BIRS/Banff 15w5142 - Statistical and Computational Challenges In Bridging Functional Genomics, Epigenomics, Molecular QTLs, and Disease Genetics

SUNDAY AUGUST 2, 2015 - Arrival

5:30pm-7:30pm: **Dinner at the Sally Borden Building**

7:30pm-8:30pm: Informal discussion & meeting planning @ **BIRS Lounge**

MONDAY AUGUST 3, 2015

Monday Morning - RNA-Seq

(Breakfast: 7am-9am)

Monday 9:15am-9:50am

Laurent Jacob

Efficient RNA isoform identification and quantification from RNA-Seq data with network flows

Several state-of-the-art methods for isoform identification and quantification are based on l1-regularized regression, such as the Lasso. However, explicitly listing the - possibly exponentially - large set of candidate transcripts is intractable for genes with many exons. For this reason, existing approaches using the l1-penalty are either restricted to genes with few exons or only run the regression algorithm on a small set of preselected isoforms. We introduce a new technique called FlipFlop, which can efficiently tackle the sparse estimation problem on the full set of candidate isoforms by using network flow optimization. Our technique removes the need of a preselection step, leading to better isoform identification while keeping a low computational cost. Experiments with synthetic and real RNA-Seq data confirm that our approach is more accurate than alternative methods and one of the fastest available.

laurent.jacob@gmail.com

Monday 9:50am-10:25am

Jeffrey Leek

Statistical analysis of RNA-seq data at different scales

RNA-seq is now the primary technology used to measure transcriptional abundance. The analysis of RNA-seq data can be done at multiple levels (genes, regions, or transcripts) and at multiple scales (small experiments or large population cohorts). I will discuss statistical challenges in developing and applying software for the analysis of RNA-seq data at multiple scales including reproducibility, statistical power, trust in genomic annotations, and detection and removal of artifacts. These issues are critical in the analysis of data from genomic experiments in general, but are particularly acute in the analysis of dynamic data from transcriptomes.

jtleek@gmail.com

Monday 10:25am-10:40am: Coffee Break

Monday 10:40am-11:15am

Rafael Irizarry

Overcoming bias and batch effects in RNAseq data

In this talk I will demonstrate the presence of bias, systematic error and unwanted variability in next generation there's sequencing. I will show the substantial downstream effects these have on downstream results and how they can lead to misleading biological conclusions. I will do this using data from the public repositories as well as our own. We will then describe some preliminary solutions to these problems.

rafa@jimmy.harvard.edu

Monday 11:15am-11:50am

Adam Olshen

Further Statistical Methods for the Analysis of Ribosome Profiling Data

During translation messenger RNA produced by transcription is decoded by ribosomes to produce specific polypeptides. Ribosome profiling, a second generation sequencing technology, was developed to measures the position and counts of ribosomes. When

combined with corresponding mRNA sequencing data, ribosome profiling data can give insights into translational efficiency. We developed the Babel framework to discover gene level changes in translational efficiency between different conditions utilizing ribosome profiling data. Here we discuss Babel and also newer statistical approaches to ribosome profiling data.
adam.olshen@ucsf.edu

Monday 11:50-12:10am

Kai Kammers (15'+5')

Genetic and transcriptomic analysis of megakaryocytes

Aggregation of platelets in the blood on ruptured or eroded atherosclerotic plaques may initiate arterial occlusions causing heart attacks, strokes, and limb ischemia. Understanding the biology of platelet aggregation is important to prevent inappropriate vascular thrombosis. GWAS studies have identified common variants associated with platelet aggregation, but because they are intronic or intergenic, it is not clear how they are linked biologically to platelet function. To examine this, we are funded to produce pluripotent stem cells (iPSCs) from people with informative genotypes, and then derive megakaryocytes (MKs), the precursor cells for anucleate platelets, from the iPSCs to determine patterns of gene transcript expression in the MKs related to specific genetic variants. To this end it is essential that the iPSC-derived MKs retain their genomic integrity during production or expansion. This was examined using three alternative measures of integrity of the MK cell lines: (1) mutation rates comparing parent cell DNA to iPSC cell DNA and onward to the differentiated MK DNA; (2) structural integrity using copy number variation (CNV) on the same; and (3) transcriptomic signatures of the derived MK cells. For the RNASeq data we extracted non-ribosomal RNA from 14 paired iPS and MK cell lines. Looking specifically for genes 'turned on' in MKs following differentiation from the iPSCs, we observed the following highly biologically relevant gene sets in the list of top 12 identified: platelet activation immune response, inflammatory response, platelet formation, and regulation of cell proliferation. Most recently, we performed extensive eQTL analyses with megakaryocytes to categorize 'functional' relevance of the GWAS-identified determinants of platelet aggregation leveraging the genotype and RNASeq data.
kai.kammers@gmail.com

