A Novel Region-Based Bayesian Approach for Genetic Association with Next Generation Sequencing (NGS) Data

Laurent Briollais

Joint work with Amelia Xu (PhD student, University of Toronto)

Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital
& Dalla Lana School of Public Health
University of Toronto

laurent@lunenfeld.ca

Aug. 8, 2018

Background: NGS

Next Generation Sequencing Studies

- The emergence of new high-throughput genotyping technologies, such as Next Generation Sequencing (NGS), allows the study of the human genome at an unprecedented depth and scale
- They provide invaluable opportunities to decipher the biological processes involved in complex human diseases
- The study of the genetic landscape of inherited and acquired mutations in cancer patients could provide invaluable insights into the essential pathways driving the progression from a normal cell to non-invasive precursor lesions, and then to advanced and metastatic diseases

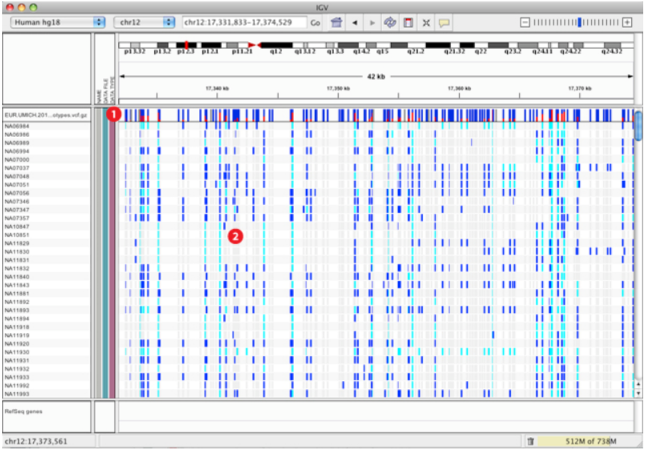
Outline of our framework

- Model setting
- Bayes Factor derivation for case-control design
- Prior definition
- Hyper-parameter specification
- Asymptotic properties
- Genome-wide inference
- Simulations with the program sim1000G
- Application on lung cancer study

Example NGS data

NGS data

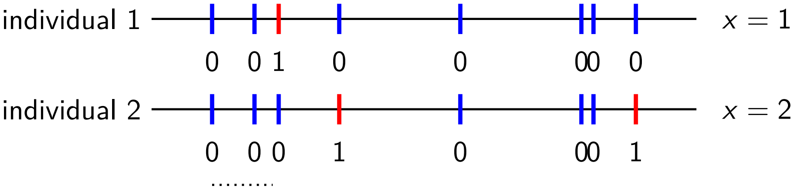
An example of sequenced genomic region is displayed below through the sequence viewer IGV.



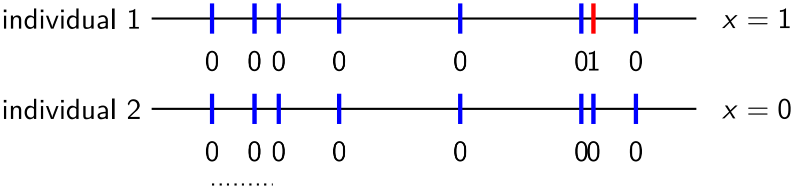
1	Each bar across the top of the plot shows the allele fraction for a single locus.
2	The genotypes for each locus in each sample. Dark blue = heterozygous, Cyan = homozygous variant, Grey = reference. Filtered entries are transparent.

Data example: a genetic region with 10 loci

Cases:



Controls:



Blue: non-mutated locus

Red: mutated locus

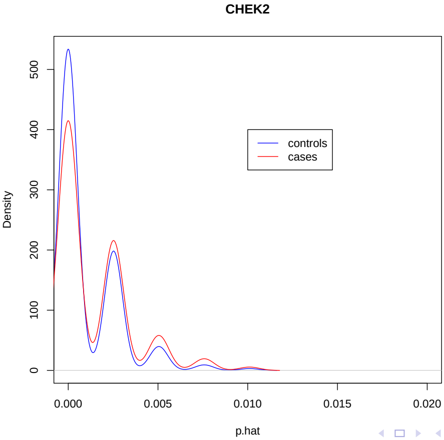
Density curve of $\hat{p}$ of real data

k: inidvidual k

n: number of loci in the region

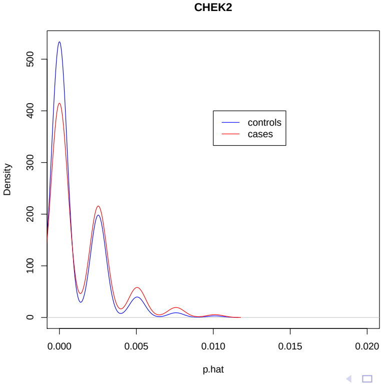
x_k: number of rare variants in the region for individual k, $x_k \sim Binomial(n, p_k)$

