



# Banff International Research Station

for Mathematical Innovation and Discovery

## Statistical Analysis of High-Throughput Genetic Data

Sunday, June 24 - Friday, June 29, 2007

### MEALS

\*Breakfast (Buffet): 7:00–9:00 am, Donald Cameron Hall, Monday–Friday

\*Lunch (Buffet): 11:30 am–1:30 pm, Donald Cameron Hall, Monday–Friday

\*Dinner (Buffet): 5:30–7:30 pm, Donald Cameron Hall, Sunday–Thursday

Coffee Breaks: As per daily schedule, 2nd floor lounge, Corbett Hall

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

### MEETING ROOMS

All lectures will be held in Max Bell 159 (Max Bell Building accessible by bridge on 2nd floor of Corbett Hall). Hours: 6 am–12 midnight. LCD projector, overhead projectors and blackboards are available for presentations. 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

#### Sunday

- 16:00 Check-in begins (Front Desk - Professional Development Centre - open 24 hours)  
Lecture rooms available after 16:00 (if desired)
- 17:30–19:30 Buffet Dinner, Donald Cameron Hall
- 20:00 Informal gathering in 2nd floor lounge, Corbett Hall (if desired)  
Beverages and small assortment of snacks available on a cash honour-system.

#### Monday

- 7:00–8:50 Breakfast
- 8:50–9:10 Introduction and Welcome to BIRS by BIRS Station Manager, Max Bell 159
- 9:10–10:10 Plenary talk: David Siegmund  
*Do complex statistical methods help in mapping complex and quantitative traits?*  
Chair: Jiahua Chen
- 10:10–10:30 Coffee break
- 10:30–11:30 Plenary talk: Heping Zhang  
*Tree and forest based approaches to genomewide association studies*  
Chair: Peter Song
- 11:30–1:00 Lunch
- 1:00–2:00 Guided Tour of The Banff Centre; meet in the 2nd floor lounge, Corbett Hall
- 2:00–2:30 Group Photo; meet on the front steps of Corbett Hall
- 2:30–3:30 Plenary talk: Sunil Rao  
*Spike and slab shrinkage for analyzing high throughput genomic data*  
Chair: Xuming He
- 3:30–4:00 Coffee break
- 4:00–4:30 Ji-Ping Wang, *Statistical method for nucleosome DNA sequence alignment and linker length preference prediction in Eukaryotic cells*
- 4:30–5:00 Peter Song, *Constructing Gene Networks with Time-Course Microarray Data*
- 5:00–5:30 Mary Lesperance, *Graphical Techniques for Gene Expression Studies*  
Organizer: Mary Lesperance
- 5:30–7:30 Dinner

## Tuesday

- 7:00–9:00 Breakfast
- 9:00–10:00 Plenary talk: Warran Ewens, *The transmission-disequilibrium test (TDT) and its generalizations*  
Chair: Fei Zou
- 10:00–10:30 Coffee break
- 10:30–11:30 Plenary talk: Shelly Bull, *Issues in Genome-wide Association: Power and Bias for Multi-stage Designs*  
Chair: Heping Zhang
- 11:30–1:30 Lunch
- 1:30–2:00 Ingo Ruczinski, *An integrated approach for the assessment of chromosomal abnormalities using SNP chip estimates of genotype, copy number, and uncertainty measurements*
- 2:00–2:30 Jian Huang, *Penalized Methods for Variable Selection and Estimation with High Dimensional Data*
- 2:30–3:00 Peng Jie, *Model Selection for QTL Mapping*  
Organizer: Heping Zhang
- 3:00–3:30 Coffee break
- 3:30–4:00 Josee Dupuis, *Mapping quantitative trait genes using high density SNP scans in extended families: challenges and partial solutions*
- 4:00–4:30 Ben Yakir, *Algorithms for Estimating Identity-By-Descent (IBD) From Dense Genotypes*
- 4:30–5:00 Benny Zee  
*Wavelet-based Prognostic Model for Gene Expression Profiling in Hepatocellular Carcinoma (HCC)*  
Organizer: David Siegmund
- 5:30–7:30 Dinner

## Wednesday

- 7:00–9:00 Breakfast
- 9:00–10:00 Plenary talk: Bruce Lindsay, *Inference for diffusion models for genomic sequence data*  
Chair: Mary Lesperance
- 10:00–10:30 Coffee break
- 10:30–11:00 Huixia Wang, *An Enhanced Quantile Approach for Assessing Differential Gene Expressions*
- 11:00–11:30 Liang Chen, *Considering Dependency among Genes for the False discovery Control of eQTL mapping*
- 11:30–12:00 Yongzhao Shao, *Power and Sample Sizes Analysis for FDR-Control in Genome-wide Studies*  
Organizer: Xuming He
- 12:00–1:30 Lunch

