13w5083: Statistical Data Integration Challenges in Computational Biology: Regulatory Networks and Personalized Medicine
Aug 11 - Aug 16, 2013

MEALS

*Breakfast (Buffet): 7:00–9:30 am, Sally Borden Building, Monday–Friday
*Lunch (Buffet): 11:30 am–1:30 pm, Sally Borden Building, Monday–Friday
*Dinner (Buffet): 5:30–7:30 pm, Sally Borden Building, Sunday–Thursday
Coffee Breaks: As per daily schedule, in the foyer of the TransCanada Pipeline Pavilion (TCPL)
***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 the lecture theater in the TransCanada Pipelines Pavilion (TCPL). An LCD projector, a laptop, a document camera, and blackboards are available for presentations.

SCHEDULE

Sunday

16:00	Check-in begins (Front Desk - Professional Development Centre - open 24 hours)
17:30–19:30	Buffet Dinner, Sally Borden Building
20:00	Informal gathering in 2nd floor lounge, Corbett Hall Beverages and a small assortment of snacks are available on a cash honor system.

Monday

7:00–8:45	Breakfast
8:45–9:00	Introduction and Welcome by BIRS Station Manager, TCPL
9:00–9:35	Chad Creighton, **Pathway-level insights from The Cancer Genome Atlas (TCGA)**
9:35–10:10	Yoav Gilad, **Understanding gene regulation (or not)**
10:10–10:45	Stephen Montgomery, **The extent and impact of rare non-coding variants in humans**
10:45–11:05	Coffee Break, TCPL
11:05–12:05	David Haussler, **Large-scale comparative genomics for cancer research**
12:05–13:00	Lunch
13:00–14:00	Guided Tour of The Banff Centre; meet in the 2nd floor lounge, Corbett Hall
14:00–14:15	Group Photo; meet in foyer of TCPL (photograph will be taken outdoors so a jacket might be required).
14:15–14:50	Barry Taylor, **Outlier genomics drives precision oncology**
14:50–15:10	Coffee Break, TCPL
15:10–16:10	Scott Boyd, **Monitoring Human Lymphocyte Populations with High-Throughput DNA Sequencing**
16:10–16:45	Benjamin Haibe-Kains, **Prediction of Drug Response in Cell Lines: Are Pharmacogenomic Datasets Consistent?**
17:30–19:30	Dinner

Tuesday

7:00–9:00	Breakfast
9:00–9:35	Pei Wang, **Regularized multivariate regression approaches for integrative genomic analysis**
9:35–10:10	Ronglai Shen, **Pattern discovery and cancer gene identification in integrated cancer genomic data**
10:10–10:45	Sunduz Keles, **Integrative analysis of *-seq datasets for a comprehensive understanding of regulatory roles of repetitive regions**
10:45–11:05	Coffee Break, TCPL
11:05–12:05	Hongyu Zhao, **Joint analysis of expression profiles from multiple cancers to identify microRNA-gene interactions**
12:05–13:25	Lunch
13:25–14:00	Christina Kendziorski, **Latent Dirichlet allocation models to enable personalized genomic medicine**
14:00–14:35	Ingo Ruczinski, **Sequencing family members to detect disease risk variants**
14:35–15:10	Venkat Seshan, **To adjust or not to adjust: the design and analysis of an epidemiologic study**
15:10–15:30	Coffee Break, TCPL
15:30–16:30	Colin Begg, **Use of Tumor Mutational Profiles to Infer Etiologic Heterogeneity of Cancers**
17:30–19:30	Dinner
20:00	Posters and Software Demo, TCPL

Wednesday

7:00–9:00	Breakfast
9:00–9:35	Pierre Neuvial, **Improved performance evaluation of DNA copy number analysis methods in cancer studies**
9:35–10:10	Laurent Jacob, **Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed**
10:10–10:45	Roger Peng, **Reproducible Research with Evidence-based Data Analysis**
10:45–11:05	Coffee Break, TCPL
11:05–12:05	Keith Baggerly, **When is Reproducibility an Ethical Issue? Genomics, Personalized Medicine, and Human Error**
12:05–13:00	Lunch
13:00–19:00	Hiking
17:30–19:30	Dinner Served
20:00	Conference Dinner, Place TBA

Thursday

7:00–9:00	Breakfast
9:00–9:35	Kasper Hansen, **A genome-wide look at DNA methylation**
9:35–10:10	Wolfgang Huber, **Differential analysis of count data from high-throughput sequencing**
10:10–10:45	Mark Segal, **Reproducibility of 3D chromatin configuration reconstructions**
10:45–11:05	Coffee Break, TCPL
11:05–12:05	Jeff Leek, **Statistical processes for facilitating personalized medicine**
12:05–13:45	Lunch
13:45–14:20	Anshul Kundaje, **Learning long-range regulatory interactions and unified gene regulation programs in diverse human cell-types**
14:20–14:55	Simon Gravel, **Personal genomics of the Mestizos**
14:55–15:15	Coffee Break, TCPL
15:15–16:15	X. Shirley Liu, **Integrating sequencing and microarray data to identify novel functions of epigenetic regulators in cancer**
17:30–19:30	Dinner
20:00	Pub evaluation led by Stephen Montgomery