p_k: probability of having a rare variant at single locus for individual k, $\hat{p}_k = \frac{x_k}{n}$



Rationale for our rare variant association test

- Goal: Develop regional association test based on the comparison of rare variant rate (p_i) distribution between cases and controls.
- This comparison is accomplished by using the Bayes Factor (BF) statistic.

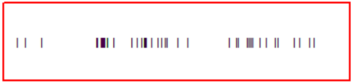


Background: Bayes Factor

Influential work on the BF: The "BayesBall"

- Albert, J. (2008), "Streaky Hitting in Baseball", Journal of Quantitative Analysis of Sports, vol. 4.
- Albert, J. (2013), "Looking at Spacings to Assess Streakiness", Journal of Quantitative Analysis of Sports, vol. 9.

Joe DiMaggio



- Bayes factor in support of true streakiness is

$$BF_K = \frac{f(y|M_K)}{f(y|M)}.$$

BF in genetic association studies

- First GWAS application = the WTCCC study (2007)
- Some review in Stephens and Balding (Nat. Rev. Genetics, 2009)
- Wakefield (2009) formalized the BF in the context of GWAS
  - Interesting discussion about informative priors (effect-MAF dependence) vs. non-informative priors (implicit p-value prior)
  - Sketches the use of BF in the Bayesian False Discovery although not detailed
- McCallum and Ionita-Laza, Biometrics 2015

Methods

Model Setting

- Let X_{ijk} be the count of rare variants in the region i, for group j and individual k

$$X_{ijk} \sim Binomial(n_{ijk}, p_{ijk})$$

- Suppose that p_{ijk} varies across genetic regions and individuals, according to a prior density function $g(p_{ijk}|\boldsymbol{\theta}_{ij})$, with $\boldsymbol{\theta}_{ij} \equiv \boldsymbol{\theta}_{i1}$ if j is in the control group and $\boldsymbol{\theta}_{ij} \equiv \boldsymbol{\theta}_{i2}$ if j is in the case group.

- Our goal is to assess whether there is a difference in rare variant counts between cases and controls in a particular region i by comparing : $H_{i0} : \boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i2} = \boldsymbol{\theta}_i$ vs. $H_{i1} : \boldsymbol{\theta}_{i1} \neq \boldsymbol{\theta}_{i2}$ using the Bayes Factor (BF) statistic.

Bayes Factor

- Bayes Factor (BF) is the ratio between the probabilities of the data (marginal likelihood) under the alternative hypothesis (association exists) and the null hypothesis (no association).

$$BF = \frac{m_1(X)}{m_0(X)}$$

- The marginal likelihood function under H_0 and H_1:

$$m_0(X) = \int_P f(X|P)g(P)dP = \int_P f(X|P) \int_\theta g(P|\theta)\pi(\theta|\eta^*, K^*)d\theta dP$$

$$m_1(X) = \int_{P_1} f(X_1|P_1) \int_{\theta_1} g(P_1|\theta_1)\pi(\theta_1|\eta_1^*, K_1^*)d\theta_1 dP_1 \times \int_{P_2} f(X_2|P_2) \int_{\theta_2} g(P_2|\theta_2)\pi(\theta_2|\eta_2^*, K_2^*)d\theta_2 dP_2$$

where θ is the parameter we want to compare between cases and controls.

- There are two definitions for the prior distribution $g(P|\theta)$.

Prior definition I

- Under the beta prior distribution, we have

$$p_{ijk}|\boldsymbol{\theta}_{ij} \sim Beta(\eta_{ij}, K_i),$$

Here the beta distribution is parametrized in terms of mean (denoted by η_{ij}) and precision (denoted by K_i). Relationship with (α, β):

$$\eta = \frac{\alpha}{(\alpha+\beta)}, \quad K = \alpha + \beta.$$

- With the Beta prior, the marginal distribution of rare variants count in the region is Beta-Binomial (BB). It assumes a similar pairwise correlation between the rare variants within the region. Our simulation studies (thanks to Fode Tounkara) showed that the BB fits the sequencing rare variants data much better than many Copula alternatives.