## Thursday

- 7:00–9:00 Breakfast
- 9:00–10:00 Plenary talk: Jun Liu, *Bayesian Inference of Haplotypes and Epistasis*  
Chair: Hongyu Zhao
- 10:00–10:30 Coffee break
- 10:30–11:30 Plenary talk: Hongyu Zhao, *Bayesian methods for reconstructing transcriptional regulatory networks*  
Chair: Jian Huang
- 11:30–1:30 Lunch
- 1:30–2:00 Hanfeng Chen, *Interval Mapping Methods for Genetic Trait Loci in Finite Regression Mixture Models*
- 2:00–2:30 Cindy Fu, *Testing homogeneity in mixtures of von Mises distributions*
- 2:30–3:00 Pengfei Li, *Iterative modified likelihood ratio test for homogeneity*  
Organizer: Cindy Fu
- 3:00–3:30 Coffee break
- 3:30–4:00 Fei Zou, *Fast Bayesian eQTL Analysis*
- 4:00–4:30 Daniel Weeks, *Linkage statistics that model relationship uncertainty*
- 4:30–5:00 Dongsheng Tu, *Identification of Patients Who Will Benefit from a Treatment Based on Their Genetic Profiles: Some Examples and Statistical Issues*  
Organizer: Hongyu Zhao
- 5:30–7:30 Dinner

## Friday

7:00–8:45 Breakfast

8:45–9:15 Steve Horvath, *Weighted Gene Co-Expression Network Analysis and Other Systems Genetic Approaches for finding Complex Disease Genes*

9:15–9:45 Jennifer Bryan, *Genome-wide studies of gene knockouts or inhibition*  
Organizer: Jiahua Chen

9:45–10:15 Coffee break

10:15–11:15 Joint presentation: Zehua Chen

*A tournament approach to model selection with applications in genome-wide association studies*

Jiahua Chen, *Extended Bayesian Information Criteria for Model Selection with Large Model Space*

Chair: Xuming He

11:15–11:30 Close conclusion

11:30–1:30 Lunch

**Checkout by 12 noon.**

\*\* 5-day workshops are welcome to use the BIRS facilities (2nd Floor Lounge, Max Bell Meeting Rooms, Reading Room) until 3 pm on Friday, although participants are still required to checkout of the guest rooms by 12 noon. \*\*



# Banff International Research Station

for Mathematical Innovation and Discovery

## Statistical Analysis of High-Throughput Genetic Data

Sunday, June 24 - Friday, June 29, 2007

### ABSTRACTS

(in alphabetic order by speaker first name)

Speaker: **Benjamin Yakir**, The Hebrew University of Jerusalem

Title: *Algorithms for Estimating Identity-By-Descent (IBD) From Dense Genotypes*

Abstract: Estimating IBD is an essential element in the construction of linkage statistics. A classical tool is the Lander-Green algorithm, which attempts to reconstruct the entire inheritance vector and fits an hidden Markov model. This approach is feasible when the pedigree is simple and the number of markers is small. Modern chip-based technology enables the collection of genotypic information over hundreds of thousands of markers. This wealth of information may raise the hope that IBD relationships can be inferred more directly without the need to model the entire inheritance history. We propose to use a robust algorithm that fits simple Markovian models to the data. Some components of these models reflect the observed genotypes and the linkage-disequilibrium dependencies while other components are hidden and reflect the states of IBD. Effort is made to execute more computationally intensive elements only at places where non-zero IBD is likely.

Speaker: **Benny Zee**, Chinese University of Hong Kong

Coauthors: Jack Lee, Nathalie Wong, Paul Lai, Winnie Yeo, Chinese University of Hong Kong

Title: *Wavelet-based Prognostic Model for Gene Expression Profiling in Hepatocellular Carcinoma (HCC)*

Abstract: Background: Hepatocellular Carcinoma (HCC) is a common cancer in Southeast Asia. In order to obtain a good prognostic model, clinical data alone may not be adequate. DNA microarray technology has enabled quantification of thousands of genes in a single assay but it has high background noise. Methods: The study includes 31 patients with cDNA microarray data containing 2398 genes. The clinical outcome is survival status at 1.8 years. Potential prognostic factors include genes expressions, albumin, ALT, bilirubin, AFP, ascites, alkaline phosphatase (APH), tumor size, and encephalopathy. The proposed method used a blocked wavelet-shrinkage principal component (BWSPCA) approach for dimension reduction with respect to the clinical outcome, and a penalized logistic regression was used to model the significant gene features and clinical prognostic factors. We compared the results to Artificial Neural Network (ANN) with multivariate Cox Model approach (Wei et al. 2004). Conclusion: Both methods have been shown good results in identifying gene expression as prognostic factors. However ANN approach is more complicated than BWSPCA approach. Acknowledgement: This study is funded by the Research Grant Council of Hong Kong #CUHK4469/03M

