Banff International Research Station

for Mathematical Innovation and Discovery

Emerging Statistical Challenges in Genome and Translational Research

BIRS Workshop 08w5062
June 01–06, 2008

ORGANIZERS

Jenny 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
Sündüz Keleş, Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison
Katherine S. Pollard, Department of Statistics and Genome Center, University of California, Davis

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.
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 Max Bell, Rooms 155–159. Please respect that all other space has been contracted to other Banff Centre guests, including any food and beverage in those areas.*

SCHEDULE

- **Keynote lectures.** Keynote lectures, highlighted in bold face on the schedule, are 50 minutes long, with a 10-minute question period. Note that these lectures do not necessarily fall under the session topic.
- **Regular lectures.** Regular lectures are 30 minutes long, with a 10-minute question period.
- **White paper.** We would like to produce a conference white paper based on the talks and posters.

Day 0 **Sunday, June 01**

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, June 02**

07:00–08:45	Breakfast
	Regulation of Gene Expression
08:45–09:00	Introduction and welcome to BIRS by BIRS Station Manager, 159 Max Bell
09:00–10:00	**John Ngai**, *Molecular and genomics approaches to the vertebrate olfactory system: biological insights from genes identified by microarray analysis*
10:00–10:40	Sündüz Keleş, *Sparse partial least squares with applications to eQTL mapping*
10:40–11:00	Coffee Break
11:00–11:40	Ru-Fang Yeh, *Preprocessing and analysis of DNA methylation bead arrays*
11:40–12:20	Jenny Bryan, *Statistical methods for genome-wide phenotypic studies of gene deletion or inhibition*
12:20–14:00	Lunch
13:00–14:00	Guided tour of the Banff Centre; meet in the 2nd floor lounge, Corbett Hall
	Statistical Genomics
14:00–15:00	**Aseem Ansari**, *TBA*
15:00–15:40	Katie Pollard, *Nonparametric approaches to QTL mapping*
15:40–16:00	Coffee Break
16:00–16:40	Karl Broman, *Mapping multiple QTL in experimental crosses*
16:40–17:20	Ingo Ruczinski, *On missing data and genotyping errors in association studies*
17:30–19:30	Dinner

Day 2 **Tuesday, June 03**

07:00–09:00	Breakfast
	Statistical Genomics
09:00–10:00	**Mark van der Laan**, *Targeted maximum likelihood estimation: applications in genomics*
10:00–10:40	Adam Olshen, *Segmentation of allele-specific DNA copy number data*
10:40–11:00	Coffee Break
11:00–11:40	Franck Picard, *Linear models for the joint analysis of multiple array CGH profiles*
11:40–12:20	Annette Molinaro, *An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP enrichment*
12:20–13:45	Lunch
13:45–14:00	Group photo; meet on the front steps of Corbett Hall
	Cancer Genomics
14:00–15:00	**Mark Segal**, *Identifying regulatory networks using multivariate random forests*
15:00–15:40	Keith Baggerly, *Cell lines, microarrays, drugs and disease: trying to predict response to chemotherapy*
15:40–16:00	Coffee Break
16:00–16:40	Jason Lieb, *An atlas of open chromatin in diverse human cell types and breast cancer using FAIRE*
16:40–17:10	Discussion for morning and afternoon sessions
17:30–19:30	Dinner
19:30–21:30	**Poster Session**

Day 3 **Wednesday, June 04**

07:00–09:00	Breakfast
	Cancer Genomics
09:00–10:00	**Simon Tavaré**, *High-throughput detection of methylation patterns for tracking cell lineages*
10:00–10:40	Joseph F. Costello, *Large scale genome-epigenome interactions in human tumorigenesis*
10:40–11:00	Coffee Break
11:00–11:40	Neil Hayes, *Classification of lung cancer: the clinical experience*
11:40–12:20	Pratyaksha (Asa) Wirapati, *Co-analyzing datasets from multiple cancer studies: incorporating hierarchical models into differential expression, prediction and cluster analysis*
12:20–14:00	Lunch
	Free Afternoon/Hike TBA
19:00	**Workshop Celebration Dinner**, TBA off-site location, self-paid

Day 4 **Thursday, June 05**

07:00–09:00	Breakfast
	High-Throughput Biotechnologies
09:00–10:00	**Gordon Mills**, *Systems approach to personalized molecular medicine*
10:00–10:40	Jian-Bing Fan, *An integrated array platform for high-throughput genetic, epigenetic and gene expression analyses*
10:40–11:00	Coffee Break
11:00–11:40	Steffen Durinck, *High-throughput transcriptome sequencing*
11:40–12:20	Hongyu Zhao, *Hierarchical models for signal transduction pathway analysis from single cell measurements*
12:20–14:00	Lunch
	High-Throughput Biotechnologies
14:00–15:00	**Tim Hughes**, *Complexity, diversity, and conservation in the protein-DNA interactome*
15:00–15:40	Rafael Irizarry, *A gene expression barcode for microarray data*
15:40–16:00	Coffee Break
16:00–16:40	Joaquín Dopazo, *Casting genomic data into biological concepts*
16:40–17:10	Discussion for morning and afternoon sessions
17:30–19:30	Dinner

