Early Career Investigators Meeting on Quantitative Problems in Human Health and Genetics

Jan 10 - Jan 15, 2016

Banff International Research Station

for Mathematical Innovation and Discovery

Organizers

Noah Zaitlen
Tuuli Lappalainen
Michael Hoffman
Julien Ayroles
Jennifer Listgarten

CONTACT INFO

The mailing list for all the participants of this workshop is 16w5078@lists.birs.ca . For contacting the organizers, email Tuuli (tlappalainen@nygenome.org) or Noah (noah.zaitlen@ucsf.edu)

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: In the foyer of the TransCanada Pipeline Pavilion (TCPL)
The 2nd floor lounge of Corbett Hall has beverages and snacks provided by Illumina
*Please remember to scan your meal card at the host/hostess station in the dining room for breakfast, lunch and dinner.

TALKS

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.

Each presenter has a 50-minute slot including time for questions/discussion. We encourage talks that make use of both the projector and the chalkboards. Please conclude your talk with the presentation of an open problem of your choosing. This could be an unanswered scientific question that needs more attention, a computational-statistical-experimental technology of high value to your research, or the next community scale data resource in the vein of 1000 Genomes, GTEx, or Encode.

OTHER ACTIVITIES

The conference venue is about 10-15 minutes walk from the Banff village with various shops and restaurants, ski rental, and ski shuttle stops. On Wednesday we have free time for skiing, hiking, and other activities. http://www.skibig3.com/ has information of skiing and related services in the area.

SPONSORED BY

illumina®

SCHEDULE

Sunday

16:00	Check-in begins (Front Desk – Professional Development Centre - open 24h)
17:30-19:30	Buffet Dinner
20:00	Informal reception in the 2nd floor lounge, Corbett Hall

Monday

7:00-8:45	Breakfast
8:45	Introduction and Welcome by BIRS Station Manager, TCPL
9:00	Michael Hoffman: Semi-automated genome annotation and an expanded epigenetic alphabet
9:50	Coffee Break, TCPL
10:30	Cole Trapnell: Manifold learning for single-cell expression data
11:20	Group Photo; meet in foyer of TCPL (photograph will be taken outdoors).
11:30-13:00	Lunch
13:00-14:00	Guided Tour of The Banff Centre; meet in the 2nd floor lounge, Corbett Hall
14:00	Dan MacArthur: Using large-scale variation data sets to interpret human gene function
14:50	Coffee Break TCPL
15:30	James Zou: Modeling the rare and missing variants reveals constraints in the human genome
16:20-17:10	Simon Gravel: The Great Migration and African-American genomic diversity
17:30-19:30	Dinner
19:30-20:30	Discussion

Tuesday

7:00-9:00	Breakfast
9:00	Jimmie Ye: ImmVar 2.0: Genetics of human immune response
9:50	Coffee Break
10:30	Alexis Battle: Understanding the impact of rare regulatory variation
11:20	Tuuli Lappalainen: Towards a multidisciplinary synthesis of function of the human genome
12:10	Lunch
13:30	Iuliana Ionita-Laza: A spectral approach integrating functional genomic annotations for coding and noncoding variants
14:20	Coffee Break
15:30	Ekta Khurana: Integrating large-scale genomics data to understand the role of non-coding regions in cancer

16:20-17:10	Bogdan Pasaniuc: Integrative methods to finemap GWAS risk loci
17:30-19:30	Dinner
19:30-20:30	Discussion

Wednesday

7:00-9:00	Breakfast
8:30 or 9:10	Ski shuttle departure to Sunshine Village from Town Lot (Behind Mt Royal Hotel) in Banff. See http://www.skibig3.com/ski-shuttle/ for full schedule
17:30-19:30	Dinner
19:30-20:30	Julien Ayroles: Shifting focus from mean to variance: The contribution of loci affecting phenotypic variability to phenotypic variation

Thursday

7:00-9:00	Breakfast
9:00	A.P. Jason de Koning: Revisiting the population genetics of molecular evolution
9:50	Coffee Break
10:30	Noah Zaitlen: Methods for genetic studies across multiple phenotypes
11:20	Ben Neale: What's coming in complex trait genetics
12:10	Lunch
13:30	Elinor Karlsson: Behavioral genetics in mixed breed dogs
14:20	Coffee Break
15:30	Joe Pickrell: Detection and interpretation of shared genetic influences on 42 human traits
16:20	Chris Cotsapas: Identifying changes to gene expression driven by disease risk variants
17:30-19:30	Dinner

Friday

7:00-9:00	Breakfast
9:00	Discussion
11:00-11:30	Checkout
11:30-13:30	Lunch

Checkout by 12 noon. We can 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.