Speaker: **Bruce Lindsay**, Penn State University

Title: *Inference for diffusion models for genomic sequence data*

Abstract: This talk is a report on a new type of model for genomic sequence data as well as inference for that model. Our interest is focused data sampled from some population, and we would like to create a tree of relationships for those sequences. Each sampled sequence is modelled as having first been sampled from a population of ancestral sequences; that population is modelled by an unknown distribution  $Q$  on sequence space. The chosen sequence then undergoes  $T$  time units of evolution using Markov Chains that describe mutation and recombination processes before it is observed. The goal is to go backward in time  $T$  units and estimate the ancestral sequence distribution  $Q$ , as well as which modern sequence maps to each ancestor (or ancestors in the case of recombination). One methodology we have tried involves using maximum likelihood inference on the model for each fixed  $T$  on a grid, then linking the estimated ancestors together over  $T$  to create an ancestral tree (Biometrika). A second method is to use modal inference. This new method, which can be motivated by a "reverse diffusion" argument, seems to give results similar to maximum likelihood with much less programming and computation time.

Speaker: **Daniel Weeks**, Human Genetics, University of Pittsburgh

Coauthor: Amrita Ray, University of Pittsburgh

Title: *Linkage statistics that model relationship uncertainty*

Abstract: In linkage analysis, the familial relationships are assumed to be correct, thus misspecified relationships can lead to erroneous results. In practice, studies either discard individuals with erroneous relationships or use the best possible alternative pedigree structure. We have developed linkage statistics that model relationship uncertainty by properly weighing over the possible true relationships. Using simulated data containing relationship errors, we compared our statistics to the maximum likelihood statistic (MLS) and the Sall non-parametric LOD score. We considered small pedigree (SP) and large pedigree (LP) datasets. SP has same apparent relationship structure - full sibling for each affected pair - and LP has several different apparent relationship types. Two of our relationship uncertainty linkage statistics (RULS) have power as high as the MLS and Sall using the true structure. Also, these two RULS have greater power than the MLS and Sall using the 'discarded' structure.

Speaker: **David Siegmund**, Department of Statistics, Stanford University

Title: *Do Complex Statistical Methods Help in Mapping Complex and Quantitative Traits?*

Abstract: For mapping quantitative trait loci (QTL), I will discuss and compare (i) standard statistical methods for genome scans and (ii) more complex methods designed to take advantage of the possibilities of gene-gene and gene-environment interactions.

Speaker: **Dongsheng Tu**, Queen's University

Title: *Identification of Patients Who Will Benefit from a Treatment Based on Their Genetic Profiles: Some Examples and Statistical Issues*

Abstract: Tissue or blood samples are now routinely collected in the clinical trials evaluating treatments and preventions of diseases. Various types of genomic information can be extracted from these samples through different types of technologies such as DNA microarray or SNP assays. These genetic data hold a great promise in the identification of patients who will benefit most from a specific treatment and, thus, in the development of individualized treatment or prevention strategies for diseases. In this talk, I am going to present some examples from cancer clinical trials and discuss some statistical issues and challenges in this area of application.

Speaker: **Fei Zou**, University of North Carolina at Chapel Hill

Coauthors: Yuling Chang, Jinze Liu, Fred Wright, University of North Carolina at Chapel Hill

Title: *Fast Bayesian eQTL Analysis*

Abstract: eQTL are loci or markers on the genome that are associated with gene expression. eQTL analysis is essential to determine whether genetic influences on the expression are cis-acting or trans-acting. Underlying the eQTL analysis is the traditional Quantitative trait (QTL) analysis. Bayesian QTL mapping approaches have been widely used in experimental crosses, and have advantages in interpretability and in constructing parameter probability intervals. Most existing Bayesian linkage methods are prohibitive for high-throughput eQTL analysis because of the time-consuming Monte Carlo sampling procedure. We present a Bayesian linkage model that offers highly interpretable posterior densities for linkage. For this model, we develop Laplace approximations which are highly accurate and fast enough so that the computation of linkage posterior densities for over 30k transcripts becomes feasible. In addition, the ability in computing the probability of data at each transcript makes it possible to estimate the global cis-acting and trans-acting probabilities. If time allows, a semi-parametric Bayesian method with application in gene-expression studies will be discussed.