Day 5 **Friday, June 06**

07:00–09:00	Breakfast
	Open Topics
09:00–10:00	**Terry Speed**, *TBA*
10:00–10:40	Yee Hwa (Jean) Yang, *Identification of candidate microRNA using matched mRNA-miRNA time course data*
10:40–11:00	Coffee Break
11:00–11:40	Darlene Goldstein, *Glycan arrays and common themes in high throughput life science research*
11:40–12:20	Concluding remarks
12:20–14:00	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 meeting rooms, reading room) until 15:00 on Friday, although they are still required to checkout of the guest rooms by 12:00 noon.*

Emerging Statistical Challenges in Genome and Translational Research

BIRS Workshop 08w5062
June 01–06, 2008

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

Keith Baggerly, MD Anderson Cancer Center

Cell lines, microarrays, drugs and disease: trying to predict response to chemotherapy

Over the past few years, microarray experiments have supplied much information about the disregulation of biological pathways associated with various types of cancer. Many studies focus on identifying subgroups of patients with particularly aggressive forms of disease, so that we know who to treat. A corresponding question is how to treat them. Given the treatment options available today, this means trying to predict which chemotherapeutic regimens will be most effective.

We can try to predict response to chemo with microarrays by defining signatures of drug sensitivity. In establishing such signatures, we would really like to use samples from cell lines, as these can be (a) grown in abundance, (b) tested with the agents under controlled conditions, and (c) assayed without poisoning patients. Recent studies have suggested how this approach might work using a widely-used panel of cell lines, the NCI60, to assemble the response signatures for several drugs. Unfortunately, ambiguities associated with analyzing the data have made these results difficult to reproduce.

In this talk, we will describe how we have analyzed the data, and the implications of the ambiguities for the clinical findings. We will also describe methods for making such analyses more reproducible, so that progress can be made more steadily.

Karl Broman, University of Wisconsin, Madison

Mapping multiple QTL in experimental crosses

We consider the problem of identifying the genetic loci (called quantitative trait loci, QTL) contributing to variation in a quantitative trait, with data on an experimental cross (such as with mice). In the traditional approach to QTL mapping, one considers each genomic position, one at a time, and tests for association between genotype and the quantitative phenotype. Great attention has been placed on the adjustment for multiple hypothesis tests. The simultaneous consideration of multiple QTL can provide greater power, can better separate linked QTL, and allows the investigation of interactions between loci. The problem is best viewed as one of model selection. We describe the key issues and propose a penalized likelihood approach for model selection. Our approach provides an automated procedure that can enable biologists with limited statistical training to obtain a more complete understanding of the set of genetic loci contributing to variation in a quantitative trait.

Jenny Bryan, University of British Columbia
Statistical methods for genome-wide phenotypic studies of gene deletion or inhibition

Researchers in functional genomics can now obtain quantitative phenotypes for large collections of organisms, each of which is characterized by the deletion of an individual gene. By observing the phenotypic consequence of deletion across diverse conditions, we obtain specific information on the functional roles of the disrupted gene and on its relationship to other genes. The most mature application of this paradigm is in yeast, for which genome-wide collections of deletion mutants are available, but a similar approach is possible in other organisms through the use of RNA interference. I will present statistical approaches I have developed for the analysis of data from these high-throughput phenotypic studies, with some coverage of low-level issues, such as normalization, and high-level analyses, such as clustering and growth curve modelling on a large scale.

Joseph F. Costello, Comprehensive Cancer Center, University of California, San Francisco
Large scale genome-epigenome interactions in human tumorigenesis

A major challenge for the future is to understand how genomic and epigenomic aberrations cooperate directly or indirectly to develop the pathophysiologies that define human malignancies. Epigenetic mechanisms can cause genomic alterations, and genomic aberrations also can influence the cancer epigenome. Several possibilities exist for interaction. For example, genomic and epigenomic aberrations may cooperate directly to complete inactivation of tumor suppressors; by methylation of one allele and either deletion or mutation of the other. Alternatively, tumor suppressor genes may be silenced primarily by epigenetic mechanisms in one tumor type, and solely by genetic mutation in other tumor types. Given the extensive genetic and epigenetic alterations in any given tumor, integrative analyses thus represent a new kind of filtering approach to cull passenger alterations from those that are drivers of tumorigenesis, which will be discussed in this presentation.

Tumors simultaneously exhibit a global decrease in 5-methylcytosine relative to matching normal tissues. In mice, hypomethylation alone is sufficient to initiate tumorigenesis, and to alter the tumor spectrum in genetic models of cancer. These data suggest the level and type of interaction between genetic and epigenetic alterations might dictate the clinical course of tumors. Indeed, in the most severe cases of brain tumors, hypomethylation can affect more than 10 million CpGs in a single tumor, and these tumors are associated with a most aggressive clinical course. However, the mechanisms and extent to which this severe hypomethylation promotes to genome instability, and to tumorigenesis are not well defined. Given the specific technical limitations, these studies indicate that the integration of several experimental strategies will be required in order to maximize the discovery of new cancer-related genes, and the most functionally and clinically relevant genome-epigenome interactions.