Friday

7:00–8:30	Breakfast
8:30–9:05	Dave Stephens, **Statistical modeling and computation for methylation profiles in the BLK gene region**
9:05–9:45	Noah Simon, **Estimating Many Effect-sizes Bayesian Estimation as a Frequentist**
9:45–10:10	Alexis Battle, **Characterizing the genetic basis of transcriptome diversity through RNA-sequencing**
10:10–10:45	Davide Risso, **The role of spike-in standards in the normalization of RNA-Seq**
11:30–13:30	Lunch
Checkout by 12 noon.	

** 5-day workshop participants are welcome to use BIRS facilities (BIRS Coffee Lounge, TCPL and Reading Room) until 3 pm on Friday, although participants are still required to checkout of the guest rooms by 12 noon. **

13w5083: Statistical Data Integration Challenges in Computational Biology: Regulatory Networks and Personalized Medicine
Aug 11 - Aug 16, 2013

ABSTRACTS
(in alphabetic order by speaker surname)

Speaker: **Keith Baggerly** (MD Anderson Cancer Center)
Title: *When is Reproducibility an Ethical Issue? Genomics, Personalized Medicine, and Human Error*
Abstract: Modern high-throughput biological assays let us ask detailed questions about how diseases operate, and promise to let us personalize therapy. Careful data processing is essential, because our intuition about what the answers should look like is very poor when we have to juggle thousands of things at once. When documentation of such processing is absent, we must apply forensic bioinformatics to work from the raw data and reported results to infer what the methods must have been. We will present several case studies where simple errors may have put patients at risk. This work has been covered in both the scientific and lay press, and has prompted several journals to revisit the types of information that must accompany publications. We discuss steps we take to avoid such errors, and lessons that can be applied to large data sets more broadly.

Speaker: **Alexis Battle** (Stanford University)
Title: *Characterizing the genetic basis of transcriptome diversity through RNA-sequencing*
Abstract: Understanding the consequences of regulatory variation in the human genome remains a major challenge, with important implications for understanding gene regulation and interpreting the many disease-risk variants that fall outside of protein-coding regions. Here, we provide a direct window into the regulatory consequences of genetic variation by sequencing RNA from 922 genotyped individuals. We present a comprehensive description of the distribution of regulatory variation by the specific expression phenotypes altered, the properties of affected genes, and the genomic characteristics of regulatory variants. We detect variants influencing expression of over ten thousand genes, and through the enhanced resolution offered by RNA-sequencing, we identify thousands of variants associated with specific phenotypes including splicing and allelic expression. Evaluating the effects of both long-range intra- chromosomal and trans (cross-chromosomal) regulation, we observe modularity in the regulatory network, with three-dimensional chromosomal configuration playing a particular role in regulatory modules within each chromosome. We also observe a significant depletion of regulatory variants affecting central and critical genes, along with a trend of reduced effect sizes as variant frequency increases, providing evidence that purifying selection and buffering have limited the deleterious impact of regulatory variation on the cell. Further, generalizing beyond observed variants, we have analyzed the genomic properties of variants affecting both expression and splicing, and developed a Bayesian model to predict regulatory consequences of novel variants, applicable to the interpretation of individual genomes and disease studies. Finally, this cohort was interviewed extensively to record medical, behavioral, and environmental variables, offering an opportunity to study their effects at a large scale. We have explored the impact of these environmental factors on transcriptional phenotypes, in addition to their relationship with regulatory variation, observing broad changes correlated with time of day, substance use, and medication, including changes in pathways relevant to disease risk. Together, these results represent a critical step toward characterizing the complete landscape of human regulatory variation.

Speaker: **Colin Begg** (Memorial Sloan-Kettering Cancer Center)
Title: *Use of Tumor Mutational Profiles to Infer Etiologic Heterogeneity of Cancers*
Abstract: Cancer has traditionally been studied using the disease site of origin as the organizing framework. Recent advances in molecular genetics have begun to challenge this taxonomy, as detailed molecular

profiling of tumors has led to discoveries of subsets of tumors that possess distinct clinical and biological characteristics. Increasingly investigators are examining whether sub-types defined by molecular or other tumor characteristics have distinct etiologies. To date, research in this field has typically involved the comparison of individual risk factors between tumors classified on the basis of candidate tumor characteristics or candidate sub-types. In this talk a more general, conceptual methodologic framework is presented, with a view to providing formal strategies for designing and analyzing epidemiologic studies to investigate etiologic heterogeneity. A unitary measure of etiologic heterogeneity is proposed that can be used to define quantitatively the degree of heterogeneity exhibited by a set of candidate tumor sub-types. It can be shown that overall risk predictability increases monotonically with etiologic heterogeneity. Candidate classification systems can be compared with respect to this measure to identify sets of sub-types with high degrees of heterogeneity. Data from case-control studies of breast cancer will be used to illustrate the ideas and corresponding analytic methods. It can also be shown that molecular profiles of double primary malignancies are uniquely informative for investigating this topic. The investigative strategy provides a structured approach to investigating the relationship between germ-line and somatic mutational profiles.