Speaker: **Hanfeng Chen**, Bowling Green State University

Title: *Interval Mapping Methods for Genetic Trait Loci in Finite Regression Mixture Models*

Abstract: There has been an explosion in the use of marker-based interval mapping method in quantitative genetics to systematically map loci underlying a genetic trait in experimental organisms. By this method, the putative genetic trait loci are assayed statistically via phenotype observations as well as genotype observations on a number of linked marker loci, thanks to modern molecular biology advances and recent developments in the statistical genetics. In this talk, we will review and discuss the latest statistical research results on the interval mapping method in finite regression mixture models with a general kernel function that includes most commonly-used kernel functions, such as exponential family mixture, logistic regression mixture and generalized linear mixture models.

Speaker: **Heping Zhang**, Yale University

Coauthors: Xiang Chen, Ching-Ti Liu, Yale University

Title: *Tree and forest based approaches to genomewide association studies*

Abstract: Tree-based analyses have become more and more useful in genetic studies. In this study, we will examine tree and forest based analyses in genomewide association studies to identify high risk genes and gene-gene interactions.

Simulation studies will be used to compare the power of our approach with existing ones. We will also apply our approach to a real data set.

Speaker: **Hongyu Zhao**, Yale Medical School

Coauthors: Ning Sun, Yale Medical School; Raymond Carroll, Texas A&M University

Title: *Bayesian methods for reconstructing transcriptional regulatory networks*

Abstract: Transcription regulation is a fundamental biological process, and extensive efforts have been made to dissect its mechanisms through direct biological experiments and regulation modeling based on physical-chemical principles and mathematical formulations. Recent advances in high throughput technologies have provided substantial amounts and diverse types of genomic data that reveal valuable information on transcription regulation, including DNA sequence data, protein-DNA binding data, microarray gene expression data, and others. In this presentation, we describe a Bayesian error analysis model to integrate protein-DNA binding data and gene expression data to reconstruct transcriptional regulatory networks. There are two unique aspects to this proposed model. First, transcription is modeled as a set of biochemical reactions, and a linear system model with clear biological interpretation is developed. Second, measurement errors in both protein-DNA binding data and gene expression data are explicitly considered in a Bayesian hierarchical model framework. Model parameters are inferred through Markov chain Monte Carlo. The usefulness of this approach is demonstrated through its application to infer transcriptional regulatory networks in the yeast cell cycle.

Speaker: **Huixia Wang**, Department of Statistics, North Carolina State University

Coauthors: Xuming He, University of Illinois at Urbana-Champaign.

Title: *An Enhanced Quantile Approach for Assessing Differential Gene Expressions*

Abstract: Due to the small number of replicates in typical gene microarray experiments, the power of statistical inference is often unsatisfactory without some form of information-sharing across genes. In this paper, we propose an enhanced quantile rank score test (EQRS) for detecting differential expression in GeneChip studies by analyzing the quantiles of gene intensity distributions through probe level measurements. A measure of sign correlation,  $\delta$ , plays an important role in the rank score tests. By sharing information across genes, we develop a calibrated estimate of  $\delta$ , which reduces the variability at small sample sizes. We compare the EQRS test with four other approaches for determining differential expression: the gene-specific quantile rank score test, the quantile rank score test assuming a common  $\delta$ , a modified t-test using summarized probe set level intensities, and the Mack-Skilings rank test on probe level data. The proposed EQRS is shown to be favorable for preserving false discovery rates and for being robust against outlying arrays. In addition, we demonstrate the merits of the proposed approach using a GeneChip study comparing gene expression in the livers of mice exposed to chronic intermittent hypoxia and of those exposed to intermittent room air.

Speaker: **Ingo Ruczinski**, Johns Hopkins University

Coauthors: Rob Scharpf, Giovanni Parmigiani, Johns Hopkins University

Jonathan Pevsner, Kennedy Krieger Institute

Title: *An integrated approach for the assessment of chromosomal abnormalities using SNP chip estimates of genotype, copy number, and uncertainty measurements*

Abstract: Many chromosomal abnormalities such as amplifications, deletions, and copy-neutral loss of heterozygosity have been associated with disease. High-throughput single nucleotide polymorphism (SNP) arrays are useful for the genome-wide assessment of such aberrations. Hidden Markov Models (HMMs) have been proposed for the analysis of such SNP array data, to explicitly utilize the correlation structure between the data derived from neighboring SNPs for the assessment of the true underlying chromosomal "states". Here we improve previously reported approaches by simultaneously integrating gene copy number estimates, genotype calls, and the corresponding confidence scores when available. Using simulated data, we demonstrate how confidence scores control smoothing in a probabilistic framework. Software for fitting these HMMs to SNP array data is available in the R package ICE.

