomes de novo for these species. This study design results in nucleotide sequence, quantitative expression, and gene structure data from thousands of genes, providing insight into gene regulation and sequence evolution across a broad spectrum of primate species.

Our analysis has revealed sets of genes whose expression level patterns are consistent with the action of natural selection along individual or ancestral primate lineages. We also investigated the relationship between alternative splicing within species and complete exon gains and losses between species. Interestingly, we also identified a large number of genes, highly conserved at the sequence level, that are expressed in the livers of some taxa, but that are unexpressed in others. Further, we have identified genes whose function and expression pattern we hypothesize may underlie specific adaptive processes. For example, SDR16C5 and AKR1B10, which are involved in retinol metabolism, are found to be highly expressed in marmosets relative to other species. Marmosets have a number of striking craniofacial adaptations, which allow them to gouge holes in trees, through which they can feed extensively on exuded latexes, saps, and gums. Retinol is one of the major derivatives of isoprene, the monomer of latex. It is believed that the initial stages of latex digestion are facilitated by bacteria in the large intestine, in which case, large quantities

of retinol may be absorbed through the intestinal wall and filtered by the liver. Therefore these increases in expression may reflect adaptations in the marmoset lineage to their latex-rich diets.

Joint work with John C. Marioni, George H. Perry, Ying Wang, Katelyn Michelini, Matthew Stephens, Jonathan K. Pritchard, and Yoav Gilad.

Katie Pollard
Gladstone Institutes, UCSF
Using metagenomics to study uncultured microbes: Who is out there, where do they live, and what are they doing?

Microorganisms are key contributors to the health of both ecosystems and humans. Despite their importance, relatively little is known about their ecological distribution and diversity. This is largely due to the difficulty of isolating microbes with modern laboratory techniques; 99% of microbial taxa are thought to be unculturable. Recent advances in shotgun sequencing of environmental DNA, a process known as metagenomics, enables direct interrogation of microbial genomic information from nature. However, the random and fragmentary nature of metagenomic sequence data, coupled with the complexity of most microbial communities, frustrates traditional sequence analysis and impedes discovery from metagenomic data. We are developing sophisticated bioinformatic methods that meet these challenges. Here, we demonstrate the application of our methods to the characterization of microbial ranges and diversity via the identification of Operational Taxonomic Units. Ultimately, we are applying our methods to study and compare microbial communities, analyze microbial genetic evolution, and understand the functional capabilities of microbes and their roles in ecosystems such as the human body.

Elodie Portales-Casamar
Centre for Molecular Medicine and Therapeutics, University of British Columbia
Analysis of Nrf2 binding sites and regulatory network through ChIP-Seq and global transcription profiling

Nrf2 (Nuclear factor E2 p45-related factor 2) is a transcription factor that activates the transcription of a plethora of cytoprotective genes to combat oxidative and electrophilic stresses. Under activating conditions, Nrf2 is released from the repressive actions of the Keap1 protein, and binds to a DNA motif called the antioxidant response element (ARE) located in the vicinity of target genes. These genes are responsible for detoxifying oxidants and electrophiles, and repairing or removing damaged macromolecules. However, many direct targets of Nrf2 remain undefined. Elucidating these direct targets of Nrf2 can provide potential therapeutic insights against chemical carcinogenesis, chronic inflammation, neurodegeneration, emphysema, asthma and sepsis. Using genome-wide chromatin immunoprecipitation followed by parallel sequencing (ChIP-Seq), we identified 1200 Nrf2-bound DNA sequences representing a 50-fold increase in the number of known Nrf2 target sites. High scoring matches to the ARE motif are over-represented in these sequences and often found in clusters, a property that had been reported in earlier studies. Integrating ChIP-Seq with global microarray profiling analyses, we identified over 600 basal and inducible direct targets of Nrf2 that participate in detoxification, proliferation, cell cycle, and survival networks. As expected, motif discovery applied to the Nrf2 ChIP-Seq regions identifies an 11 bp motif consistent with the current ARE model. In addition, we identified a significant enrichment for the dinucleotides AA or TT to be present either 2 or 3bp downstream of the AREs. These new properties will help define a better Nrf2 binding model for improved ARE predictions, which will both improve the analysis of functional variation across human genome sequences and lead to a better understanding of the Nrf2 regulatory network.