Speaker: **Scott Boyd** (Stanford University)
Title: *Monitoring Human Lymphocyte Populations with High-Throughput DNA Sequencing*
Abstract: Next-generation DNA sequencing (NGS) of immunoglobulin or T cell receptor gene rearrangements provide a new method for evaluating the diversity, clonality, and function of a patients B cell or T cell populations in immune system responses to vaccination or pathogen exposure. In addition, abnormal lymphocyte populations are a feature of allergic and autoimmune disorders, and can be detected and tracked using this methodology. Similarly, the immune receptor gene rearrangements in a previously diagnosed lymphoid malignancy represent highly specific tumor markers that can be used to monitor for relapse. I will discuss experimental designs and data analysis approaches that increase the interpretability and value of these complex data sets in human clinical studies. As one example, measurement of B cell clonal expansions in the blood following influenza vaccination provides an early and predictive metric of whether or not an individual will seroconvert and increase virus-specific antibody titers. Deep sequencing data identifying expanded B cell clonal lineages following vaccination correlate well with the results of single cell flow cytometric sorting and recombinant antibody synthesis identifying influenza-specific plasmablasts, supporting the biological relevance of overall B cell repertoire monitoring. Strikingly, B cell responses to influenza vaccination in different people show a prominent family of influenza-specific IgH rearrangements which are different at the DNA sequence level but highly similar at the protein level, indicating that convergent selection of antibodies to specific antigens is a common feature of human immune responses.

Speaker: **Chad Creighton** (Baylor College of Medicine)
Title: *Pathway-level insights from The Cancer Genome Atlas (TCGA)*
Abstract: Sequencing and microarray-based technologies are generating large amounts of high quality molecular data. A mandate of The Cancer Genome Atlas (TCGA) has been to collect and make available comprehensive genomic data sets on human cancers, representing multiple levels of data (DNA mutation, DNA copy, DNA methylation, mRNA, miRNA, and protein). This presentation will focus on integrative analyses of TCGA datasets, towards the goal of providing a more complete view of pathway deregulation in cancer.

Speaker: **Yoav Gilad** (University of Chicago)
Title :*Understanding gene regulation (or not)*
Abstract: Histone modifications are important markers of function and chromatin state, yet the DNA elements that direct them to specific locations in the genome are poorly understood. Here we use the genetic variation in Yoruba lymphoblas- toid cell lines as a natural experiment to identify genetic differences that af- fect histone marks and to better understand their relationship with transcrip- tional regulation. Across the genome, we identified hundreds of quantitative trait loci that impact histone modification or RNA polymerase (PolII) occu- pancy. In many cases the same variant is associated with quantitative

changes in multiple histone marks and PolII, as well as in DNaseI sensitivity and nu- cleosome positioning, indicating that these molecular phenotypes often share a single underlying genetic cause. Polymorphisms in some transcription factor binding sites cause differences in local histone modification and we identify specific transcription factors whose binding leads to histone modification in lymphoblastoid cells. Finally, we find that variants that impact chromatin at distal regulatory sites frequently also direct changes in chromatin and gene ex- pression at associated promoters. In summary, the class of variants identified here generate coordinated changes in chromatin both locally and sometimes at distant locations, frequently drive changes in gene expression, and likely play an important role in the genetics of complex traits.

Speaker: **Simon Gravel** (McGill University)
Title: *Personal genomics of the Mestizos*
Abstract: There is great scientific and popular interest in understanding the genetic history of populations in the Americas. We wish to understand when different regions of the continent were inhabited, where settlers came from, and how current inhabitants relate genetically to earlier populations. Because of the important migrations that marked the history of the continent over the last few hundred years, many individuals derive ancestry from multiple continental groups, predominantly African, European, and Native American. To develop personalized medicine for such diverse populations, we must understand how this recent admixture of previously isolated populations impacted individual genomes.

I will focus on the Mestizos of Latin America and the Caribbean, and discuss how we can integrate multiple genetic datasets to learn about the historical processes that led to the observed diversity within these populations, and within individuals. I will present methods to overcome statistical challenges caused by biases in high-throughput sequence data, propose precise estimates of the human mutation rate, discuss the most likely origins of the Taino people, and speculate on consequences for personalized medicine.

Speaker: **Kasper Hansen** (Johns Hopkins University)
Title: *A genome-wide look at DNA methylation*
Abstract: DNA methylation is an important epigenetic mark in mammalian cells, implicated in tissue differentiation and cancer. Whole-genome bisulfite sequencing (WGBS) is a recent technological breakthrough which has, for the first time, enabled true genome-wide measurement of this epigenetic mark. We discuss our recent work on analyzing this type of data and discuss changes in DNA methylation associated with carcinogenesis as well as global differences between tissues.