Speaker: **Jennifer Bryan**, University of British Columbia

Title: *Genome-wide studies of gene knockouts or inhibition*

Abstract: In many organisms, the function of each gene in the genome is being studied by observing the phenotype after gene deletion or inhibition (such as with RNA interference), perhaps in the context of one or more other environmental or genetic perturbations. I will describe this new high throughput platform, as it is being realized through the Yeast Deletion Set and will highlight key statistical challenges and some of the solutions being developed by myself and my collaborators.

Speaker: **Ji-Ping Wang**, Department of Statistics, Northwestern University

Title: *Statistical method for nucleosome DNA sequence alignment and linker length preference prediction in Eukaryotic cells*

Abstract: Eukaryotic DNAs exist in a highly compacted form known as chromatin. The nucleosome is the fundamental repeating subunit of chromatin, formed by wrapping a short stretch of DNA, 147bp in length, around four pairs of histone proteins. Nucleosome DNA obtained by experiments however varies in length due to imperfect digestion. We develop a mixture model that characterizes the known dinucleotide periodicity probabilistically to improve the alignment of nucleosomal DNAs. To further investigate chromatin structure, we experimentally cloned and sequenced di-nucleosome sequences from yeast, chicken and human. Each dinucleosome sequence roughly cover two nucleosomes (located toward the two ends) with a linker DNA in between. A HMM model is trained based on the nucleosome sequence alignment for prediction of nucleosome positioning. Results show that Eukaryotic cells do favor periodic linker length in chromatin forming on a roughly 10 bp basis, however with two different forms, i.e. with peaks around 5?s or 10?s.

Speaker: **Jiahua Chen**, Department of Statistics, University of British Columbia

Coauthor: Zehua Chen, National University of Singapore

Title: *Extended Bayesian Information Criteria for Model Selection with Large Model Space*

Abstract: It has been observed that the ordinary Bayes information criterion is too liberal for model selection when the model space is large. In this talk, we re-examine the Bayesian paradigm for model selection and propose an extended family of Bayes information criteria. Unlike the original Bayes information criterion, which balances the log likelihood by a penalty on the number of unknown parameters, the extended Bayes information criteria take into account both the number of unknown parameters and the complexity of the model space. The consistency of the extended Bayes information criteria is established. Their performance in various situations is evaluated by simulation studies. They are compared with the original Bayes information criterion in terms of positive selection rate and false discovery rate in problems of variable selection. It is demonstrated that the extended Bayes information criteria incurs a little loss in positive selection rate but tightly controls false discovery rate, a desirable property in many applications. The extended Bayes information criteria are extremely useful for variable selection in problems with moderate sample size but huge number of covariates, especially, in genome-wide association studies which is now a hot area in genetics research.

Speaker: **Jian Huang**, University of Iowa

Coauthors: Shuangge Ma, Yale University; Huiliang Xie, University of Iowa; Cun-Hui Zhang, Rutgers University

Title: *Penalized Methods for Variable Selection and Estimation with High Dimensional Data*

Abstract: A general approach for fitting sparse, high-dimensional models is to use regularization penalties. Several important penalized methods, including lasso and bridge penalties, for variable selection and estimation have been proposed, but the properties of these methods have not been systematically studied. To apply the methods in scientific investigations, it is helpful to understand their properties. In particular, it is helpful to know under what conditions, the methods correctly select the important variables and estimate their effects consistently and efficiently. In this talk, we present some preliminary results concerning the variable selection consistency and asymptotic oracle properties of lasso and bridge methods in high-dimensional settings, and discuss some extensions of these methods. We also illustrate applications of these methods to the analysis of binary and censored outcomes with high-dimensional genomic covariate data.

Speaker: **Jie Peng**, UC Davis, Department of Statistics

Coauthors: Jiming Jiang, Thuan Nguyen, UC Davis

Title: *Model Selection for QTL Mapping*

Abstract: As pointed out by many authors, the QTL mapping problem can be viewed as a model selection problem. This view is especially valuable nowadays when we have highthroughput marker data. In this paper, we propose two methods for model selection in QTL mapping. One is the “fence method” which is useful for model selection in both regression and random effects models and enjoys some computational advantages over the BIC, AIC type of methods in searching of the model space. The other is a lasso type penalized regression approach. The lasso regression (or more general sparse regression) exploits the sparsity of the model which is the case in QTL mapping since much fewer QTLs exist compared to the large number of markers. Simulation results as well as application on real data sets will be used to demonstrate the performance of the proposed methods and the comparison with existing ones.