Our ability to define the genomic and epigenomic events that contribute to cancer pathophysiology and response to therapy is determined by the analytical technologies that can be used to discover them. The power and genomic precision of analytical approaches are increasing dramatically as the technologies and information from the human genome project are harnessed for these purposes. I will discuss emerging technologies for increasingly comprehensive analyses of tumor genomes and epigenomes.

Joaquín Dopazo, Centro de Investigación Príncipe Felipe, Valencia, Spain
Casting genomic data into biological concepts

The ultimate goal of any genome-scale experiment is to provide a functional interpretation of the results, relating the available genomic information to the hypotheses that originated the experiment. Initially, this interpretation has been made on a pre-selection of relevant genes, based on the experimental values, followed by the study of the enrichment in some functional properties. Nevertheless, functional enrichment methods demonstrated to have a flaw: the first step of gene selection results too stringent given that the

cooperation among genes (within the modules aimed to find) was implicitly (and paradoxically) ignored. The assumption that modules of genes related by relevant biological properties (functionality, co-regulation, chromosomal location, physical interaction between proteins, etc.), and not the genes alone, are the real actors of the cell biology dynamics, lead to the development of new procedures implicitly closer to systems biology concepts. Such procedures, generically known as gene set methods, have successfully been used to analyze transcriptomic and large-scale genotyping experiments as well as to test other different genome-scale hypothesis in other fields such as phylogenomics. The use of modules has, however, some limitations that deserve to be commented.

Steffen Durinck, Illumina

High-throughput transcriptome sequencing

For the last decade microarrays have been the major technology used to study gene expression. Despite their popularity, microarrays have known limitations such as cross-hybridization, probe affinity effects, availability for sequenced genomes only, and limited ability to study alternative transcription. Recent advances in sequencing technologies have significantly reduced the cost of sequencing. These advances make it possible to now use sequencing for transcriptome studies. A single transcriptome sequencing experiment can potentially deliver information on alternative transcription, transcript-level expression profiles, gene mutation profiles, allele-specific expression patterns, presence of known and unknown pathogens, and gene fusion events. Sequencing transcriptomes on this scale is new and there is a tremendous need for development of new statistical and computational methods. One sequencing run on Illumina's 1G Genome Analysis System generates millions of short reads that have to be mapped to the genome or assembled together. New methods are needed to for example convert sequence data into exon and transcript-level expression measurements and to study differential transcript expression when comparing samples. In this talk I will give an overview of Illumina's Solexa sequencing technology as applied to sequencing transcriptomes, highlight the statistical and computational challenges, and discuss initial transcriptome sequencing results.

Jian-Bing Fan, Illumina

An integrated array platform for high-throughput genetic, epigenetic and gene expression analyses

At Illumina, our goal is to apply innovative technologies and revolutionary assays to the analysis of genetic variation and function, making studies possible that were not even imaginable just a few years ago.

We have developed a comprehensive line of products that address the scale of experimentation and the breadth of functional analysis required to achieve the goals of molecular medicine. This offering includes flexible and scalable, array-based technologies for: SNP genotyping, copy number variation detection, DNA methylation studies, gene expression profiling, and low-multiplex analysis of DNA, RNA, and protein. They serve as tools for disease research, drug development, and the development of molecular tests in the clinic.

Darlene Goldstein, École Polytechnique Fédérale de Lausanne, Switzerland

Glycan arrays and common themes in high throughput life science research

During the last decade, an important theme in life sciences research has been the emergence of high throughput assays which have produced data on a genomewide scale. A recent addition is the glycan array, used to study the biological roles for oligosaccharides. Glycomics represents another strategy for biomarker discovery.

In this talk, I will present some background in glycobiology and array-based glycomics, and give some applications in HIV and cancer. I will also touch on some of the recurring themes in high throughput life science research.

Keynote lecture

Tim Hughes, University of Toronto

Complexity, diversity, and conservation in the protein-DNA interactome

Mapping the complete spectrum of protein-DNA interactions is paramount to understanding global gene regulation and to fully decoding the genome and interpreting its evolution. We are undertaking a brute force effort to determine the binding preferences of as many individual mouse transcription factors as possible, by cloning and purifying transcription factor DNA-binding domains and then determining their binding specificity using a microarray technique that surveys relative affinity to all possible 8-mer sequences. To date, we have binding profiles for proteins in 23 different structural classes, including the majority of SOX, ETS, ARID, IRF, AP-2, GCM, and homeodomain family members. These data reveal a surprisingly rich landscape of DNA sequence preferences, with many proteins exhibiting what appear to be multiple binding modes. Homeodomains provide a striking example of the biochemical repertoire that can be achieved with a single, simple domain structure: there are at least 65 distinct homeodomain DNA-binding activities in the mouse alone. Since the binding preferences correlate with conserved protein sequence features, the mouse data can be used to predict relative 8-mer binding sequences for homeodomains in species as distant as Drosophila and C. elegans, with the highest-scoring sequences corresponding to known in vivo binding sites. Our results suggest that variation in sequence recognition may be a factor in the functional diversity and evolutionary success of many transcription factor DNA-binding domain families, but also show that the sequence specificity of many animal regulatory factors has not changed substantially since the Cambrian era. Ongoing efforts are aimed at a more comprehensive effort, focusing on families such as zinc-fingers that have undergone the greatest expansions in mammals.