Speaker: **David Haussler** (UC Santa Cruz)
Title: *Large-scale comparative genomics for cancer research*
Abstract: UCSC has built the Cancer Genomics Hub (CGHub) for the US National Cancer Institute, designed to hold up to 5 petabytes of research genomics data (up to 50,000 whole genomes), including data for all major NCI projects. In its first year it has served 6 petabytes of data to more than 100 research labs. Cancer is exceedingly complex, with thousands of subtypes involving an immense number of different combinations of mutations. The only way we will understand it is to gather together DNA data from many thousands of cancer genomes so that we have the statistical power to distinguish between recurring combinations of mutations that drive cancer progression and "passenger" mutations that occur by random chance. Currently, with the exception of a few projects such as ICGC and TCGA, most cancer genomics research is taking place in research silos, with little opportunity for data sharing. If this trend continues, we lose an incredible opportunity. Soon cancer genome sequencing will be widespread in clinical practice, making it possible in principle to study as many as a million cancer genomes. For these data to also have impact on understanding cancer, we must begin soon to move data into a global cloud storage and computing system, and design mechanisms that allow clinical data to be used in research with appropriate patient consent. A global alliance for sharing genomic and clinical data is emerging to address this problem. This is an opportunity we cannot turn away from, but involves both social and technical challenges.

Reference: http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html

Speaker: **Wolfgang Huber** (EMBL)
Title: *Differential analysis of count data from high-throughput sequencing*
Abstract: Many applications of high throughput sequencing require statistical inference based on count data. Mapped reads are often summarised by counting their overlaps with genomic features of interest (genes, exons, binding regions) in samples from different experimental conditions. Applications include differential gene expression, differential exon usage, HiC, ChIP-Seq, CLIP-Seq; similar counting problems are also posed in proteomics.

In this talk, I will describe some of our recent work on the use of generalised linear models of the Negative Binomial family for this task, in particular shrinkage estimation of treatment effects and dispersion parameters in the small sample situation, and robustness to outlier data. Accompanying software is available in the DESeq2 package in Bioconductor. I will also briefly present an application to the detection of evolutionarily conserved patterns of tissue dependent exon usage.

Speaker: **Laurent Jacob** (UC Berkeley)
Title: *Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed*
Abstract: When dealing with large scale gene expression studies, observations are commonly contaminated by sources of unwanted variation such as platforms or batches. Not taking this unwanted variation into account when analyzing the data can lead to spurious associations and to missing important signals. When the analysis is unsupervised, e.g. when the goal is to cluster the samples or to build a corrected version of the dataset — as opposed to the study of an observed factor of interest — taking unwanted variation into account can become a difficult task. The unwanted factors may be correlated with the unobserved factor of interest, so that correcting for the former can remove the effect of the latter if not done carefully. We show how negative control genes and replicate samples can be used to estimate unwanted variation in gene expression, and discuss how this information can be used to correct the expression data or build estimators for unsupervised problems. The proposed methods are then evaluated on three gene expression datasets. They generally manage to remove unwanted variation without losing the signal of interest and compare favorably to state of the art corrections.

Speaker: **Benjamin Haibe-Kains** (Institut de Recherches Cliniques de Montral)
Title: *Prediction of Drug Response in Cell Lines: Are Pharmacogenomic Datasets Consistent?*
Abstract: Cancer cell line studies have long been used to test the efficacy of therapeutic agents and to explore the genomic factors predictive of response. Several large-scale pharmacogenomic studies were published recently; each assayed a panel of several hundred cancer cell lines for gene expression, copy number, genome sequence, and pharmacological response to multiple anti-cancer drugs. The resulting datasets present a unique opportunity to characterize mechanisms associated with drug response. In this talk I will show that in comparing these datasets high-throughput genomic data are well correlated, however the measured pharmacologic response to drugs is highly discordant. The poor correspondence is surprising as these studies assessed drug response using common estimators: the IC50 (concentration at which the drug inhibited 50% of the maximal cellular growth), and the AUC (area under the activity curve measuring dose response). For response for drugs screened in these studies, only one drug had a correlation coefficient between studies greater than 0.6; these results are also reflected in the gene-drug associations where inconsistent results were found. However, the results improved when we assessed pathway-drug correspondence (a few drugs with a correlation greater than 0.6), suggesting that analyzing the genomic basis of drug response at the pathway level may yield greater consistency between studies. The discrepancy in pharmacologic response in well-controlled experiments makes drawing firm conclusions from them very difficult and has potential implications for using these outcome measures to assess gene-drug relationships or select potential anti-cancer drugs.