Speaker: **Josee Dupuis**, Boston University School of Public Health

Coauthors: Kathryn Lunetta, Alisa Manning, L. Adrienne Cupples, Boston University School of Public Health

James Meigs, Massachusetts General Hospital and Harvard Medical School

Title: *Mapping quantitative trait genes using high density SNP scans in extended families: challenges and partial solutions*

Abstract: High density SNP scans are currently being performed on several family-based population samples with multiple phenotypes available. There are several statistical challenges related to finding genes influencing quantitative traits in such rich datasets. Is there any value to performing linkage analysis using dense SNPs, when it is believed that genome-wide association (GWA) analysis is the answer to all questions? With the huge number of tests that result from a GWA scan with multiple phenotypic outcomes, how does one control type-I error and still retain some power to detect effects of modest size? Should one perform analyses to safe-guard against false positive results arising from population stratification, at a cost of a possible reduction in power, or should one rely on replication studies to limit false positive associations? We motivate and present some partial solutions to these questions in the context of the Framingham Heart Study.

Speaker: **Jun Liu**, Department of Statistics, Harvard University

Coauthors: Yu Zhang, Tim Niu, Harvard University

Title: *Bayesian Inference of Haplotypes and Epistasis*

Abstract: Haplotypes provide complete information of inheritance, which are very useful in population genetics and association studies. Since experimentally determining haplotype data is expensive, much effort has been devoted to develop computational tools for inferring haplotypes from genotype data. I will present a few Bayesian and semi-Bayesian models that have been formulated over the past few years for this task, including new hierarchical Bayes model developed in our group that incorporates the coalescence effect in a prior distribution. The prediction accuracy of the new method is uniformly improved compared to existing methods such as HAPLOTYER and PHASE. I will further discuss a Bayesian approach in detecting multi-locus interactions (epistasis) for case-control association studies. Existing methods are either of low power or computationally infeasible when facing of a large number of markers. Using MCMC sampling techniques, the method can efficiently detect interactions among thousands of markers. Using simulation results, I will discuss the power of our approach and the importance to consider epistasis in association mapping.

Speaker: **Liang Chen**, University of Southern California

Title: *Considering Dependency among Genes for the False discovery Control of eQTL mapping*

Abstract: “Genetical genomics” searches for DNA variants associated with gene expression variations among segregated populations. Such DNA variants are called expression Quantitative Trait Loci (eQTL). Although genetical genomics is a promising approach in many fields, the statistical analysis of eQTL data involving thousands of genes and thousands of markers presents many statistical and computational challenges. Here, we focus on one challenge: multiple testing. We treat the eQTL mapping as a problem of identifying differentially expressed genes across different marker genotypes. In previous studies, the dependency among genes is largely ignored in consideration of multiple comparison adjustments. However, such dependency may be strong for eQTL data. It can have significant impact on data analysis and interpretation. We introduce a weighted version of false discovery control to improve the statistical power to identify eQTLs. Different genes are assigned to different weights according to the dependence structure. The relative performance of weighted false discovery control in eQTL studies is illustrated through simulation studies and real data analysis.

Speaker: **Mary Lesperance**, Department of Mathematics and Statistics, University of Victoria

Title: *Graphical Techniques for Gene Expression Studies*

Abstract: Correspondence analysis (CA) is a descriptive technique designed for investigating the association between row and column variables by graphically displaying the patterns in the data. It has been widely applied to categorical data. We explore and develop variations of CA techniques to identify differentially expressed genes and to assess the quality of replicate DNA arrays.

Multiple correspondence analysis (MCA) and a related technique called joint correspondence analysis (JCA) are methods for visualizing the joint features of 2 or more categorical variables. We have been working with the Genetic Pathology Evaluation Centre (GPEC) at UBC and the Breast Outcomes Unit (BCOU) at the B.C. Cancer Agency (BCCA) to study relationships between molecular markers and outcomes for breast cancer. Molecular markers and diagnostic variables are typically categorized as positive/negative by pathologists and oncologists, whereas outcome measures such as time to recurrence or breast cancer specific survival time are continuous and possibly censored. Some researchers have incorporated survival information in an MCA analysis without regard for any inherent censoring.