Joint work with M. Berger, G. Badis-Breard, A. Gehrke, S. Talukder, L. Pena-Castillo, S. Jeager, E. Chan, T. Alleyne, Q.D. Morris, and M.L. Bulyk.

Rafael Irizarry, Johns Hopkins University

A gene expression barcode for microarray data

The ability to measure genome-wide gene expression holds great promise for characterizing cells and distinguishing diseased from normal tissues. Thus far, microarray technology has only been useful for measuring relative expression between two or more samples, which has handicapped the ability of microarrays to classify tissue types. This paper presents the first method that can successfully predict tissue type based on data from a single microarray hybridization. We achieved this by developing a statistical procedure that is able to accurately demarcate expressed from unexpressed genes and therefore defines a unique gene expression barcode for each tissue type. The utility of the method is demonstrated by defining a barcode-based classification algorithm with better predictive power than the best existing algorithms. Hundreds of publicly available human and mouse arrays were used to define and assess the performance of the barcode. With clinical data, we find near perfect predictability of normal from diseased tissue for three cancer studies and one Alzheimer s disease study. The barcode method also discovers new tumor subsets in previously published breast cancer studies that can be used for the prognosis of tumor recurrence and survival time. A preliminary web-tool, that when given a raw data file predicts tissue type, is available at `http://rafalab.jhsph.edu/barcode`.

Sündüz Keleş, University of Wisconsin, Madison

Sparse partial least squares with applications to eQTL mapping

TBA

Keynote lecture

Mark van der Laan, University of California, Berkeley

Targeted maximum likelihood estimation: applications in genomics

We present a general maximum likelihood based approach targeting a user supplied parameter of the data generating distribution. This approach results in locally efficient estimators fully tailored for the parameter of interest, and have been shown to be more robust than maximum likelihood estimators.

We illustrate this method for the purpose of assessing the effect of mutations in the HIV virus that cause resistance to a particular drug, for detecting binding sites in the regulatory region of the yeast genome based on gene expression experiments, for assessing the effect of gene expressions on response to treatment in breast cancer patients, and we illustrate the performance of targeted maximum likelihood estimation in a simulation study. We also illustrate the use of targeted maximum likelihood variable importance analysis for assessing the effect of SNP's on some disease outcome in case control studies.

Jason Lieb, University of North Carolina, Chapel Hill

An atlas of open chromatin in diverse human cell types and breast cancer using FAIRE

FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) is a simple low-cost genomic method for the isolation and identification of nucleosome-depleted regions in eukaryotic cells. 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. I will present preliminary results from two studies. In the first project, we are using FAIRE, coupled with high-resolution DNA microarrays and Solexa sequencing, to identify at active gene regulatory elements among cell types representative of most human tissues. This data should serve to greatly reduce the sequence space considered when searching for regulatory information encoded in DNA. In the second project, we are using FAIRE to identify the set of DNA-encoded regulatory elements active in human breast tumors excised from women treated at UNC Hospitals. We ultimately hope to identify which of these activities are anomalies characteristic of cancer, and the mechanisms by which they contribute to carcinogenesis.

Keynote lecture

Gordon Mills, MD Anderson Cancer Center

Systems approach to personalized molecular medicine

The realization of the promise of personalized molecular medicine will require the efficient development and implementation of novel targeted therapeutics. The goal will be to deliver the right drug to the right patient at the right time at the right dose. This effort will require a integration of information from the DNA, RNA and protein level into predictors of which patients are likely to respond to particular therapies. The overall likelihood of response to particular drugs represents the interaction between predictors of sensitivity with predictors of resistance. Efficient clinical trials testing these precepts will require the development and implementation of novel trial designs. It is likely that we will need to increase the size of phase I and II trials to allow the identification and validation of molecular markers at the same time as the initial evaluation the toxicity and efficacy of targeted therapeutics. This will come with the advantage of being able to deliver targeted therapeutics to enroll a much smaller population of patients selected for the likelihood to respond in phase III trials accelerating the approval of effective targeted therapeutics.

The phosphatidylinositol 3kinase (PI3K) pathway is aberrant at multiple levels across a wide variety of tumors making it the most common activating aberration in cancer. This has led to the development and now early clinical testing of drugs targeting multiple components of the pathway. The efficient utilization of these drugs will require the ability to accurately determine mutation and activation status in tumors as well as determining the interaction between the PI3K pathway and other pathways in driving tumor pathophysiology. Using a novel accurate and sensitive mass spectroscopy based sequencing approach, we have evaluated mutations in the PI3K pathway across more than 500 breast cancer samples. We have also implemented a high throughput functional proteomics approach designated reverse phase protein arrays to characterize the level and activity of multiple signaling pathways. We demonstrate than an integrated

analysis of mutation, proteins levels and protein activity is able to predict lack of response to trastuzumab in patients and to novel drugs targeting the PI3K pathway in vitro. This demonstrates that the response to targeted therapeutics is due to an interaction of markers of sensitivity and markers of resistance and provides important approaches for patient selection.