Monday 12:15am-1:30pm: Lunch Break

Monday afternoon - eQTLs

Monday 1:25pm-2:00pm

Tuuli Lappalainen

Genomic imprinting across diverse human tissues

Large-scale functional genomics data sets provide opportunities not only for mapping functional genetic variants but also for systematic genome-wide analysis of diverse biological phenomena. Imprinting is an epigenetic mechanism that leads to parent-of-origin effects via imbalanced expression of the maternally and paternally inherited alleles. We have used multiple population-scale genetic and RNA-sequencing data sets, including the GTEx project pilot data, to create a tissue-specific map of imprinting in human adults. We characterized imprinting in 42 genes, including both novel and previously identified genes. Tissue specificity of imprinting is widespread, and gender-specific effects are revealed in a small number of genes in muscle. IGF2 shows maternal expression in the brain instead of the canonical paternal expression elsewhere. In summary, our systematic characterization of imprinting in adult tissues highlights variation in imprinting between genes, individuals, and tissues, and motivates future research into mechanisms of imprinting and its relevance in human disease.
tlappalainen@nygenome.org

2pm: Group photo outside the TransCanada Pipelines Pavilion

Monday 2:05pm-2:40pm

Yoav Gilad

eQTL mapping in iPSC lines

Human induced pluripotent stem cells (iPSCs) provide unprecedented potential to study multiple human cell

types from a single individual. This greatly impacts human genomics by allowing researchers to study cell type specific gene regulation and perform experiments on living samples. A necessary step in this process is the generation of a panel of iPSCs large enough to study the effects of genetic variation on gene regulation. To this end, we reprogrammed lyphoblastoid cell lines (LCLs) from 70 individuals into iPSCs. The individuals in this panel have been used extensively in previous studies, making them attractive for further studies on gene regulation. In this study, we characterized regulatory variation, by collecting gene expression data and identifying expression quantitative trait loci (eQTLs). In our hands, the number of eQTLs identified in iPSCs are comparable to those identified in somatic cell types with similar sample sizes. This result was at first surprising, given that we also observed a high degree of homogeneity in gene expression between individuals. Yet, we found that eQTL effect sizes are smaller on average in iPSCs compared to eQTLs in somatic cells. Standard errors of the estimates of effect size are also an order of magnitude lower in iPSCs. I will discuss these results in the context of statistical issues and the biology of eQTLs in iPSCs and across other tissues.
gilad@uchicago.edu

Monday 2:45pm-3:20pm

Philip Awadalla

High-coverage RNA-sequencing reveals substantial variation associated with geography, environment and endophenotypic variation

Phenotypic variation is the result of the combined effect of genetic variation with environmental influences. Gene-by-environment interactions are thought to be pervasive and may be responsible for a large fraction of the unexplained variance in heritability and disease risk. However, it has been particularly difficult to reliably identify robust gene-by-environment effects in humans. Studies mapping gene expression variation in humans have established that there is an abundant amount of inter-individual regulatory variation and that a significant fraction of this variation is heritable. Yet, a general understanding of the extent of variation of gene expression and how genetic regulatory variation is modulated by environmental factors is lacking. To systematically survey genetic, environmental and interaction effects on whole blood transcriptome, we combined whole transcriptome RNASeq profiling with whole genome genotyping on deeply endophenotyped individuals selected from over 40,000 participants in the CARTaGENE resource. We document substantial geographical variation in whole blood gene expression in this founder population that follows a south-north cline in the province of Quebec. Using haplotype-based methods on genome-wide genotyping, we detected fine-scale genetic structure within the province, and we were able to identify individuals that have migrated within the province from their ancestral region. In addition to the strong signature of geographic regional effects on gene expression, we reveal a substantial impact of environmental factors on global gene expression profiles overpowering that of genotype. Expression profiles of migrants are more similar to those of individuals presently living in the same region than to those of individuals with the same ancestry, but living in a different region. Genes involved in oxygen transport and inflammation are enriched among the differentially expressed genes between regions, suggesting an impact the highly urbanized environments on expression profiles. We also report several instances of genome-wide significant transcriptional gene-environment interactions (environmental eQTLs) that may have a clinical impact for individuals carrying specific genotypes in a given environment. These findings suggest that environmental variation can significantly alter disease genetic risk in both direct and indirect fashion and call for placing regulatory variants in the context of their geographical distribution and associated environmental exposures.
philip.awadalla@oicr.on.ca

Monday 3:20pm-3:35pm: Coffee Break

Monday 3:35pm-4:10pm

Barbara Engelhardt

Heteroskedastic linear models for functional genomics

Heteroskedasticity in linear models refers to the situation where a predictor -- here, a single nucleotide

polymorphism (SNP) -- is correlated with the residual error of a linear model fitted to that SNP and a quantitative trait response. Previous work in functional genomics has identified examples of a genotype affecting not (only) the mean of the quantitative trait, but (also) its variance, implying a heteroskedastic association. Methods to identify these variance QTLs (vQTLs) are based on ANOVA tests or two-stage linear models. Here, we develop a test for heteroskedasticity based on a Bayesian heteroskedastic linear model. We show the power of this test for identifying vQTLs for various cellular traits. We extend the model to pairs of quantitative traits to address questions about causal relationships among cellular phenotypes.
bee@princeton.edu

Monday 4:10pm-4:45pm

Shamil Sunyaev

Can we rely on eQTLs to understand GWAS peaks?