ABSTRACTS

(in alphabetic order by speaker surname)

Julien Ayroles
jayroles@princeton.edu
Princeton University

Shifting focus from mean to variance: The contribution of loci affecting phenotypic variability to phenotypic variation

[Abstract TBD]

Alexis Battle
ajbattle@cs.jhu.edu
Johns Hopkins University

Understanding the impact of rare regulatory variation

The increase in availability of full human genome sequences presents great opportunity for understanding the impact of rare genetic variants. Based on current knowledge, however, we are still limited in our ability to interpret or predict consequences of rare and private variants in non-coding regions of the genome. The availability of RNA-seq and other cellular measurements for the same individuals with genome sequencing offers a new avenue for integrated methods for prioritizing rare regulatory variants. We demonstrate that diverse signals from RNA-seq including allelic imbalance, and expression of both proximal and distal genes are informative for identifying the impact of rare non-coding variants largely excluded from previous analyses. Here, I will discuss the use of Bayesian machine learning to integrate whole genome sequencing with such molecular phenotypes. Such approaches for analysis of rare regulatory genetic variants offer great potential for identifying potentially deleterious non-coding genetic variants from individual genomes.

Don Conrad
dconrad@genetics.wustl.edu
Washington University

An analysis of the n=1 problem in human genetics

The diagnosis of rare, idiopathic diseases is emerging as a primary application of medical genome sequencing. However, the application of standard tools from genetic epidemiology for many of these cases is frustrated by a combination of small sample sizes, genetic heterogeneity and the large number of singleton variants found by genome sequencing. I will present novel statistical and experimental approaches for identifying unusual functional variation from a single genome, what I refer to as the n-of-1 problem. One such approach is a framework for calculating the population sampling probability (PSAP) of an arbitrary genetic variant that reflects the functional effect of the variant, and the selective constraint and local mutation rate of the underlying sequence. I will show that PSAPs behave like well-calibrated p-values when applied to single genomes, and that they can be used to rapidly and sensitively infer the identity of known disease mutations from the background of typical human genetic variation. I will show how PSAPs can be easily combined with other data types, such as transcriptome data from the GTEx project, and what improvements may be realized from integrating other data types. Finally I will discuss application of these approaches to a number of real life n=1 cases from cohorts of infertile men, children with congential urinary tract anomalies, and families from the NIH Undiagnosed Diseases Program.

Chris Cotsapas
cotsapas@broadinstitute.org

Yale School of Medicine
Identifying changes to gene expression driven by disease risk variants
[Abstract TBD]

Simon Gravel
simon.gravel@mcgill.ca
McGill university
The Great Migration and African-American genomic diversity
We present a detailed population genetic study of 3 African-American cohorts comprising over 3000 genotyped individuals across US urban and rural communities: two nation-wide longitudinal cohorts, and the 1000 Genomes ASW cohort. Ancestry analysis reveals a uniform breakdown of continental ancestry proportions across regions and urban/rural status, with 79% African, 19% European, and 1.5% Native American/Asian ancestries, with substantial between-individual variation. The Native American ancestry proportion is higher than previous estimates and is maintained after self-identified Hispanics and individuals with substantial inferred Spanish ancestry are removed. This supports direct admixture between Native Americans and African Americans on US territory, and linkage patterns suggest contact early after African-American arrival to the Americas. Local ancestry patterns and variation in ancestry proportions across individuals are broadly consistent with a single African-American population model with early Native American admixture and ongoing European gene flow in the South. The size and broad geographic sampling of our cohorts enable detailed analysis of the geographic and cultural determinants of finer-scale population structure. Recent identity-by-descent analysis reveals fine-scale geographic structure consistent with the routes used during slavery and in the great African-American migrations of the twentieth century: east-to-west migrations in the south, and distinct south-to-north migrations into New England and the Midwest. These migrations follow transit routes available at the time and are in stark contrast with European-American relatedness patterns.

Michael Hoffman
michael.hoffman@utoronto.ca
Princess Margaret Cancer Centre/University of Toronto
Semi-automated genome annotation and an expanded epigenetic alphabet
First, we will discuss Segway, an integrative method to identify patterns from multiple functional genomics experiments, discovering joint patterns across different assay types. We apply Segway to ENCODE ChIP-seq and DNase-seq data and identify patterns associated with transcription start sites, gene ends, enhancers, CTCF elements, and repressed regions. Segway yields a model which elucidates the relationship between assay observations and functional elements in the genome.