The PI3K pathway is critically important to cellular function and is thus under exquisite homeostatic control. The feedforward and feedback loops in the pathway determine the response to perturbation of the pathway by mutation or therapeutic intervention. Strikingly inhibition of the pathway at the level of mTOR or AKT results in the activation of potent feedback loops resulting in activation of multiple cell surface tyrosine kinases, PI3K itself and in the case of mTOR inhibitors, AKT. This may contribute to the observation that mTOR inhibitors appear to make some patient tumors grow more rapidly an unexpected and disappointing consequence of targeted therapeutics. Our preliminary systems biology-based mathematical and experimental models of the PI3K signaling network accurately predict these consequences as well as the biochemical processes involved. Further, the models suggest combinations of targeted therapeutics likely to reverse the negative effects of the mTOR inhibitors converting the outcome from negative to positive in terms of tumor growth.

Systems biology is the study of the emergence of functional properties that are present in a biological system but that are not obvious from a study of its individual components. Systems biology is a data-driven process requiring comprehensive databases at the DNA, RNA, and protein level to integrate systems biology with cancer biology. Combining these patient and model-based databases with the ability to interrogate functional networks by a systematic analysis using siRNA libraries and chemical genomics provides an ability to link in silico modeling, computational biology, and interventional approaches to develop robust predictive models applicable to patient management.

Annette Molinaro, Yale University

An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP enrichment

DNA methylation is an important component of epigenetic modification that influences the transcriptional machinery and is aberrant in many human diseases. Several methods have been developed to map DNA methylation for either limited regions or genome-wide. In particular, antibodies specific for methylated CpG have been successfully applied in genome-wide studies. However, despite the relevance of the obtained results, the interpretation of antibody enrichment is not trivial. Of greatest importance, the coupling of antibody-enriched methylated fragments with microarrays generates DNA methylation estimates that are not linearly related to the true methylation level. Here, we present an experimental and analytical methodology to obtain enhanced estimates which better describe the true values of DNA methylation level throughout the genome. We propose an experimental scenario for evaluating the true relationship in a high-throughput setting and a model-based estimate of the absolute and relative DNA methylation levels. We successfully applied this model to evaluate DNA methylation status of normal human melanocytes compared to a melanoma cell strain. Despite the low resolution typical of methods based on immunoprecipitation, we show that model derived estimates of DNA methylation provide relatively high correlation with measured absolute and relative levels, as validated by bisulfite genomic DNA sequencing. Importantly, the model derived DNA methylation estimates simplify the interpretation of the results both at single-loci and at chromosome-wide levels.

Joint work with Mattia Pelizzola, Yasuo Koga, Michael Krauthammer, Sherman Weissman, and Ruth Halaban.

Keynote lecture

John Ngai, University of California, Berkeley

Molecular and genomics approaches to the vertebrate olfactory system: biological insights from genes identified by microarray analysis

The vertebrate olfactory system receives and decodes sensory information from thousands of different chemical cues. The first step in this process is the recognition of these cues by receptors expressed on olfactory neurons, the primary sensory neurons in the olfactory epithelium. Receptor-mediated activity within the population of olfactory neurons is then interpreted by the brain to identify the molecular nature of the odorant stimulus. How is this process of molecular recognition accomplished? Current models of olfactory coding in vertebrates embody the hypothesis that individual olfactory neurons express a single odorant receptor gene (according to the one receptor, one neuron rule), and the axons of neurons expressing common receptors converge upon spatially invariant glomeruli in the olfactory bulb. In this manner, it is thought that an individual odorant, by activating a subset of odorant receptor molecules (and therefore a corresponding subset of olfactory neurons), elicits a spatial pattern of activity in the olfactory bulb characteristic for that odorant. Studies in my laboratory are focused on understanding the molecular principles of chemical recognition by the odorant receptors, and also on the mechanisms underlying the specification and development of the peripheral olfactory system. We use molecular, genomics, and computational approaches to study these issues, using the zebrafish and mouse as model experimental systems. In my talk I will highlight two studies that utilized genome-wide approaches to identify genes involved in two aspects of olfactory development: the differentiation of olfactory progenitor cells and the targeting of olfactory axons in the olfactory bulb.

Adam Olshen, Memorial Sloan Kettering Cancer Center

Segmentation of allele-specific DNA copy number data

Segmentation methods have gained popularity for identifying gains and losses in DNA copy number data. Traditionally segmentation has been applied to total copy number data, as that is all that is available for most copy number platforms. For SNP platforms, however, there are separate estimates of copy number for the two parental chromosomes, which necessitates new segmentation methodology. I will present our segmentation approach. I will demonstrate its utility on data from the Cancer Genome Atlas.

Joint work with Venkat Seshan and Richard Olshen.

Franck Picard, Université Claude Bernard Lyon I, France

Linear models for the joint analysis of multiple array CGH profiles

Segmentation methods have been successfully applied to the mapping of chromosomal abnormalities when using CGH microarrays. Current methods can deal with one CGH profile only, and do not integrate multiple arrays, whereas the CGH microarray technology becomes widely used to characterize chromosomal defaults at the cohort level. In this work, we propose a new statistical model to jointly segment multiple CGH profiles based on linear models. This strategy is very powerful for the joint segmentation of multiple profiles, as well as for the joint characterization of aberration types (status assignment of regions based on the cohort). One major difficulty lies in the limitation of dynamic programming whose complexity is limiting when the size of the profiles is important (for tiling arrays for instance), and when the number of samples increases. We solve the first issue using a CART-based strategy, where CART is used to select candidate breakpoints. The second part is solved with another Dynamic Programming algorithm. Overall, linear models offer a unified framework for the joint analysis of multiple CGH profiles, and we will show how they can be used to link the experience acquired in the field of expression arrays (normalization, experimental design) with array CGH data analysis.