Abstract pending

Monday **4:50pm-6:20pm**

Tunnel mountain hike

Monday 6:20pm-7:30pm - Dinner

TUESDAY AUGUST 4, 2015

Tuesday Morning - GWAS

(Breakfast 7am-9am)

Tuesday 9am-9:35am

Michael Snyder

Differences among individuals and between species

We have been analyzing differences in gene regulation and expression between humans using lymphoblastoid lines and also between species. The latest results will be presented.
mpsnyder@stanford.edu

Tuesday 9:35am-10:10am

Stephen Montgomery

Rare regulatory variation in individuals, families and populations

Genome and transcriptome sequencing in population samples provides the opportunity to identify rare and causal non-coding variants. Both total expression and allele-specific expression outliers act as informative priors for identifying genes harboring impactful rare and non-coding variants. I will present analyses of large population cohort, SardiNIA and GTEx, where integrative approaches have identified individuals, families and tissues with rare non-coding variants.
smontgom@stanford.edu

Tuesday 10:10am-10:30am: Coffee Break

Tuesday 10:30am-11:05am

Jennifer Listgarten

Linear Mixed Models for Genome and Epigenome-Wide Association Studies

Understanding the genetic underpinnings of disease is important for screening, treatment, drug development, and basic biological insight. Genome-wide associations, wherein individual or sets of genetic markers are systematically scanned for association with disease are one window into disease processes. Naively, these associations can be found by use of a simple statistical test. However, a wide variety of confounders lie hidden in the data, leading to both spurious associations and missed associations if not properly addressed. These confounders include population structure, family relatedness, cell type heterogeneity, and environmental confounders. I will discuss the state-of-the art approaches (based on linear mixed models) for conducting these analyses, in which the confounders are automatically deduced, and then corrected for, by the data and model.
jennl@microsoft.com

Tuesday 11:05am-11:40am

Anna Goldenberg

Data integration, variant aggregation and combined annotation

Majority of human diseases are complex, arising due to a multitude of factors. Identifying these factors is critical to understanding diseases and improving health care, yet it is a very difficult computational problem: low

signal-to-noise ratio (only a few variants out of millions are likely to be causal), heterogeneity of reasons (e.g. coding, regulatory, epigenetic), epistasis (gene interaction patterns), etc. We propose to combine two mostly complementary data sources: coding variants and gene expression. These two data sources are responsible for different kinds of protein aberrations. Combining them allows us to survey both coding and regulatory aberrations genome wide without underpowering the model. We developed a biologically motivated hierarchical factor graph model which efficiently combines these two sources of data. We use variant harmfulness and gene interactions as priors, to increase the likelihood of identifying the genes correctly. To our knowledge, this is the first work that takes into account complementarity of exome and gene expression data sources in a principled way, integrating variant harmfulness and gene interaction information in the inference process of the model. Our approach a) allows to integrate different data modalities; b) provides a principled way to aggregate rare (and common) variants; c) improves the power of detecting genes associated with a given disease; d) implicates proteins that have been affected in the population in a variety of ways, rather than solely through the coding DNA sequence. Our extensive simulations confirm that our method has superior sensitivity and precision compared to other methods that aggregate rare variants. We have tested our approach in a large breast cancer dataset as a proof of concept and found that our method is able to identify important breast cancer genes. Interestingly, we find genes that have DNA mutations or coding variants in some patients and gene expression aberrations in other patients, indicating that our method is able to effectively explain the disease in more patients.
anna.goldenberg@utoronto.ca

Tuesday 12:15-12:50pm

Casey Brown

Allele specific regulatory activity in the liver

Thousands of human genetic variants have been associated with risk of dozens of common diseases. The vast majority of the identified disease risk loci lie within cis-regulatory elements, however the resolution of GWAS do not allow for the identification of functional genetic variants or their target genes. To address this limitation in the context of cardiometabolic disease mapping, we have built allele specific maps of histone modification states and gene expression in a large collection of liver biopsies. We have developed improved statistical models to jointly model these heterogenous data types with the aim of fine-mapping causal variants. Using these improved causal variant predictions, we have screened thousands of regulatory SNPs with a novel parallelized reporter assay to functionally validate our model predictions.

Tuesday 12:40pm-1:30pm: Lunch Break

Tuesday afternoon - Networks

Tue 1:30pm-2:15pm

John Quackenbush (with John Platig)

Using Networks to Probe Biological Systems

As new technologies are providing us with ever-richer, more complex data sets that capture diverse yet complementary information about biological systems, our challenge is to move beyond simple correlations and explore the complex networks that drive biological systems. I will draw upon separate, but interwoven threads of network analysis that have proven very fruitful in our work, including inference and comparison of gene regulatory networks using a message-passing framework, exploring the diversity and distribution of networks within a phenotype and in state transitions, and exploring the structure of networks as a means of interpreting the factors driving complex, multifactorial phenotypes.