Prior definition II

- Under the mixture prior distribution, we assume that p_{ijk} follows a mixture distribution of a point mass at zero and a beta distribution with probability w_{0ij} and $w_{1ij} = 1 - w_{0ij}$, respectively:

$$X_{ijk} = \begin{cases} 0, & \text{if } p_{ijk} = 0 \text{ with } P(p_{ijk} = 0) = w_{0ij} \\ X_{ijk} \sim Bin(n_{ijk}, p_{ijk}), & \text{if } p_{ijk} > 0 \text{ with } P(p_{ijk} > 0) = 1 - w_{0ij} \end{cases}$$

Also when $p_{ijk} > 0$, the prior density for p_{ijk} is $Beta(\eta_{ij}, K_i)$.

Hierarchical hyper-parameter specification

- Our hyper parameters of interest are η, η_1, η_2, w_{01}, w_{02}, and w_0.
- We assume a hierarchical prior structure where each hyper-parameter is assumed to follow a beta distribution with new mean and precision parameters η^*, η_1^*, η_2^*, K^*, K_1^*, K_2^*.
- The parameters of the prior and hyperprior distributions are estimated empirically from the data by using MLE.

BF distribution under the null

- Ideal parameters η^* and K^* should lead to:
    - BF is independent of gene size
    - BF ($\log BF$) has a known theoretical distribution

- *Theorem 1*. Assume that $\eta^* = \hat{\eta}$, $K^* = \hat{\eta}\hat{\Sigma}^{-1}$, $\eta_1^* = \hat{\eta}_1$, $K_1^* = \hat{\eta}_1\hat{\Sigma}_1^{-1}$, $\eta_2^* = \hat{\eta}_2$ and $K_2^* = \hat{\eta}_2\hat{\Sigma}_2^{-1}$, for gene i, when sample size $N_1 \to \infty$ and $N_2 \to \infty$,

$$2 \log BF = \frac{(\hat{\eta}_1 - \hat{\eta}_2)^2}{\hat{\Sigma}_1 + \hat{\Sigma}_2} \sim \chi^2(1)$$

BF with individual-level covariates

For group j (j=1 or 2, j=1, control group, j=2, case group), individual k, $p_{jk} \sim Beta(\eta_{jk}, K)$. We build Beta regression to model the relationship between covariate vector w_{jk} with length equal to c and the rare variant rate at single locus p_{jk}.

- Version 1

$$logit(\eta_{jk}) = \beta_{0j} + w_{jk}\beta$$

$$\beta_{0j} \sim Normal(\mu_j, \sigma_j^2)$$

$$\beta \sim MVN(\mu_\beta, B)$$

- Version 2

$$logit(\eta_{jk}) = logit(\eta_j) + R_{jk},$$

where $R_{jk} = \beta w_{jk}$ and w_{jk} is a vector of PCs or ethinic group indicator variables.

$$\eta_j \sim beta(\eta_j^*, K_j^*)$$

Bayesian FDR

Bayesian control of False Discovery Rate (FDR) for genome wide inference

- The goal of genome-wide inference is to perform a simultaneous testing of multiple hypotheses (i.e. all the genes or genomic regions) $= m$ null hypotheses $H_i, i = 1, \cdots, m$, using data Y

- Let $Z_i = 1$ if H_i is true and $Z_i = 0$ if H_i is false, $i = 1, \cdots, m$, and π_0 the proportion of regions/genes generated under the null

- We have $Z_i|\pi_0 \sim Bernoulli(1 - \pi_0)$

- We also define δ_i denote a decision rule in $(0, 1)$ on Z_i based on the data and $D = \sum_{i=1}^{m} \delta_i$

Bayesian control of False Discovery Rate (FDR) for genome wide inference

Following Muller et al. (2006), the False Discovery Proportion (FDP) is defined as

$$\text{FDP} \equiv \frac{\sum_{i=1}^{m} \delta_i (1 - Z_i)}{D \bigvee 1},$$

and the Bayesian FDR as:

$$\overline{FDR} \equiv E(FDP|Y) = \frac{\sum_{i=1}^{m} \delta_i (1 - v_i)}{D \bigvee 1}.$$