Katie Pollard, University of California, Davis

Nonparametric approaches to QTL mapping

Understanding the genetic basis and architecture of phenotypic variation is the primary focus of quantitative genetics research. Quantitative trait locus (QTL) mapping is the process of identifying genetic loci that contribute to variation in a trait of interest and estimating their individual effects. Existing statistical methods for QTL mapping include single marker mapping, interval mapping, and composite interval mapping. These methods analyze traits individually and typically assume traits are normally distributed. We propose using a hidden Markov model with empirical likelihood-based emission probability distributions to identify QTLs. By exploiting information from all traits in a genomic study in parameter estimation, this novel approach may improve the power to detect QTLs. Preliminary results of applying this method to gene expression data from Arabidopsis thaliana are presented. In addition, because many traits are not normally distributed, we propose using an empirical likelihood approach for interval and composite interval mapping.

Ingo Ruczinski, Johns Hopkins University
On missing data and genotyping errors in association studies

Many algorithms to improve genotype estimates using raw data from high-throughput platforms have been proposed in the recent literature. Missing genotypes typically arise when the respective algorithms indicate that the confidence in certain genotype estimates is low. Thus, missing genotype data and genotyping errors are linked: for any particular genotype calling algorithm, higher (lower) genotype calling rates come with an increase (decrease) in false genotype calls, using different confidence thresholds for genotype calling. In this presentation, we discuss approaches for dealing with missing data and genotype uncertainties in association studies, and also show that accounting for genotype uncertainty can be crucial when inferring possible copy number variants.

Keynote lecture
Mark Segal, University of California, San Francisco
Identifying regulatory networks using multivariate random forests

Significant gains in prediction accuracy have been achieved by using ensembles of decision trees in the context of a single continuous or categorical response (cf random forests: Breiman, 2001, Statistical Science). Here we develop and illustrate the extension of random forests to multivariate responses. Multivariate response (single) regression tree methodology was first developed to handle longitudinal data (Segal, 1992, JASA) and has been applied in several ecological / environmental settings (e.g. Larsen and Speckman, 2002, Biometrics). More recently, select bioinformatics applications have appeared. In particular, Phuong (2004, Bioinformatics) used abundances of sequence motifs as covariates, and microarray gene expression levels as responses, to identify key regulatory elements in yeast under various experimental conditions. We revisit such applications using our (ensemble) random forest extension, focussing on several yeast microarray experiments, including cell cycle, and various stresses. We demonstrate that random forest derived covariate importance measures more reliably identify key regulators compared to relying on a single tree. Further, utilizing the proximity matrix from the forest output to cluster genes into homogeneous groups based on both motifs and expression values, we show that the multivariate response random forest effectively reveals high-order motif combinations that influence gene expression patterns, thereby obviating the need for examining the entire combinatorial space of all motif pairs, as previously undertaken (Pilpel, 2001, Nature Genetics).

Joint work with Yuanyuan Xiao.

Keynote lecture
Simon Tavaré, University of Cambridge/University of Southern California
High-throughput detection of methylation patterns for tracking cell lineages

Mutations within somatic cell genomes encode historical information about the relationships among the cells. It should therefore be possible to reconstruct aspects of these relationships by studying the appropriate mutations. We exploit methylation patterns at particular CpG islands to reconstruct ancestral information about stem cells and their lineages. I will describe a high-throughput bead-based technology that replaces cloning and sequencing of bisulfite treated DNA to identify such methylation patterns in single cells, and describe how it can be used to study tumour evolution.

Pratyaksha (Asa) Wirapati, Swiss Institute of Bioinformatics

Co-analyzing datasets from multiple cancer studies: incorporating hierarchical models into differential expression, prediction and cluster analysis

Publicly available clinical and genomics data are rapidly accumulating and the increased sample sizes promise more stable and consolidated results from genome-wide studies. However, combined analysis are still hampered by incommensurabilities due to disparate measurement platforms, data representation and study designs. Hierarchical sampling models (such as those based on meta-analysis, empirical Bayes or random-effect/random-coefficient models) should naturally be used to account for between-study heterogeneities. Although some solutions for two-sample differential expression problems have been proposed, extensions to other data types, such as survival, are still not clear. Furthermore, existing methods for more complex analysis modes, such as prediction and cluster analysis, assume single-study (one-level sampling) models. I will present a framework for modifying these commonly used "expression analysis workhorses" to accommodate datasets from multiple studies.