Speaker: **Pengfei Li**, University of Waterloo

Coauthor: Jiahua Chen, University of Waterloo and University of British Columbia

Title: *Iterative modified likelihood ratio test for homogeneity*

Abstract: Testing for homogeneity in finite mixture models has attracted substantial research recently. Various methods have been proposed and investigated. Modified likelihood ratio test (MLRT) is a nice method because it has an asymptotically distribution-free test statistic and is locally most powerful. Interestingly, the mixture of exponential distributions or mixture models in scale distribution families do not satisfy the regularity conditions prescribed by many methods including the MLRT. To overcome this difficulty, we propose an iterative modified likelihood ratio test (IMLRT) in this paper. The IMLRT statistic has the same simple limiting distribution as MLRT statistic. The result is applicable to much more general mixture models and it does not require the parameter space to be bounded. Simulations show that the IMLRT has more accurate type I errors and higher powers under various models compared to existing methods. We also apply the IMLRT to a real data example.

Speaker: **Peter Song**, University of Waterloo

Coauthors: Xin Gao and Qiang Pu, York University

Title: *Constructing Gene Networks with Time-Course Microarray Data*

Abstract: Gene-gene interaction provides the basis for the construction of gene networks that are crucial for the understanding of many underlying biological mechanisms. Time-course microarray data gives rise to an interesting platform to reveal how gene-gene interactions evolve over time. We invoke hidden Markov models (HMM) for measurements of gene expression, in which the expression process is driven by a process of hidden states. This HMM framework enables us to explicitly account for gene-gene dependency by jointly modeling a set of hidden states associated with multiple genes. The transition matrix is estimated through a new EM algorithm based on composite likelihood. To form edges in a gene network, significant interactions are determined by composite-likelihood ratio test, with the control of a reasonable false discovery rate. Simulation studies and real biological data analysis will be used to demonstrate the application of the proposed method.

Speaker: **Shelley Bull**, University of Toronto

Coauthors: Lei Sun, University of Toronto; Andrew Paterson, Hospital for Sick Children Research Institute

Xinlei Xie, Samuel Lunenfeld Research Institute of Mount Sinai Hospital

LongYang Wu, University of Waterloo

Title: *Issues in Genome-wide Association: Power and Bias for Multi-stage Designs*

Abstract: Because of improved efficiency with respect to genotyping costs and/or sample accrual, high-density genome-wide association (GWA) studies are typically designed with multiple stages. Whether the second stage involves an independent sample of individuals or a family-based design, genetic effect estimates at a second stage will be less optimistic than those obtained at the first stage, due to selection bias arising from genome-wide screening. Motivated by a GWA study of the genetics of complications of type I diabetes, we evaluate implications for power and bias in alternative designs and analytic strategies for detection and mapping of gene regions using high-density SNP arrays. The bias of genetic effect estimates depends on sample size, true effect size, minor allele frequency, and screening stringency. Application of computationally-intensive bootstrap estimation yields less biased effect estimates, and hence more realistic specifications for replication in a subsequent stage.

Speaker: **Steve Horvath**, University of California

Title: *Weighted Gene Co-Expression Network Analysis and Other Systems Genetic Approaches for finding Complex Disease Genes*

Abstract: Weighted gene co-expression network analysis (WGCNA) facilitates a systems biologic view of gene expression data. The network framework makes it straightforward to integrate gene expression data with other types of data, e.g. clinical traits, proteomics data or genetic marker data. The talk covers several theoretical topics including network construction, module definition, network based gene screening, and differential network analysis. The methods are illustrated using several applications including i) screening for biomarkers of kidney transplantation success ii) finding obesity related genes in mice, and iii) complex disease gene mapping in humans. Related articles and material can be found at the following webpage <http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/>

Speaker: **Sunil Rao**, Case Western Reserve University

Coauthor: Hemant Ishwaran, Cleveland Clinic

Title: *Spike and slab shrinkage for analyzing high throughput genomic data*

Abstract: High throughput genomic and proteomic technologies are revolutionizing the way we now study chronic diseases like cancer. Typical examples of such technologies are gene expression arrays, fine map SNPs, and proteomic mass spectra. Much of the excitement is driven by the potential for a deeper understanding of mechanistic behaviour, as well as the potential to identify novel biomarkers for diagnostics and therapeutics. A hallmark of these data however

is that the sheer number of variables (i.e. genes, SNPs or proteins) that one has to accommodate can get very large; yet only a relatively small number of these variables might prove truly insightful (i.e. a sparse underlying true relationship). In this joint talk, we discuss a theoretical approach to such problems making use of a type of regularization known as spike and slab shrinkage. Applications originally involved the detection of differentially expressing genes in stagewise colon cancer progression, but the methodology can also be applied to other scenarios like the identification of gene expression signatures that correlate with survival and the combination of genotype and gene expression profiles to study the genetic determinants of the natural variation of gene expression in an otherwise healthy population.