The interest in the Bayesian control of the FDR, is to estimate $v_i \equiv Pr(Z_i = 1|Y)$ by

$$\hat{v}_i = \frac{(1 - \hat{\pi}_0) BF_i}{\hat{\pi}_0 + (1 - \hat{\pi}_0) BF_i}$$

Estimate of $\hat{\pi}_0$

- Wen et al. (2016) showed that an upper bound estimation of π_0 can be obtained by

$$\hat{\pi}_0 = \frac{\sum_{i=1}^{m} I(BF_i \leq q_{i,\gamma})}{m\gamma}.$$

  $=>$ requires permutations to assess the null distribution of the BF for each gene
  $=>$ lacks well study of impact of γ

- Since we proved that $2\log BF_i \xrightarrow{d} \chi^2(1)$, we can then estimate π_0 by

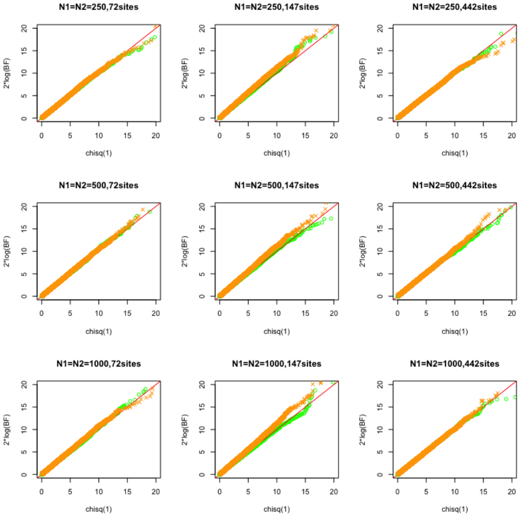
$$\hat{\pi}_0 = \frac{\sum_{i=1}^{m} I(2\log BF_i \leq q^*_\gamma)}{m\gamma},$$

  where q^*_γ is the γ-quantile of a $\chi^2(1)$ distribution
  $=>$ which avoids the need for permutations
  $=>$ Try to find optimal value of γ

Simulation Procedure

- R package "sim1000G" is used to simulate the rare variant genotype data.
  - Now available on the CRAN, credit to Apostolos Dimitromanolakis
  - The simulated data can capture the allele frequencies and LD patterns in the genome, as well as recombination hotspots.
  - Only choose variants with $MAF \in (1e - 6, 0.01)$ for data analysis.
- Number of causal variants is proportional to the region size. We assume all causal variants are deleterious, with OR $= 2.63$ to 3.73, inversely related to MAF.
- Each simulated dataset has same number of cases and controls.

QQ plot: BF simulated under the null



Simulation Results

Table: Statistical power of different methods for different gene sizes and sample sizes with 1,000 replicates (reject null hypothesis when $p < 0.05$)

Statistical Test		$N_1 = N_2 = 250$			$N_1 = N_2 = 500$			$N_1 = N_2 = 1000$		
		72 sites	147 sites	442 sites	72 sites	147 sites	442 sites	72 sites	147 sites	442 sites
BF method										
Beta prior	Compare η	23.8	41.3	87.2	35.3	58.9	98.3	59.4	82.2	100.0
Mixture prior	Compare η	25.6	44.4	88.7	37.1	61.7	98.2	62.2	83.5	100.0
SKAT		13.1	22.0	50.2	24.9	45.2	86.1	55.7	79.1	99.9
Burden		16.9	30.2	83.5	25.8	50.2	96.6	48.4	75.1	100.0
SKAT-O		16.8	32.6	82.5	29.9	57.6	98.0	61.8	88.5	100.0

Lung cancer data application

Lung Cancer Study

- Our data is from lung cancer exome-sequencing consortium study, including 4 different cohorts.
- After removing the duplicated individuals, sample size of different cohorts

Cohort	cases	controls	Total
Toronto	260	258	518
Liverpool	65	69	134
HSPH	426	269	695
IARC	293	284	577