Jared Roach
Institute for Systems Biology
Family genome analysis

Lei Sun
DLSPH and Department of Statistics, University of Toronto
BR-squared: A practical solution to the winners curse in genome-wide scans

In genome-wide scans, the most significant variants detected in the original discovery study tend to have inflated effect size estimates due to the winners curse phenomenon. The winners curse has recently gained much attention in Genome-Wide Association Studies (GWAS), because it has been recognized as one of the major contributing factors to the failure of attempted replication studies. For example, five Nature Genetic publications in the first three months of 2009 acknowledged the effect of winners curse in their discovery samples (e.g. Nair et al., 2009). However, none made statistical adjustments to the nave estimates.

We extend our previous work (Sun and Bull, 2005) developed in the context of genome-wide linkage analyses to provide Bias-Reduced estimates via Bootstrap Re-sampling (BR-squared) for GWAS without collecting additional data. In contrast to the likelihood-based approaches (e.g. Zllner and Pritchard, 2007; Ghosh et al., 2008; Zhong and Prentice, 2008), the proposed method adjusts for the effects of selection due to both stringent genome-wide thresholds and ranking of the association statistics over the genome. In addition, our method explicitly accounts for the effect of allele frequency because the expected bias is inversely related to power of the association test.

We implemented the method to provide Bias-Reduced estimates via Bootstrap Re-sampling (BR-squared) for association studies of both disease status and quantitative traits, and we applied it in genome-wide association studies of Psoriasis and HbA1c. We observed over 50% reduction in the genetic-effect-size estimation for many associated SNPs which translates into a greater than 4-fold increase in sample size requirements for replication studies. Thus, adjusting for the effects of the winners curse is crucial for interpreting findings from genome-wide scans, and in planning replication studies, as well as attempts to translate findings into the clinical setting.

Jean-Philippe Vert
Mines ParisTech
Including prior knowledge in shrinkage classifiers for genomic data

Estimating predictive models from high-dimensional and structured genomic data, such as gene expression of CGH data, measured on a small number of samples is one of the most challenging statistical problems raised by current needs in post-genomics. Popular tools in statistics and machine learning to address this issue are shrinkage estimators, which minimize an empirical risk regularized by a penalty term, and which include for example support vector machines or the LASSO. In this talk I will discuss new penalty functions for shrinkage estimators, including generalizations of the LASSO which lead to particular sparsity patterns, and which can be seen as a way to include problem-specific prior information in the estimator. I will illustrate the approach by several examples such as the classification of gene expression data using gene networks as prior knowledge, or the classification and detection of frequent breakpoints in CGH profiles.

Pratyaksha Wirapati
Swiss Institute of Bioinformatics
Fast hierarchical clustering with small memory requirement for omics data analysis

We have implemented an R package for performing high-performance hierarchical clustering. The algorithm produces exactly the same results as the classical algorithm for arbitrary distance measures and for most link functions, including average linkage. However, it runs in quadratic time with respect to the number of objects clustered, unlike cubic time required by many implementations, such as hclust in R, and does not need quadratic memory space to temporarily store pairwise distances. The running time is still linear

with respect to the number of features for dense data matrix, and can be much faster for sparse data such as gene annotations. Some additional features are 1) built-in multi-level meta-analytical correlations for clustering variables in multi-platform datasets, 2) fast graphical display in R, and 3) branch reordering based on nearest-nephew rule. Data matrix with 20,000 variables and several hundred samples can be clustered in a few minutes on a laptop.