Speaker: **Warren Ewens**, University of Pennsylvania

Title: *The transmission-disequilibrium test (TDT) and its generalizations*

Abstract: The transmission-disequilibrium test was developed by Spielman, McGinnis and the author as a test of linkage between a marker locus and a purported disease susceptibility locus. It is however more frequently used today as a test of association between the alleles at these respective loci. Complications to the test arise with the current availability of millions of marker loci. Many authors have developed generalizations of the test to handle this and other novel situations. A review of the conceptual thinking behind the test will be given, followed by a description of these developments.

Speaker: **Yongzhao Shao**, New York University School of Medicine

Coauthor: Chi-Hong Tseng, New York University

Title: *Power and Sample Sizes Analysis for FDR-Control in Genome-wide Studies*

Abstract: In fields as diverse as genetics, genomics, and medical imaging, researchers may test up to hundreds of thousands of hypotheses at a time, leading to the problem of multiple comparisons for high dimensional data. For example, DNA microarrays have been widely used for the purpose of monitoring expression levels of thousands of genes simultaneously to identify those genes that are differentially expressed. The false identification probability can increase sharply when the number of tested genes gets large. In addition to the challenge of adjusting for multiple testing while maintain high statistical power, a further challenge is the existence of dependence between the test statistics due to reasons such as gene co-regulation and/or correlation in the measurement errors. Appropriate adjustment for dependency among test statistics is critical to avoid serious loss of power or an abundance of false positive results. We introduce a general approach to calculate the sample size needed to ensure adequate overall power while controlling the false discovery rates (FDR) to avoid an abundance of false positive results. In particular, we investigate practical methods to adjust for dependence among test statistics. The usefulness of the proposed sample size formula is demonstrated using both simulated data as well as real data.

Speaker: **Yuejiao Cindy Fu**, York University

Coauthors: Jiahua Chen, University of British Columbia; Pengfei Li, University of Waterloo

Title: *Testing homogeneity in mixtures of von Mises distributions*

Abstract: Testing homogeneity has always been an important and difficult research problem in finite mixture models, especially in the presence of a structural parameter. The von Mises distribution and its mixture are widely used for circular data arising naturally from many sciences. The modified likelihood ratio test has been successfully applied for the homogeneity test in a variety of mixture models. We propose the use of the modified likelihood ratio test and the iterative modified likelihood ratio test in general two-component von Mises mixture with a structural parameter. Two accuracy enhancing methods are developed. The limiting distributions of the resulting test statistics are derived. Simulations show that the test statistics have accurate type I errors and adequate power. Two real data examples are also provided.

Speaker: **Zehua Chen**, National University of Singapore

Coauthor: Jiahua Chen, University of British Columbia

Title: *A tournament approach to model selection with applications in genome-wide association studies*

Abstract: Recent genome-wide association studies in genetics pose many challenging problems for statisticians. One of such problems is model selection with a huge number of covariates. The data from genome-wide association studies typically involves tens or hundreds of thousands SNPs, but the sample size is much less than the number of SNPs. The sheer amount of the covariates under concern and a relatively small sample size make the currently existing model selection methods infeasible. To take the challenge, we propose a novel tournament approach to model selection for the situation that the number of covariates under consideration far exceeds the sample size. The approach consists of a stage-wise screening procedure, which mimics the rounds of competitions in a tournament, and an extended Bayes information criterion. In the stage-wise screening procedure, the collection of all the covariates under consideration

is randomly partitioned into small size groups. For each group, a non-quadratic penalized likelihood with a properly tuned penalty parameter is maximized to select a specified number of relatively more important covariates. The selected covariates from all the groups are mixed together and re-partitioned. The same selection process repeats for the newly partitioning groups. The process continues until a tentative final group of candidate covariates with a specified size is reached. To reduce the possible erratic nature of the random partitioning, a permutation aggregation is introduced into the screening procedure; that is, the stage-wise screening procedure is repeated for many times with a different random partition at each time, and the candidate covariates are then selected according to their frequencies of appearance in all the tentative final groups. The eventual final group of covariates then undergo a refined model selection procedure with the extended Bayes information criterion. The extended Bayes information criterion differs from the original Bayes information criterion by that it takes into account not only the number of unknown parameters of a model but also the complexity of the model space which the model belongs. The consistency of the extended Bayes information criterion has been established. In this talk, we discuss the tournament procedure and the extended Bayes information criterion in detail. We also present some simulation results and a real data analysis for a genome-wide association study where the tournament approach is applied.