John Platig

Genome Wide Association Studies (GWAS) and eQTL analyses are producing huge numbers of associations and show no signs of slowing. There are now more than 8,500 SNPs associated with more than 350 complex traits reported in the NHGRI GWAS Catalog. However, interpreting these associations collectively in a functional context remains a challenge. Using genotyping and gene expression data from 163 lung tissue samples in a lower respiratory disease study, we calculated eQTL associations between SNPs and genes and cast significant associations as links in a bipartite network. We identified biological function by focusing on densely linked communities, which comprise *groups* of SNPs associated with *groups* of genes. By investigating the intermediate scale of network organization, we found GWAS SNPs enriched at the cores of these communities, including GWAS hits for COPD, asthma, and pulmonary function, among others. We believe these methods are widely applicable to any data set that can be represented as a bipartite network with a giant connected component.

johnq@jimmy.harvard.edu
jplatig@jimmy.harvard.edu

Tue 2:15pm-2:40pm

Gerald Quon

Tissue-specific enhancer networks underlying complex traits

Despite great recent advances, the regulatory architecture of complex traits remains uncharacterized, and the causal variants, target genes, and potential master regulators underlying disease-associated remain uncharacterized. Here, we address these three challenges jointly, using a new graphical probabilistic model for combining large-scale experimental evidence, genetic evidence, and predicted regulatory circuitry. We construct tissue-specific networks linking genetic variants to regulators and target genes through the enhancer regions each is linked to. We develop a graphical probabilistic model that utilizes these networks for joint inference of causal variants, driver enhancers, master regulators, and target genes, using an iterative expectation maximization framework. We apply our framework to the largest collection of genetic association studies for which genome-wide summary statistics are available. We assemble and curate a set of 42 case-control and quantitative trait studies, and analyze them using regulatory networks for 127 cell and tissue types. We find significant regulatory enrichments for 19 traits, including type 1 diabetes, type 2 diabetes, Alzheimer's disease, and schizophrenia. Most genetic variants overlap enhancers with cell type-restricted activity, emphasizing the importance of tissue-specific regulatory annotations. We recover known trait-associated regulators and predict additional regulators that were not previously known but show biologically-relevant annotations. We directly validate our predictions for a subset of traits and show that the predicted target regions show increased evolutionary conservation across both mammals and primates, that the predicted target genes show trait-relevant deletion phenotypes in mouse, and that the predicted regulators show trait-relevant phenotypes in primary human cells upon siRNA knock-down. We implement our method in a software package, CONVERGE (for Complex trait Networks for Variant, Enhancer, Regulator, and Gene Elucidation) a new method for the joint inference of causal variants, driver enhancers, upstream regulators and downstream target genes underlying complex traits. We make our software, code, intermediate datasets, and predictions publicly available, providing important tools for studying the regulatory architecture of any trait. Our trait-specific predictions of GWAS target genes, GWAS target enhancers and upstream master regulators form important starting points for the systematic experimental dissection of human disease.
gquon@mit.edu

Tuesday 2:40pm-3pm: Coffee Break

Tue 3pm-3:35pm

Brenda Andrews

Genetic networks: General properties and complex phenotypes

To define the general principles of genetic networks, our group developed a unique functional genomics platform called 'synthetic genetic array' (SGA) analysis that automates yeast genetics and enables the systematic construction of double and triple mutants. One of our major goals has been to use a simple phenotypic readout of cell growth rate – colony size – to produce the first complete genetic interaction map for any cell, and to empirically delineate the properties of genetic networks. Application of our automated pipeline has enabled systematic analysis of the majority of all possible 18 million yeast gene pairs. The resultant network consists of ~560,000 negative and positive genetic interactions, spanning 93% of all yeast genes. Analysis of the network has revealed: [1] a central role for and unique properties of essential genes, which we consider analogous to disease-associated genes in humans; [2] hubs and pleotropic genes on the network which show a clear association with several

fundamental physiological and evolutionary properties that are predictive of genetic interactions in other organisms; [3] functional modules that we use to predict and test conservation of interactions in other systems.
brenda.andrews@utoronto.ca

Tue 3:35pm-4:10pm

Benjamin Haibe-Kains

Ensemble framework to infer large--scale causal gene regulatory networks from transcriptomic data

It is now established that complex biological phenotypes are not governed by single genes but instead by networks of interacting genes and gene products. As a consequence, deciphering the structure of the gene regulatory network (GRN) is crucial to further our understanding of fundamental processes in human cells. However, the mapping of molecular interactions in the intracellular realm remains a major bottleneck in the pipeline to produce biological knowledge from high-throughput biological data. Multiple methods exist to infer undirected large-scale regulatory networks from collections of transcriptomic data. However very few network inference methods can infer the directionality of predicted gene interactions, despite this being key in the process of better interpreting GRNs. Another challenge when inferring large-scale GRNs consists in quantitatively assessing their validity. Popular, however weak, validation procedures include (i) simulation; (ii) using incomplete 'gold standard' datasets, such as known transcription factors and their targets, which only partially recapitulate the interactions that can be inferred from transcriptomic data; and (iii) using low-throughput laboratory experiments to validate a few predicted interactions, which represent only a very small and potentially biased part of the inferred GRN. To address these issues, we have developed mRMRe, an ensemble approach for network and causality inference, and their integration of priors extracted from the biomedical literature. We applied our new method on a large collection of nearly 500,000 shRNA experiments with gene expression profiles of cancer cell lines before and after knockdown of 3500 genes. This unique dataset allowed us to infer a regulatory networks for 978 landmark genes in multiple cell types, and quantitatively assess their quality. Our results suggest that the complexity of the underlying biology, and the noise present in the shRNA experiments and gene expression profiling make it very challenging to infer meaningful gene-gene interactions. Not only our study highlights the need for quantitatively assessing the predictive value of regulatory networks, but also provides evidence that very large sample size does not necessarily yield high quality networks. These results may open new avenues of research for integrative analysis of multiple data types.
benjamin.haibe.kains@utoronto.ca

Tue 4:10pm-5:30pm

Group discussion

The changing face of scientific discussion, publishing, peer-review, and dissemination.