Second, we will discuss a new method to discover transcription factor motifs and identify transcription factor binding sites in DNA with covalent modifications such as methylation. Just as transcription factors distinguish one standard nucleobase from another, they also distinguish unmodified and modified bases. To represent the modified bases in a sequence, we replace cytosine (C) with symbols for 5-methylcytosine (5mC), 5-hydroxylmethylcytosine (5hmC), 5-formylcytosine (5fC). Similarly, we adapted the well-established position weight matrix model of transcription factor binding affinity to an expanded alphabet. We created an expanded-alphabet genome sequence using genome-wide maps of 5mC, 5hmC, and 5fC in mouse embryonic stem cells. Using this sequence and expanded-alphabet position weight matrixes, we reproduced various known methylation binding preferences, including the preference of ZFP57 and C/EBPβ for methylated motifs and the preference of c-Myc for unmethylated motifs. Using these known binding preferences to tune model parameters enables discovery of novel modified motifs.

Iuliana Ionita-Laza

ii2135@columbia.edu
Columbia University

A spectral approach integrating functional genomic annotations for coding and noncoding variants

Over the past few years, substantial effort has been put into the functional annotation of variation in human genome sequence. Indeed, for any genetic variant, whether protein coding or noncoding, a diverse set of functional annotations is available from projects such as Ensembl, ENCODE and Roadmap Epigenomics. Such annotations can play a critical role in identifying putatively causal variants among the abundant natural variation that occurs at a locus of interest. The main challenges in using these various annotations include their large numbers, and their diversity. Here we discuss an unsupervised approach to derive a meta-score (Eigen), that, unlike most existing methods, is not based on any labeled training data. The proposed method produces estimates of predictive accuracy for each functional annotation score, and subsequently uses these estimates of accuracy to derive the aggregate functional score for variants of interest as a weighted linear combination of individual annotations. We show that the resulting meta-score has better discriminatory ability using disease associated and putatively benign variants from published studies (for both Mendelian and complex diseases) compared with the recently proposed CADD score. Furthermore, an important advantage of the Eigen score is that it can be easily adapted to a specific tissue or cell type.

Elinor Karlsson
elinor@broadinstitute.org
U. Mass. Medical School

Behavioral genetics in mixed breed dogs

The domesticated dog is one of the best natural models for human behavioral and psychiatric disorders. The strong artificial selection on behavioral traits in dogs pushed associated variants of large effect up in prevalence, makes them particularly tractable to genomewide association mapping. Our work on canine compulsive disorder identified genes and candidate functional variants implicating dysfunction in the same glutamatergic signaling pathways associated with human OCD.

Until now, gene mapping in dogs has focused almost exclusively on dog breeds – genetically isolated created within the last few hundred years. We are implementing, for the first time, GWAS in the mixed breed pet dog population. We are soliciting detailed behavioral phenotype information from dog owners as citizen scientists, allowing association studies to be done quickly with much larger sample sizes. New analytical approaches that leverage the complex population history of pet dogs could prove particularly powerful for mapping causal variants.

Eimear Kenny
eimear.kenny@mssm.edu
Mount Sinai School of Medicine

Ekta Khurana
ekk2003@med.cornell.edu
Weill Cornell Medicine

Integrating large-scale genomics data to understand the role of non-coding regions in cancer

Most variants obtained from whole-genome sequencing occur in noncoding regions of the genome. Although variants in protein-coding regions have received the majority of attention, numerous studies have now noted the importance of noncoding variants in cancer. Identification of functional noncoding variants that drive tumor growth remains a challenge and a bottleneck for the use of whole-genome sequencing in the clinic. The overall theme of my talk will be the identification of functional variants in

somatic tumor genomes and applying the lessons learned from germline genomes. I will discuss the various modes in which noncoding sequence variants can cause oncogenesis and then I will present a computational framework to annotate and prioritize cancer regulatory mutations. Using this scheme, we identified candidate noncoding drivers in ~600 samples from 10 different cancer types. I will also discuss the ongoing efforts to apply this approach to analyze ~2700 tumor whole-genomes in the 'Pan-Cancer Analysis of Whole Genomes, PCAWG' consortium.