Speaker: **Sunduz Keles** (University of Wisconsin)
Title: *Integrative analysis of *-seq datasets for a comprehensive understanding of regulatory roles of repetitive regions*
Abstract: A fundamental question in molecular biology is how cell type specific gene expression programs are established and maintained through gene regulation. Main drivers of cell-specific gene expression are regulatory elements (e.g., promoters, transcription factor (TF) binding sites, chromatin/epigenomic marks, enhancers, silencers). Identifying genomic locations of these elements and unraveling exactly how they control gene expression in different cell types has been a major challenge. The ENCODE projects have generated exceedingly large amounts of genomic data towards this end. A formidable impediment to comprehensively understanding of these ENCODE data is the lack of statistical and computational methods required to identify functional elements in repetitive regions of genomes. Although next generation sequencing (NGS) technologies, embraced by the ENCODE projects, are enabling interrogation of genomes in an unbiased manner, the data analysis efforts by the ENCODE projects have thus far focused on mappable regions with unique sequence contents. This is especially true for the analysis of ChIP-seq data in which all ENCODE-adapted methods discard reads that map to multiple locations (multi-reads). This is a highly critical barrier to the advancement of ENCODE data because significant fractions of complex genomes are composed of repetitive regions; strikingly, more than half of the human genome is repetitive.

We present a unified statistical model for utilizing multi-reads in *-seq datasets (ChIP-, DNase-, and FAIRE-seq) with either diffused or a combination of diffused and point source enrichment patterns. Our model efficiently integrates multiple *-seq datasets and significantly advances multi-read analysis of ENCODE and related datasets.

Speaker: **Christina Kendziorski** (University of Wisconsin)
Title: *Latent Dirichlet allocation models to enable personalized genomic medicine*
Abstract: Genomic based studies of disease now involve highly diverse types of data collected on large groups of patients. A major challenge facing statistical scientists is how best to combine the data, extract important features, and comprehensively characterize the ways in which the features affect an individual's disease course and likelihood of response to treatment. In this talk, I will review methods that we have developed to address this challenge. Drawing an analogy from information retrieval, we consider each patient as a document, and data on each patient as text; and we extend the latent Dirichlet allocation model (LDA) to our application domain. Documents are constructed using data from multiple clinical sources and high-throughput assays. By introducing priors that accommodate known structure among subsets of genomic variables, the LDA based model allows for discovery of distinct topics across the patient population (collections of genomic aberrations, clinical variables, and treatments) as well as determination of patient-specific mixtures over topics. Further model extensions provide for survival-related responses to supervise model fit. The approach facilitates data integration across multiple platforms and scales to enable powerful patient-specific inference, as demonstrated in studies of cancer from the cancer genome atlas (TCGA) project.

Speaker: **Anshyul Kundaje** (MIT, Broad Institue and Stanford University)
Title: *Learning long-range regulatory interactions and unified gene regulation programs in diverse human cell-types*
Abstract: In multicellular organisms, epigenetic information is a key enabler of dynamic regulatory regions shaping the identity of each cell. This information is encoded in distinct combinations of epigenetic modifications defining chromatin states specific to different types of functional elements such as promoters, enhancers, transcribed elements and repressed domains. First, we used multivariate Hidden Markov models to jointly learn the largest collection of gene-proximal and distal regulatory elements from histone modification ChIP-seq data in 120 diverse human cell-types from the Roadmap Epigenomics and ENCODE consortia. Next, we developed a novel probabilistic model based on Latent Dirichlet Allocation to computationally infer putative target genes of cell-type specific enhancers based on the associated chro-

matin state and gene expression dynamics. We automatically discovered co-activated transcriptional and enhancer modules that are strongly enriched for lineage specific functional annotations and biochemical pathways; as well as the complex, non-linear, cell-type specific interactions between these modules. The resulting model showed significant improvements in prediction of transcriptional responses compared to simple correlation-based linking methods. The accuracy and cell-type specificity of our predicted links were further validated by experimental ChIA-PET chromatin interaction data in matching cell-types and eQTL predictions. Finally, we developed a novel ensemble learning framework based on Boosting algorithms to learn context-specific predictive models of gene regulation by integrating DNA binding sequence motifs of a comprehensive collection of transcription factors with gene expression data. We dissect these models to highlight cell-type specific regulatory elements, transcription factors and pathways. Together, these analyses provide a unified, multi-faceted view of dynamic gene regulation in humans.

Speaker: **Jeff Leek** (Johns Hopkins University)
Title: *Statistical processes for facilitating personalized medicine*
Abstract: The promise of personalized medicine has been tempered by high-profile errors in the application of genomic biomarkers.Time permitting I will discuss statistical processes for facilitating personalized medicine: single sample normalization and artifact correction, self-normalizing biomarkers for data integration, locked down biomarker development, and interactive visualization. I will illustrate these ideas with a case study in genomic biomarker development.