Yee Hwa (Jean) Yang, University of Sydney

Identification of candidate microRNA using matched mRNA-miRNA time course data

Analysis of microarray experiments often generate long gene lists that are not any easier to interpret compared to the original microarray experiments. This makes experimental verification of results extremely hard and of late, some attempts have been made to integrate lab-specific expression studies with other biological meta data and/or other biotechnologies to facilitate better interpretation and understanding of the biological question under investigation. In this talk, I will examine various aspects related to the analysis of a matched gene expression and microRNA expression timecourse experiment to understand and identify a small number of potential candidate microRNAs.

Ru-Fang Yeh, University of California, San Francisco

Preprocessing and analysis of DNA methylation bead arrays

Epigenetic alteration, specifically DNA methylation, is increasingly recognized as a major mechanism for gene regulation. Aberrant cytosine methylation in CpG dinucleotides is associated with silencing of tumor suppressor genes in many cancers. Using a microarray platform originally developed for SNP genotyping by Illumina, one can determine the methylation status of up to 1505 CpG sites with 96 samples at a time. In this talk, we will discuss the analysis issues for such data. We extended our multi-site, multi-array SNP genotyping algorithm, MAMS, to make dichotomized methylation calls and derive associated confidence measures as an alternative for the manufacture-recommended metric, relative intensity ratio beta. We also developed a likelihood ratio test and a model-based clustering algorithm based on the underlying beta distribution for differential methylation and clustering analysis. Applications on a large collection of normal and tumor tissue samples will be discussed.

Joint work with Yuanyuan Xiao, E Andres Houseman, and investigators in the BUCKDM3 Consortium.

Emerging Statistical Challenges in Genome and Translational Research

BIRS Workshop 08w5062
June 01–06, 2008

ABSTRACTS: POSTERS
(in alphabetic order by presenter surname)

Karl Broman, University of Wisconsin, Madison

Crossover interference and the sex difference in recombination

Many organisms exhibit large differences in recombination rate between the sexes. Why? The lengths of chromosomes at the key stage of meiosis can be quite different between males and females, and analysis of extremely large mouse backcrosses, with high density genotype data on a single chromosome, indicates that crossover interference (the tendency for crossovers to not occur too close together) is similar in the two sexes, but on a physical level, and by physical think m not bp. Longer female chromosomes, with a constant level of interference, would then allow more crossovers.

Sean Hanlon and Jason Lieb, University of North Carolina, Chapel Hill

Promoter architecture dynamically modulates transcription factor binding and transcriptional output over time, facilitating efficient responses to changing environments

Regulation of transcription factor targeting and transcriptional output is a key mechanism by which organisms respond to changing environments and developmental cues. To explore how targeting and transcriptional output relate to each other in a simple developmental context, we determined the genomic distribution of a yeast transcription factor, Rap1, during vegetative growth, respiratory growth, and throughout meiosis and sporulation in S. cerevisiae. Simultaneously, we monitored the expression of the genes downstream of binding events.

The majority of Rap1 targets were bound under all conditions tested, but some targets were bound specifically during respiratory growth and early meiosis, while at other targets binding was lost during sporulation. Within each of these target sets were examples of genes that were repressed during meiosis, and examples of genes that were induced. By depleting Rap1 from cells, we demonstrate that the presence of Rap1 is required simultaneously for the activation and repression of different targets. While there is no simple relationship between Rap1 binding and downstream transcriptional activity, differences in binding dynamics and transcriptional outcome can be partially explained by the presence of additional transcription factor binding motifs in target promoters and the composition and orientation of Rap1 motif contained within the promoters. Analysis of the genes comprising each class of Rap1 targets suggests that the cell utilizes the complex promoter architecture to allow this versatile transcription factor to direct a number of key pathways in the direction appropriate for the conditions or stresses that the cell is experiencing.

Sündüz Keleş, Christopher L. Warren, Clayton D. Carlson, and Aseem Z. Ansari, University of Wisconsin, Madison

CSI-Tree: A regression tree approach for modeling binding properties of DNA binding molecules based on cognate site identification (CSI) data

The identification and characterization of binding sites of DNA binding molecules, including transcription factors, is a critical problem at the interface of chemistry, biology and molecular medicine. The Cognate Site Identification (CSI) array is a high-throughput microarray platform for measuring comprehensive recognition profiles of DNA binding molecules Warren et al. (2006, PNAS). This technique produces datasets that are useful not only for identifying binding sites of previously uncharacterized transcription factors but also for elucidating dependencies, both local and non-local, between the nucleotides at different positions of the recognition sites.

We have developed a regression tree technique, CSI-Tree, for exploring the spectrum of binding sites of DNA binding molecules. Our approach constructs regression trees utilizing the CSI data of unaligned sequences. The resulting model partitions the binding spectrum into homogeneous regions of position specific nucleotide effects. Each homogeneous partition is then summarized by a position weight matrix. Hence, the final outcome is a binding intensity rank-ordered collection of position weight matrices each of which spans a different region in the binding spectrum. Nodes of the regression tree depict the critical position/nucleotide combinations.