Jason de Kooning
jason.dekoning@ucalgary.ca
University of Calgary, School of Medicine

Revisiting the population genetics of molecular evolution

One of the most important ways we can predict the functional impact of human mutations is by carefully inferring what sequence variations worked throughout human and vertebrate evolution. I will discuss our recent work on solving two aspects of this problem: 1) the development of theory for correctly accounting for how population-level genetic processes affect the long-term rate of molecular evolution under variable mutation rates; and 2) scalable computational approaches for enabling large-scale Bayesian inference of heterogeneous mutation and selection in across-species genome comparisons.

Tuuli Lappalainen
tlappalainen@nygenome.org
New York Genome Center & Columbia University

Towards a multidisciplinary synthesis of function of the human genome

Functional genomics analysis of genetic variants in the human genome has become an increasingly important discipline to bridge together quantitative understanding of genetic variation and molecular study of genome function. In this talk, I will discuss what the eQTL research of the recent years has taught us about the tissue- and condition-specific function of the human genome, genetic interactions, and interpretation of GWAS loci. In particular, I will discuss the discordances of the insights obtained from eQTL analysis and other fields such as molecular biology and GWAS-related research. Many of these questions lack definitive answers and the reasons for any possible discordances are rarely fully understood or even discussed. I claim that to reach a true understanding of the human genome, each field digging deeper on its own must be complemented with a systematic attempt to bridge the gaps to reach a comprehensive, multidisciplinary synthesis.

Daniel MacArthur
danmac@broadinstitute.org
Broad Institute

Using large-scale variation data sets to interpret human gene function

To date over a million humans worldwide have been exome or genome-sequenced, representing a potential treasure trove of data about the distribution of genetic variation in human genes; however, much of these data remain inaccessible for various technical, political and ethical reasons. Here I describe the efforts of the Exome Aggregation Consortium (ExAC) to assemble a large, harmonized call set of genetic variation across over 90,000 individuals. I will focus on the uses of the resulting data set to better understand the impact of genetic variation on disease, to identify genes showing high sensitivity to genetic perturbation, and finally its use (in combination with other resources) to identify and deeply phenotype human "knockouts". Finally, I will discuss the planned extensions of this resource, including larger sample sizes, the incorporation of whole-genome sequence data, and improved access to phenotype data and genotype-based recall of samples.

Benjamin Neale
bneale@broadinstitute.org
Massachusetts General Hospital
What's coming in complex trait genetics
The next few years will see dramatic increases in sample size for all forms of genetic enquiry. In this talk, I will outline clear opportunities for increasing the size and scope of genetic datasets and how these might be effectively used for gaining insight into the biological and epidemiological phenomena for complex traits.

Bogdan Pasaniuc
pasaniuc@ucla.edu
UCLA
Integrative methods to finemap GWAS risk loci
Although GWAS have been extremely successful in identifying numerous risk loci for complex traits and diseases, at the vast majority of these loci, the causal mechanism between genetic variation and disease risk remains largely unknown. In the quest to address this gap, post-GWAS studies are experiencing a "big data" revolution driven by the exponentially decreasing costs of high-throughput genomic assays. Multiple layers of data (genetic variation, transcriptome levels, epigenetic modifications, localization of tissue-specific regulatory sites, etc.) are routinely collected in increasingly large cohorts of individuals. I will discuss new methods that integrate various types of data (genetic, epigenetic, transcriptomic) to understand the causal mechanism of disease at GWAS risk loci.

Joe Pickrell
jkpickrell@nygenome.org
New York Genome Center
Detection and interpretation of shared genetic influences on 42 human traits
We performed a genome-wide scan for genetic variants that influence multiple human phenotypes by comparing large genome-wide association studies (GWAS) of 42 traits or diseases, including anthropometric traits (e.g. nose size and male pattern baldness), immune traits (e.g. susceptibility to childhood ear infections and Crohn's disease), and psychiatric diseases (e.g. schizophrenia and Parkinson's disease). First, we identified 341 loci (at a false discovery rate of 10%) that influence multiple traits. Several loci influence a large number of phenotypes; for example, variants near the histo-blood group gene ABO influence eleven of these traits, including risk of childhood ear infections and allergies, among others. Similarly, a nonsynonymous variant in the zinc transporter SLC39A8 influences seven of these traits, including risk of schizophrenia and Parkinson's disease, among others. Second, we used these loci to identify traits that share multiple genetic causes in common. For example, genetic variants that delay age of menarche in women also, on average, decrease risk of male pattern baldness, and variants that increase risk of schizophrenia also tend to increase risk of inflammatory bowel disease. Finally, we developed a method to identify pairs of traits that show evidence of a causal relationship, and used this to identify four such pairs. For example, we show evidence that increased BMI causally increases triglyceride levels, and that increased liability to hypothyroidism causally decreases adult height.