Speaker: **Xiaole Shirley Liu** (Harvard University and Dana-Farber Cancer Institute)
Title: *Integrating sequencing and microarray data to identify novel functions of epigenetic regulators in cancer*
Abstract: There have been growing appreciation of the role of epigenetic alteration in tumorigenesis and cancer progression. The integration of recent genomic techniques and massive public data is a useful approach to study epigenetic gene regulation in cancer. To this end, we use chromatin dynamics from ChIP-seq and DNase-seq profiles identify driving transcription factors in cancer progression and find novel functions of chromatin regulators. We also integrate large scale tumor expression data to identify novel lncRNAs with oncogenic functions and unkown partners that mediate the novel function of chromatin regulators.

Speaker: **Stephen Montgomery** (Stanford University)
Title: *The extent and impact of rare non-coding variants in humans*
Abstract: Recent and rapid human population expansion has led to an excess of rare genetic variants that are expected to contribute to an individual's genetic burden of disease risk. To date, large-scale exome sequencing studies have highlighted the abundance of rare and deleterious variants within protein-coding sequences. However, in addition to protein-coding variants, rare non-coding variants are likely to be enriched in functional consequences. I will discuss our effort to characterize the impact of rare non-coding variation in a large human family and an isolated population. Further, I will discuss our effort to understand the systemic (multi-tissue) impact of highly-deleterious coding variants (or variants of unknown significance). To address this, we have developed a multiplex, microfluidics-based method for assessing the interaction of regulatory variation on deleterious protein-coding alleles identified through exome sequencing. Finally, I will discuss our efforts to understand rare and common regulatory variants underlying complex disease and will highlight new analytical approaches for the analysis of RNA sequencing data that we have applied to understanding cardiovascular and lung disease.

Speaker: **Pierre Neuvial** (CNRS and University of Evry)
Title: *Improved performance evaluation of DNA copy number analysis methods in cancer studies*
Abstract: Changes in DNA copy numbers are a hallmark of cancer cells. Therefore, the accurate detection and interpretation of such changes are two important steps toward improved diagnosis and treatment. The analysis of copy number profiles measured from high-throughput technologies such as SNP microarray

and DNAseq data raises a number of statistical and bioinformatic challenges. Evaluating existing analysis methods is particularly challenging in the absence of gold standard data sets.

We have designed and implemented a framework to generate realistic DNA copy number profiles of cancer samples with known parent-specific copy-number state. This talk illustrates some of the benefits of this approach in a practical use case: a comparison study between methods for segmenting SNP array data into regions of constant parent-specific copy number. This study helps identifying the pros and cons of the compared methods in terms of biologically informative parameters, such as the signal length, the number of breakpoints, the fraction of tumor cells in the sample, or the chip type.

Speaker: **Roger Peng** (Johns Hopkins University)
Title: *Reproducible Research with Evidence-based Data Analysis*
Abstract: tatistical software is plentiful today, with new procedures and algorithms constantly being developed, implemented, and optimized. Traditional statistical software tends to focus on solving a relatively self-contained task, often something that is a single piece of a much larger data analysis. Data analysts are subsequently free to combine the various pieces of statistical software out there in any number of combinations to analyze their data as they see fit. Hence, the number of "degrees of freedom" given to the analyst in most situations is enormous. But why is this so? Statistical software is typically written with a specific interface where certain parameters are modifiable by the user but most others are not. A similar approach needs to be taken at the much higher level of the entire data analysis. Data analysis pipelines can be built using pre-determined combinations of procedures that have been chosen based on sound statistical evidence of their fitness or superiority. Such analysis pipelines—"transparent boxes"—would have relatively few options available to the user and would be deterministic in their operation. We call these general pipelines "deterministic statistical machines" and present an example of one in the context of air pollution epidemiology. We further discuss how these machines can be used to encourage reproducible research in biomedical science.

Speaker: **Davide Risso** (UC Berkeley)
Title: *The role of spike-in standards in the normalization of RNA-Seq*
Abstract: Normalization of RNA-Seq data has proven to be an essential step to ensure accurate inference of expression levels, by correcting for sequencing depth and other distributional differences within and between replicate samples. Recently, the External RNA Control Consortium (ERCC) has developed a set of 92 synthetic spike-in standards that are now commercially available and relatively easy to add to a standard library preparation. In this talk, we evaluate the performance of the ERCC spike-ins and we use them as controls to compare different normalization strategies. Moreover, we investigate the possibility of directly using spike-in expression measures to normalize the data. We show that although spike-in standards are a useful resource for evaluating accuracy in RNA-Seq experiments, their expression measures are not stable enough to be used to estimate even a global scaling parameter to normalize the data.

We propose a novel normalization strategy that aims at removing unwanted variation from the data by performing a factor analysis on a suitable set of control genes and that can exploit spike-in controls when they are present in the library, without relying exclusively on them. Our novel approach leads to more accurate estimates of expression fold-changes and tests for differential expression, compared with state-of-the-art normalization methods.

Speaker: **Ingo Ruczinski** (Johns Hopkins University)
Title: *Sequencing family members to detect disease risk variants*
Abstract: We present some new statistical methods and software to detect disease risk variants, sequencing affected only or affected and non-affected individuals in families. Examples are mostly drawn from a study of oral clefts with probands of Asian and European descent.