We analyze the CSI data of the eukaryotic transcription factor Nkx-2.5 and two engineered small molecule DNA ligands and obtain unique insights into their binding properties. The CSI tree for Nkx-2.5 reveals an interaction between two positions of the binding profile and elucidates how different nucleotide combinations at these two positions lead to different binding affinities. The CSI trees for the engineered DNA ligands exhibit a common preference for the dinucleotide AA in the first two positions which is consistent with preference for a narrow and relatively flat minor groove. We carry out a reanalysis of these data with a mixture of position weight matrices approach. This approach is an extension of the simple position weight matrix method and accommodates position dependencies based on only sequence data. Our analysis indicates that the dependencies revealed by the CSI-Tree are challenging to discover without the actual binding intensities. Moreover, such a mixture model is highly sensitive to the number and length of the sequences analyzed. In contrast, CSI-Tree provides interpretable and concise summaries of the complete recognition profiles of DNA binding molecules by utilizing binding affinities.

Reference: `http://nar.oxfordjournals.org/cgi/content/abstract/gkn057`

Pei Fen Kuan, Dana Huebert, Audrey Gasch, and Sündüz Keleş, University of Wisconsin, Madison

A non-homogeneous hidden Markov model on first order differences for automatic detection of nucleosome positions

The heterogeneity of nucleosome densities across genomes and short linker regions are the two main challenges in mapping nucleosome occupancies based on chromatin immunoprecipitation on microarrays (ChIP-chip) data. Previous works rely on heuristic detrending and careful visual examination to detect low density nucleosomes, which may exist in subpopulation of cells. We propose a non-homogeneous hidden Markov model based on first order differences of experimental data along genomic coordinates that bypasses the need for local detrending and can automatically detect nucleosome positions of various occupancy levels. Our proposed approach is applicable to both ChIP-chip and ChIP-Seq (Chromatin Immunoprecipitation and Sequencing) data, and is able to map nucleosome-linker boundaries accurately. This automated algorithm is also computationally efficient and only requires a simple preprocessing step. We provide several examples illustrating the pitfalls of existing methods, the difficulties of detrending the observed hybridization signals and demonstrate the advantages of utilizing first order differences in detecting nucleosome occupancies via simulations and case studies involving ChIP-chip and ChIP-Seq data on nucleosome occupancy in yeast.

Katie Pollard and Dennis Kostka, University of California, Davis

Evolution of regulatory sequences after gene duplications in the ape lineage

Given the high level of sequence similarity between the human and chimpanzee genomes, rapid expansion of gene families has been put forward as a possible evolutionary mechanism to explain species-specific traits. Consistent with this model, Hahn and colleagues (Genetics 2007) observed accelerated gene turnover in the ape lineage, despite slower rates of nucleotide evolution compared to other mammals. They also found evidence for positive selection in the protein-coding sequences of genes in rapidly expanding gene families. We hypothesized that these gene families might also show accelerated evolution in upstream non-coding sequences, reflecting functional adaptation at the level of gene regulation. To investigate this hypothesis, we comprehensively analyzed substitution rates in non-coding sequences within 5Kb of the transcription start site of all human Ensembl genes. We applied a likelihood ratio test (LRT) to carefully filtered multiple alignments of up to 28 vertebrates in order to identify genes whose upstream regions have accelerated or decelerated substitution rates in the ape lineage. Approximately half of all loci show evidence of significant rate variation in the ape lineage (FDR adjusted $p < 0.05$). Strikingly, accelerated evolution is much more common among duplicated loci. To better understand the role of regulatory divergence in adaptive evolution of duplicated genes, we characterized the regulatory potential of significantly accelerated loci, identifying candidates for ape-specific regulatory modules. Overall, our findings are in line with a model in which duplicated genes rapidly adapt to a new regulatory context.

Mark Segal, University of California, San Francisco

Re-cracking the nucleosome positioning code

Nucleosomes, the fundamental repeating subunits of all eukaryotic chromatin, are responsible for packaging DNA into chromosomes inside the cell nucleus and controlling gene expression. While it has been well established that nucleosomes exhibit higher affinity for select DNA sequences, until recently it was unclear whether such preferences exerted a significant, genome-wide effect on nucleosome positioning in vivo. This question was seemingly and recently resolved in the affirmative: a wide-ranging series of experimental and computational analyses provided extensive evidence that the instructions for wrapping DNA around nucleosomes are contained in the DNA itself. This subsequently labelled "second genetic code" was based on data-driven, structural, and biophysical considerations. It was subjected to an extensive suite of validation procedures, with one conclusion being that intrinsic, genome-encoded, nucleosome organization explains approximately 50% of in vivo nucleosome positioning. Here, we revisit both the nature of the underlying sequence preferences, and the performance of the proposed code. A series of new analyses, employing spectral envelope (Fourier transform) methods for assessing key sequence periodicities, classification techniques for evaluating predictive performance, and discriminatory motif finding methods for devising alternate models, are applied. The findings from the respective analyses indicate that signature dinucleotide periodicities are absent from the bulk of the high affinity nucleosome-bound sequences, and that the predictive performance of the code is modest. We conclude that further exploration of the role of sequence-based preferences in genome-wide nucleosome positioning is warranted. This work offers a methodologic counterpart to a recent, high resolution determination of nucleosome positioning that also questions the accuracy of the proposed code and, further, provides illustration of techniques useful in assessing sequence periodicity and predictive performance.

Natalie Thorne, Cambridge University

DNA methylation array data analysis

Breaking the waves: improved CNV detection