Introductory statements by: Tuuli Lappalainen, Stephen Montgomery, Yoav Gilad, Mike Snyder, Manolis Kellis, Jeff Leek.

- research, ideas, collaboration, sharing, credit attribution, openness
- reproducibility, code reuse, extending work, error detection, problem detection
- publishing, peer review, pre-prints, non-prints, blogs, establishing priority
- post publication peer review, letters to the editor, pubpeer, pubmed, Twitter, blogs
- dealing with rogue science, manipulation, plagiarism, witch hunts, mob effect
- dissemination, journal news & views, press releases, news stories, university/journal/author blogs

Tue 5:30pm-7:30pm - Dinner + continued

WEDNESDAY AUGUST 5, 2015

Wednesday Morning - 3D

(Breakfast 7am-9am)

Wed 9am-9:35am

Mark Segal

A Two-Stage Algorithm for 3D Genome Reconstruction

The three-dimensional (3D) configuration of chromosomes within the eukaryote nucleus is consequential for several cellular functions including gene expression regulation and is also strongly associated with cancer-causing translocation events. While visualization of such architecture remains limited to low resolutions (due to compaction, dynamics and scale), the ability to infer structures at high resolution has been enabled by recently-devised chromosome conformation capture techniques. In particular, when coupled with next generation sequencing, such methods yield an unbiased inventory of genome-wide chromatin interactions. Various algorithms have been advanced to operate on such data to produce reconstructed 3D configurations. Several studies have shown that such reconstructions provide added value over raw interaction data with respect to downstream biological insights. However, such added value has yet to be fully realized for higher eukaryotes since no genome-wide reconstructions have been inferred for these organisms because of computational bottlenecks and organismal complexity. Here we propose a two-stage algorithm, deploying multi-dimensional scaling and Procrustes transformation, that overcomes these barriers. After showcasing 3D architectures for mouse embryonic stem cells and human lymphoblastoid cells we discuss methods for evaluating these solutions. Downstream deployment of 3D reconstructions to identify ""3D hotspots"" with respect to superposed functional data is also illustrated.
mark@biostat.ucsf.edu

Wed 9:35am-10:10am

Kasper Hansen

Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data

Analysis of Hi-C data has shown that the genome can be divided into two compartments called A/B compartments. These compartments are cell-type specific and are associated with open and closed chromatin. We show that A/B compartments can be reliably estimated using epigenetic data from several different platforms: the Illumina 450k DNA methylation microarray, DNase hypersensitivity sequencing, single-cell ATAC sequencing and single-cell whole-genome bisulfite sequencing. We do this by exploiting the fact that the structure of long range correlations differs between open and closed compartments. This work makes A/B compartments readily available in a wide variety of cell types, including many human cancers.
khansen@jhsph.edu

Wed 10:10am-10:30am: Coffee Break

Wed 10:30am-11:05am

Alan Moses

Statistical methods for automated analysis of high-throughput protein localization data

Advances in high-throughput genetics and automated microscopy have led to biological image collections of unprecedented size and scale. These data are biologically rich, but also high-dimensional and highly heterogeneous. Open data analysis questions include: How do we compare quantitative measurements between experiments? How do we identify rare patterns where few (or no) known examples are available for training data? How do we obtain statistical confidence in non-independent measurements? I will describe our efforts to used unsupervised approaches to address several of these challenges. We have developed quantitative, biologically interpretable image-based measurements (features) that we can make for each cell in a microscope image, which allows us to quantitatively compare patterns and perform analysis in the feature space. We found that most previously known subcellular localization patterns can be identified in unsupervised analysis, and that rare, complex patterns of localization can also be identified. We have also explored kernel-based approaches to model cell-cell variability in image data, and use these to perform the first systematic search for genes with cell-cell variability in subcellular localization. In gener5al, I believe that putting the subcellular localization data from images in an unbiased quantitative framework will facilitate discovery and integration with other large-scale biological data.
alan.moses@utoronto.ca

Wed 11:05am-11:40am

Sohrab Shah (25'+10')

Somatic mutations in two cancer contexts: evolutionary dynamics and gene expression impact

I will present two important emergent concepts in the interpretation of somatic mutations in cancer genomes. First, I will show evolutionary dynamics in cancer across temporal and spatial axes as measured by the prevalence of mutations marking clonal genotypes. I will show results from time series measurements of breast cancer patient derived xenograft models as well as patterns of clonal spread as inferred from synchronously sampled intra-peritoneal deposits in ovarian cancers. The second part of the talk will highlight cancer gene discovery, patient stratification and mutation interpretation through simultaneous integration of mutations and gene expression profiles.