Speaker: **Mark Segal** (UCSF)
Title: *Reproducibility of 3D chromatin configuration reconstructions*
Abstract: It is widely recognized that the three dimensional (3D) architecture of eukaryotic chromatin plays critical roles in nuclear and cellular function. However, until a few years ago, observing 3D structure at even modest resolutions was problematic, because genomes are highly condensed and assays were low-throughput. Recently devised high-throughput molecular techniques are changing this situation. Notably, the development of chromatin conformation capture (CCC) assays has enabled elicitation of contacts: spatially close chromosomal loci. These techniques have provided insight into chromatin organization at unprecedented resolutions, and permitted exploration of the downstream influence of such organization on a variety of biological processes, including gene regulation and cancer-driving gene fusions. Accordingly, obtaining high resolution 3D reconstructions of genome architecture is a compelling biological quest. However, most analysis of CCC data has focussed on the one dimensional contact level, with appreciably less effort directed toward evaluating accuracy and reproducibility of 3D reconstructions, and deploying such structures to analyze consequent biological processes. Questions of accuracy must be addressed experimentally. However, questions of reproducibility can be addressed statistically. After describing and applying a constrained optimization technique to reconstruct chromatin configurations for a number of closely related yeast datasets we assess the reproducibility thereof using three relevant metrics that measure the distance between 3D configurations. The first of these, Procrustes fitting, measures configuration closeness after applying reflection, rotation, translation and scaling based alignment of the structures. The other two, congruence among distance matrices and distance differencing, base comparisons on the within-configuration inter-point distance matrix. Inferential results for these metrics rely on suitable resampling schemes. Preliminary findings indicate that distance matrix based approaches are preferable to Procrustes analysis, not because of the metrics per se but rather on account of attendant inferential (permutation) schemes.

It has recently been emphasized that the use of constrained optimization approaches to 3D architecture reconstruction, as employed here, can be prone to becoming trapped in local minima. Our methods of reproducibility assessment provide a means for comparing 3D reconstruction solutions so that we can discern between local and global optima by contrasting solutions under perturbed inputs.

Speaker: **Venkat Seshan** (Memorial Sloan-Kettering Cancer Center)
Title: *To adjust or not to adjust: the design and analysis of an epidemiologic study*
Abstract: Women who survive their first primary breast cancer are at an increased risk of developing a second primary cancer in their contralateral breast. In a recent study Reiner et al (2012) showed that younger age at diagnosis, family history of breast cancer and degree of relationship to affected relative are associated with the risk of contralateral breast cancer. Of interest is whether the molecular characteristics of the primary tumor can give a better predictor. Since the molecular characteristics such as expression, methylation etc. are high-dimensional feature selection is an important step. The risk score derived from the molecular characteristics can only be considered useful if it adds value to the existing risk predictors. An important issue is whether the features used in the molecular risk predictor should be selected adjusting upfront for known risk factors or selected unadjusted and adjusted post development of score. We will compare the performance of the two approaches using simulations in a simple logistic regression framework.

Speaker: **Ronglai Shen** (Memorial Sloan-Kettering Cancer Center)
Title: *Pattern discovery and cancer gene identification in integrated cancer genomic data*
Abstract: Large-scale integrated cancer genome characterization efforts including the cancer genome atlas have created unprecedented opportunities to study cancer biology in the context of knowing the entire catalog of genetic alterations. A clinically important challenge is to discover cancer subtypes and their molecular drivers in a comprehensive genetic context. Curtis et al. [Nature (2012) 486(7403):346352] has recently shown that integrative clustering of copy number and gene expression in 2,000 breast tumors reveals novel subgroups beyond the classic expression subtypes that show distinct clinical outcomes. To extend the scope of integrative analysis for the inclusion of somatic mutation data by massively parallel

sequencing, we propose a framework for joint modeling of discrete and continuous variables that arise from integrated genomic, epigenomic, and transcriptomic profiling. The core idea is motivated by the hypothesis that diverse molecular phenotypes can be predicted by a set of orthogonal latent variables that represent distinct molecular drivers, and thus can reveal tumor subgroups of biological and clinical importance. To identify genomic features that contribute most to the biological variation and thus have direct relevance for characterizing the molecular subgroups, we apply a penalized likelihood approach. We show application of the method to the TCGA pan-cancer cohort with whole-exome DNA sequencing, SNP6.0 array, mRNA sequencing data in 3,000 patient samples spanning 12 cancer types.