Oliver Stegle
stegle@ebi.ac.uk
EMBl-EBI
Genetic analysis of correlated traits

I will discuss a range of different challenges that occur in the analysis of correlated phenotypes. In this talk, I will start with simple models that use a handful of traits for genetic analysis to then consider high-dimensional phenotypes as they occur in molecular association genetics.

One thought is to discuss the integration of phenotypes of the same type (e.g. genes across pathways) and distinctly different traits that are measured though different technologies. I hope to give outlook on ongoing work in the genetic analysis of stem cells and other studies we are currently involved in.

Cole Trapnell
coletrap@uw.edu
University of Washington

Manifold learning for single-cell expression data

Single-cell transcriptomics and epigenomics reveal the global cellular state of thousands of individual cells in a single experiment, opening the door to analyses that are unavailable with bulk genomics. For example, tracking cell differentiation with single-cell gene expression can reveal the complete "trajectory" of intermediate cell states on a path from stem cell to terminal cell type. Many analyses require clustering cells by type or state in an unsupervised manner. However, early single-cell genomics experiments have revealed far more cell-to-cell variation than anticipated, uncovering rare cell subtypes and cryptic intermediate states. Even enumerating the number of distinct cell types in a tissue has proved challenging from a statistical perspective. Here, I will discuss our recent efforts to formulate single-cell type and state classification as a manifold learning problem. I will demonstrate that learning an embedded cell graph directly from single-cell RNA-Seq data reveals cryptic cellular decision points that are masked by simple linear dimensionality reduction.

Jimmie Ye
jimmie.ye@ucsf.edu
UCSF

ImmVar 2.0: Genetics of human immune response

GWAS of autoimmune and neurodegenerative disease have unequivocally implicated innate and adaptive immune response in disease pathogenesis. In order to better interpret GWAS associations, we established the ImmVar Consortium to map the genetic basis of human immune response. Here, we present new results on the genetic control of gene regulation in dendritic cells (DCs) and CD4+ T cells using next generation sequencing. In CD4+ T cells, we show widespread cis genetic control of gene expression (eQTLs) and chromatin accessibility (caQTL) with significant overlap between eQTLs and caQTLs. In DCs, in addition to eQTLs, we detected widespread cis genetic control of isoform usage, many of which overlapped known GWAS loci. These results suggest that genetic control of transcription and splicing factor binding are key mechanisms for establishing transcriptome variability. We will conclude with a hypothesis about the functional basis for common non-coding disease associations and discuss what experimental resources and computational methods are required for the interpretation of their biological function.

Noah Zaitlen
noah.zaitlen@ucsf.edu
UCSF

Methods for genetic studies across multiple phenotypes

Testing for associations in big data faces the problem of multiple comparisons, with true signals buried inside the noise of all associations queried. This is particularly true in genetic association studies where a substantial proportion of the variation of human phenotypes is driven by numerous genetic variants of small effect. The current strategy to improve power to identify these weak associations consists of

applying standard marginal statistical approaches and increasing study sample sizes. While successful, this approach does not leverage the environmental and genetic factors shared between the multiple phenotypes collected in contemporary cohorts. Here we develop two methods that improve the power of detecting associations when a large number of correlated variables have been measured on the same samples. Our analyses over real and simulated data provide direct support that large sets of correlated variables can be leveraged to achieve dramatic increases in statistical power equivalent to a two or even three or four fold increase in sample size.

James Zou
jzou@fas.harvard.edu
Microsoft Research and MIT

Modeling the rare and missing variants reveals constraints in the human genome

I will describe our recent work leveraging the largest collection of human exomes (ExAC) to model the landscape of harmful genetic variations in healthy individuals. We developed an algorithm that uses the variants identified in ExAC to accurately estimate statistics of the variants that are not in this cohort but exist in the general population. Our linear program algorithm has strong mathematical guarantees. The inferred statistics of rare and unobserved variants provide a framework to quantify the discovery power of future sequencing projects, such as the Precision Medicine Initiative. Our model also quantifies constraints on pathways, genes, protein domains and individual codons, and we estimated the selection coefficients corresponding to the observed constraints.