Results will be shown from analysis of the TCGA cohort across twelve tumour types which revealed ~30 novel tumour suppressor candidate genes and ~90 genes with previously undescribed trans-effects. Taken together, results from evolutionary dynamics and gene expression integration implicate temporal, spatial and transcriptional contexts as critical factors in the interpretation of somatic mutations in cancer.
sshah@bccrc.ca

Wed 11:40am-12:15pm

Annabelle Haudry (25'+10')

An evolutionary perspective of genome size variations in Drosophila

Understanding why Eukaryote species exhibit such a huge range in their genome size appears as a central and longstanding question in evolutionary biology. Despite an extensive body of research, there is still little consensus on the evolutionary mechanisms driving genome size evolution. One potentially important contributing factor is mobile selfish genetic elements named transposable elements (TEs) that represent half of our own genome. As they are thought to be essentially neutral or deleterious, TEs' accumulation in a host genome is expected to be restrained by natural selection, which intensity is positively correlated with the effective population size. According to this assumption, it was proposed that TE genomic content depends on the effective population size, which is in turn correlated with some life-history traits and is affected by demographic history. However, other overlooked factors could affect TE content and genome size. In particular, interspecific exchanges of genetic elements (horizontal transfers) may help the spread of selfish DNA in a "naïve" host-species from which they were previously absent. The development of a novel, integrative and evolutionary approach is therefore crucial in order to estimate the relative impact of all those different factors on genome size evolution.
annabelle.haudry@univ-lyon1.fr

Thu 12:15pm-12:35pm

Marieke Kuijjer (15'+5')

Estimating sample-specific regulatory networks

Biological systems are driven by intricate interactions among the complex array of molecules that comprise the cell. Many methods have been developed to reconstruct network models that attempt to capture those interactions. These methods often draw on large numbers of measured expression samples to tease out subtle signals and infer connections between genes (or gene products). The result is an aggregate network model representing a single estimate for edge likelihoods. While informative, aggregate models fail to capture the heterogeneity that is often represented in a population. Here we propose a method to reverse engineer sample-specific networks from aggregate network models. We demonstrate the accuracy and applicability of our approach in several datasets, including simulated data, microarray expression data from synchronized yeast cells, and RNA-seq data collected from human subjects. We show that these sample-specific networks can be used to study the evolution of network topology across time and to characterize shifts in gene regulation that may not be apparent in the expression data. We believe the ability to generate sample-specific networks will revolutionize the field of network biology and has the potential to usher in an era of precision network medicine.

Wed 12:35pm-1:30pm: Lunch

Wednesday Afternoon - Johnston canyon hike

THURSDAY AUGUST 6, 2015

Thursday Morning - Cancer

(Breakfast: 7am-9am)

Thu 9am-9:35am

Rob Scharpf

Genomic heterogeneity of structural variants in ovarian cancer

Despite the success of targeted therapies for many solid tumors in recent years, the treatment and overall survival of patients diagnosed with ovarian cancer is largely unchanged. In part, the poor prognosis reflects the gap between the recognition of ovarian cancers as a collection of distinct cancer subtypes and our ability to molecularly characterize these distinctions in terms of mutations and structural variant profiles that could be used to stratify risk and guide therapies. Towards this end, we performed whole genome sequencing of over 50 ovarian cancer cell lines and primary tumors, including serous, mucinous, clear cell, and endometriod subtypes. Nearly all ovarian cancers harbor deletions, amplifications, inversions, intra- and inter-chromosomal translocations, and in-frame gene fusions. Our analyses highlight the importance of structural alterations in ovarian cancer and identify key driver genes and pathways that may be clinically useful. This talk will emphasise the computational and statistical analyses underpinning the estimation and interpretation of structural variant profiles for ovarian cancers.
rscharpf@jhu.edu

Thu 9:35am-10:10am

Venkatraman Seshan

Copy Number Profile from Tumor Sequencing

Next generation sequencing is employed regularly in cancer both for clinical and research objectives. DNA sequencing of the whole exome or a panel of cancer genes are used to get information on somatic changes to the DNA. The sequencing reads can be used to obtain both the total and allele specific copy number information using both total and allele specific coverage depths. This in turn can be used to estimate the cellular fraction of the tumor as a whole as well as the cellular fractions of all the somatic changes. In earlier work we developed the Circular Binary Segmentation algorithm for the analysis of array based copy number data and extended it to parent specific copy numbers (PSCBS) for SNP array data. In this talk I will present CBS methodology adapted to sequencing data and show examples of its application to sequencing experiments.
seshanv@mskcc.org