Speaker: **Noah Simon** (Stanford University)
Title: *Estimating Many Effect-sizes Bayesian Estimation as a Frequentist*
Abstract: With the advent of high-throughput technologies, the multiple testing problem has become pervasive in the analysis of biomedical and genomic data. Much attention has been devoted to this issue, and in many cases the field has developed good solutions. However, there is an equally important but more overlooked problem: estimating the effect-sizes of the significant features. The standard for estimating effect-sizes in high-throughput problems has been the empirical bayes methods of Robbins (Robbins [1951] and others), which was recently brought back into the limelight by Efron (Efron [2010] and others). Combined with new developments in flexible density estimation these methods perform somewhat astoundingly well. Unfortunately these methods are still not widely used in practice — they are often seen as unintuitive, or clever but impractical (eg. James-Stein). In this talk I will give a simple intuitive reformulation of the frequentist effect-size estimation problem. This reformulation will lead directly to the empirical bayes approach. Along the way I will include a number of simulated and real data examples. At the end I will discuss some unresolved issues and future research directions.

B. Efron. Large Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge, 2010.
H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems. Stanford-Berkeley joint symposium, 1951.

Speaker: **David Stephens** (McGill University)
Title: *Statistical modeling and computation for methylation profiles in the BLK gene region*
Abstract: I will discuss some work in progress on statistical methods for extracting patterns from methylation profiles, particularly sequencing data, in an example based on the BLK gene region. The assay involves a high throughput technology that can output strand-specific reads and methylation proportions over relatively large regions. Our approach uses hidden Markov models, and I will discuss discrete and continuous latent representations of the methylation patterns.

This is joint work with Asad Haris, Celia Greenwood and Aurelie Labbe.

Speaker: **Barry Taylor** (UCSF)
Title: *Outlier genomics drives precision oncology*
Abstract: Curative therapy for patients with advanced-stage solid tumors remains elusive. Even with the much-heralded advent of targeted inhibitors of oncogenic signaling pathways, drug resistance and disease progression occur in essentially all patients. Indeed, little is known about the molecular genetic basis of exceptional and curative responses to cancer therapy. We have begun to investigate, with whole-genome sequencing and associated approaches, the genetic basis of complete and durable responses to both targeted and systemic anti-cancer therapies, an outlier phenotype. Here, we discuss early successes and challenges as well as the opportunities for variant interpretation in clinical specimens from patients with established phenotypes. These studies have revealed not only individual sensitizing mutations, but also synergistically acting genetic interactions and the contribution of tumor clonality to the durability of treatment response. Together, these data have yielded unprecedented insights into the molecular genetic basis of exceptional

responses, leading to the discovery of (i) novel pathway biology, (ii) previously occult biomarkers of clinical benefit, and (iii) rational polytherapeutic strategies to interdict in a manner that extends such profound, life-altering activity in molecular defined populations.

Speaker: **Pei Wang** (Fred Hutchinson Cancer Research Center)
Title: *Regularized multivariate regression approaches for integrative genomic analysis*
Abstract: Understanding expression quantitative trait loci (eQTL) provides important clues to genetic basis of gene expression regulation. In this talk, we introduce a new statistical method, GroupRemMap, for identifying eQTLs. We model the dependent relationship between gene expression and single nucleotide variants (SNVs) through a multivariate linear regression model, in which gene expression levels are treated as outcomes and SNV genotypes are treated as predictors. To handle the high-dimensionality as well as to incorpo- rate the intrinsic group structure of SNV data, we introduce a new regularization scheme to (1) control the overall sparsity of the model; (2) encourage the group selection of SNVs from the same gene; and (3) facilitate the detection of trans-hub-eQTLs. We apply the proposed method to the colorectal and breast cancer data sets from the cancer genome atlas (TCGA), and identify several biologically interesting eQTLs. These findings could potentially inform the underlying biological processes of cancers and generate hypotheses for future studies.

Speaker: **Hongyu Zhao** (Yale University)
Title: *Joint analysis of expression profiles from multiple cancers to identify microRNA-gene interactions*
Abstract: MicroRNAs (miRNAs) play a crucial role in tumorigenesis and development through their effects on target genes. The characterization of miRNA-gene interactions will lead to a better understanding of cancer mechanisms. Many computational methods have been developed to infer miRNA targets with/without expression data. Since expression data sets are in general limited in size, most existing methods concatenate datasets from multiple studies to form one aggregated dataset to increase sample size and power. However, such simple aggregation analysis results in identifying miRNA-geneinteractions that are mostly common across data sets, whereas specific interactions may be missed by these methods. Recent releases of The Cancer Genome Atlas (TCGA) data provide paired expression profiling of miRNAs and genes in multiple tumors with sufficiently large sample size. To study both common and cancer specific interactions, it is desirable to develop a method that can jointly analyze multiple cancers to study miRNA-gene interactions without combining all the data into one single data set. In this presentation, we describe a novel statistical method to jointly analyze expression profiles from multiple cancers to identify miRNA-gene interactions that are both common across cancers and specific to certain cancers. The benefit of this joint analysis approach is demonstrated by both simulation studies and real data analysis of TCGA datasets. Compared to simple aggregate analysis or single sample analysis, our method can effectively use the shared information among different but related cancers to improve the identification of miRNA-gene interactions. Another useful property of our method is that it can estimate similarity among cancers through their shared miRNA-gene interactions. This is joint work with Xiaowei Chen and Frank Slack.