Thu 10:10am-10:30am: Coffee Break

Thu 10:30am-11:05am

Paul Scheet

Surveys of Subtle Allelic Imbalance in Tissue

Somatically-acquired allelic imbalance (AI) is an established factor in cancer initiation and has recently been implicated as a marker for cancer risk. While DNA microarrays and next-generation sequencing are effective for whole-genome profiling of AI, in typical settings their sensitivities become extremely limited when the aberrant cell fraction (or tumor purity) is below 10-20%. Yet, this range may be critical for early detection and diagnostics, since often for such applications the samples of interest will be comprised of heterogeneous mixtures of cells with a large component of DNA from normal (i.e. the germline) rather than aberrant (e.g. the tumor) sources. Here we introduce a powerful haplotype-based computational technique (Vattathil & Scheet, 2013, Gen Res) and use it to characterize AI in several difficult settings. We start with a reanalysis of a study of over 35,000 samples of healthy tissue from recent genome-wide association studies and find a 2-fold higher rate of somatic mosaicism (within-individual genomic heterogeneity), which may indicate a wider applicability for the use of mosaicism as a biomarker for cancer risk. We next examine premalignant tissue, profiling polyps from individuals at risk for colorectal cancer to show subtle levels of AI across critical loci; we also demonstrate extensive mosaicism in the lung field (normal-appearing tissue surrounding the tumor), consistent with recent studies of expression (Kadara et. al., 2014, JNCI). Finally, we study lymph node tissue (of lung cancer

patients), sampled via endobronchial ultrasound, and discover chromosomal aberrations in samples that were deemed negative by pathology review but that were ultimately determined to be positive following surgical extraction, thus demonstrating potential for molecular diagnostics.
permutations@gmail.com

Thu 11:05am-11:25am

Richard Cowper Sallari (15’+5’)

Convergence of dispersed regulatory mutations reveals candidate driver genes in prostate cancer.

Cancer genome sequencing has revealed cancer-associated genes based on recurrent mutations across independent tumors, but has been largely restricted to protein-coding alterations. Here, we extend recurrence analysis to dispersed regulatory mutations, based on transcriptome and genome sequencing of cancer-normal sample pairs in prostate cancer. We infer the regulatory plexus of each gene in healthy prostate, defined as its three-dimensional neighborhood of regulatory regions, combining chromosome looping and epigenomic annotations. Cancer-dysregulated genes show an increased mutation rate in non-coding regions, enriched in predicted enhancers active in prostate and diverse other tissues, suggesting out-of-context de-repression. Controlling for mutational heterogeneity across tumors, genomic regions, and chromatin states, we identify 15 genes showing significant regulatory recurrence, with roles in androgen-insulin signaling, immune system evasion, and mitochondrial function, suggesting higher-order pathway-level convergence. Our results provide a model for both cancer and personal genome regulatory analysis, by coalescing low-frequency scattered mutations into high-frequency regulatory events.
sallari@mit.edu

Thu 11:25am-12:00pm

Elli Papaemmanuil

Dissecting genetic and phenotypic heterogeneity to deliver personalized predictions in cancer patients

Systematic sequencing screens of thousands of cancer genomes has delivered a near complete catalogue of somatic mutations in cancer implicating >400 cancer genes. Molecular profiling within distinct tumour types recurrently unravels a long tail of infrequently mutated genes. Coupled with in depth genomic profiling approaches, these have revealed that cancer is a disease of extensive genomic diversity and clonal complexity. In each patient, mutations are acquired over time, resulting in clonal diversification, and parallel evolution. This has profound implications for understanding the clonal origins of disease, the molecular underpinnings of progression and predicting response to therapy. We present a in depth molecular characterization of 1540 AML patients enrolled in clinical trials of the German Austrian Study Group. Custom capture analysis of 111 genes together with recurrent cytogenetic alterations results in the characterization of 5234 pathogenic lesions. We formally model genomic structure and find that AML is subdivided in at least 11 molecular subgroups, each defined by recurrent second- and third order genetic interactions and associated with distinct clinical presentation and clinical outlooks. Perceived rare driver genes are significantly enriched within molecular subsets and often dictate class membership. Importantly we show that second and third order interactions markedly redefined clinical phenotype and long term clinical response. To this effect, we apply global statistical models to study the relative contributions of independent variables including demographic (age, gender), diagnostic, treatment and genomic variables and build personally tailored prognostication models. We show that large knowledge banks of well annotated clinical cohorts, coupled with detailed molecular annotation can deliver refined risk estimates tailored to individual patient status and that inform from the composite molecular architecture that defines a patient tumour.
papaemme@mskcc.org

Thu 12:00pm-12:20pm

Marianne DeGorter (15’+5’)

Whole genome sequencing of diverse human populations resolves causal regulatory variants

A major challenge to identifying causal regulatory variation is distinguishing the causal variant from multiple tightly-linked alternatives. By leveraging differing patterns of linkage disequilibrium across populations, it is possible to localize the causal variant. Now, with the availability of whole genomes from multiple populations from Phase 3 of the 1000 Genomes Project, coupled with gene expression data in 414 individuals from six populations, we leveraged genetic variability to identify causal regulatory variants shared among human populations. We found that nearly half of all expression quantitative loci (eQTLs) discovered (FDR < 0.05) in each of the six populations represent blocks of tied variants, some with up to several hundred variants in perfect linkage disequilibrium in our sample. By meta-analysis of gene expression in all the populations, we can break ties and assign a single candidate causal variant in the majority of loci with tied variants. We further demonstrate that these candidate causal variants are more likely to overlap transcription factor binding sites and H3K27ac marks from ENCODE, compared to a SNP chosen naively from the haplotype block (p < 0.001). When considering the relative utility of combinations of populations, we detect that inclusion of African populations provides the largest improvement in our ability to detect functional variants due to increased genetic diversity in these populations. Applying this approach, we further refine and report the properties of causal variants underlying several GWA studies. Overall, we demonstrate that a multiple population approach will be an important and efficient design for follow-up investigations which aim to localize causal variants from eQTL and GWA studies.

Thu 12:20pm-1:30pm: Lunch Break

Thu afternoon - Epigenomics and Statistical Data Integration

Thu 1:30pm-2:05pm

Jean-Philippe Vert

Some new methods for robust high-dimensional classification

Learning predictive models from genomic data remains challenging due to the high dimensionality and the complexity of the data. I will discuss a few techniques that we have investigated recently to try to overcome some of the challenges: (1) the Kendall and Mallows kernels, which learn a predictive model based on pairwise comparisons between features, and (2) new atomic matrix norms, to learn models with particular sparsity structures such as disjoint support or sparse latent factors.
Jean-Philippe.Vert@mines-paristech.fr

Thu 2:05pm-2:40pm

Aurelie Labbe

Component-based models for statistical data integration

Multi-block component methods are a class of statistical methods with a tremendous potential for data integration. Several particular cases of these methods, such as Partial Least Square (PLS) regression or Canonical Correlation Analysis (CCA) for example, have been widely used for the analysis of multivariate phenotypes. Because such methods directly focus on the correlation between datasets, we believe that they merit attention in the genomic field. Specifically, we consider methods for analysis of several sets of data measured on the same subjects. Datasets are assumed to come from at least three different sources; for example, genotypes, gene expression, DNA methylation, brain imaging or clinical traits. One of the sets of data is the “phenotype data”, or the outcome(s) of interest. We present a new component based method to predict phenotypes jointly with the most highly shared information among the datasets. We’ll also provide some recommendations about the use of component-based methods in general for data integration
aurelie.labbe@mcgill.ca

Thursday 2:40pm-3pm: Coffee Break

Thu 3:00pm-3:35pm

Alexis Battle

The complex and cascading impact of regulatory variation

Regulatory genetic variation is believed to play a significant role in human disease, but while we have identified thousands of variants affecting mRNA levels, we do not yet fully understand their downstream consequences or the factors that affect their impact. In recent work, we have evaluated the modulation of genetic effects by common environmental and behavioral risk factors, such as substance use, using novel models of allele-specific expression. Additionally, we are currently developing integrated machine learning methods to identify non-coding variants likely to have large effects on the cell and higher-level phenotypes. Together, the methods we are developing will help characterize the complex consequences of regulatory variation.

Thu 3:35pm-3:55pm

David Knowles (15’+5’)

Joint modeling of cellular and disease QTLs
The majority of known GWAS associations fall in non-coding genomic regions.

QTLs for cellular phenotypes such as gene expression, splicing and translation rates are known to be enriched in GWAS hits, suggesting that some proportion of GWAS variants are mediated through these phenotypes. However, this enrichment is usually shown simply as a post-hoc analysis. We consider here three extensions over standard enrichment analysis: (i) we jointly model QTLs across different transcriptomic phenotypes (for example, RNA synthesis, expression, and ribo-profiling); (ii) we use an explicitly polygenic model of effect sizes in GWAS; (iii) we deconvolve eQTLs for mixed tissues into cell type specific eQTLs using known expression signatures and chromatin data for purified cell types This methodology offers the potential to improve power in GWAS, provide more mechanistic understanding of GWAS variants, and better identify cell types, genes and pathways relevant to a particular disease.

Thu 3:55pm-4:30pm

Manolis Kellis

Dissecting non-coding associations with human disease

Perhaps the greatest surprise of genome-wide association studies (GWAS) of human disease is that 90% of top-scoring disease-associated loci lie outside protein-coding regions. This has increased the urgency of mapping non-coding DNA elements and regulatory circuits, in order to understand the molecular basis of human disease. To address this challenge, we have developed and applying new methods to systematically characterize the epigenomic landscape of diverse primary human cells and tissues, resulting in the annotation of enhancer elements across primary human tissues and cell types. We also predicted tissue-specific regulatory networks linking these enhancers to their upstream regulators and target genes, and enable us to weave genetic information from GWAS through these networks to recognize preferentially-disrupted genes, regulators, and biological processes. In this talk, I will describe the use of non-coding annotations and circuits for understanding the molecular basis of genetic differences underlying disease.
manoli@mit.edu

Thu 4:30pm-5:30pm

Session reports + Discussion

RNA, eQTLs, GWAS, Networks, 3D, Cancer

Future slides compilation, discussion, directions, ideas

Thu 5:30pm-7:30pm - Dinner

FRIDAY AUGUST 7, 2015

Friday Morning - Discussion

(Breakfast: 7am-9am)

9am: Lake Moraine Hike:

Fri 12:15pm-1:30pm
Adjourn + Lunch + Departures