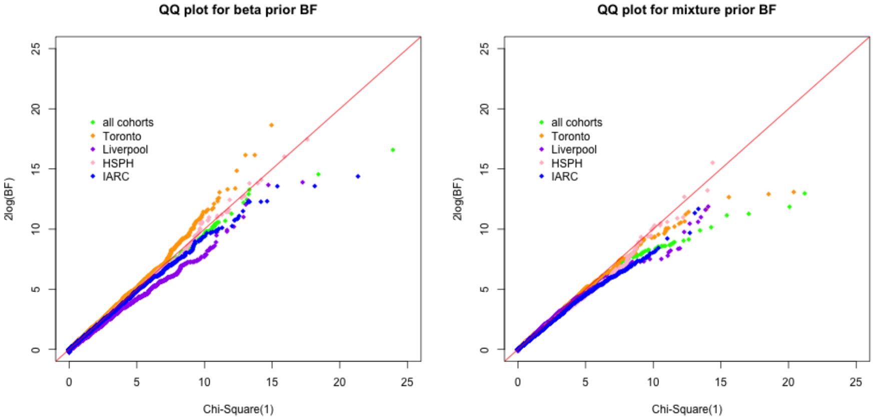
Data summary

- After filtering out multi-allelic variants, the MAF distribution for the bi-allelic variants are

MAF	0	(0,0.01]	(0.01,0.05]	(0.05,0.5]	Total
#(Variants)	62,940	1,095,794	60,204	129,412	1,348,350
Proportion (%)	4.7	81.3	4.5	9.6	

- In the analysis, the number of sites within the gene is at least 20 for beta prior BF and 50 for mixture prior BF.
- The number of genes used for beta prior BF and mixture prior BF are 14,321 and 7,454 respectively.

QQ plot: include all variants

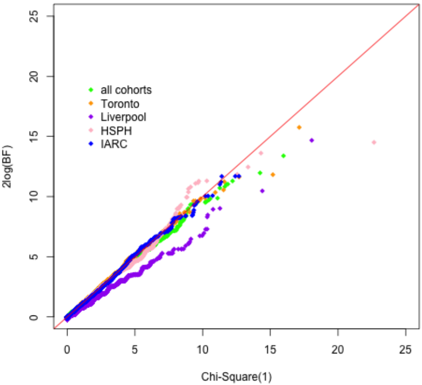


QQ plot for beta prior BF

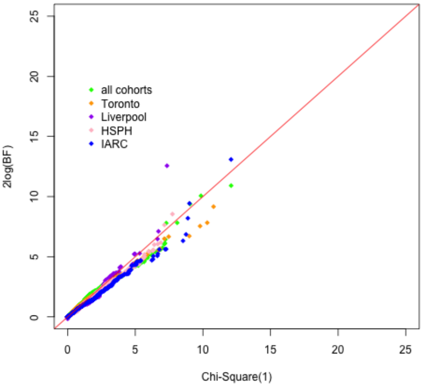
QQ plot for mixture prior BF

QQ plot: include high impact variants

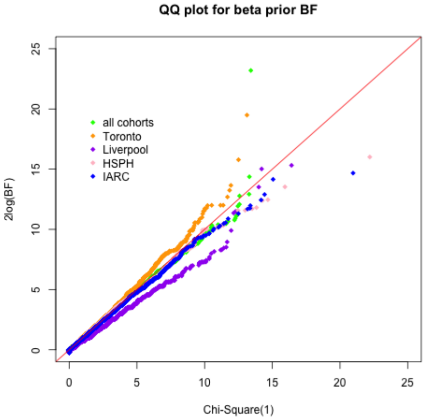
QQ plot for beta prior BF



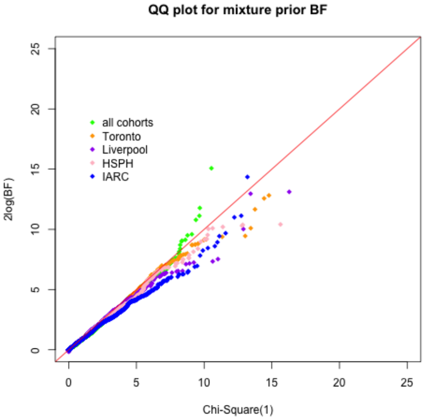
QQ plot for mixture prior BF



QQ plot: include high and moderate impact variants



QQ plot for beta prior BF



QQ plot for mixture prior BF

Bayesian FDR application

Table: Estimate of π_0

	$\gamma = 1 - \frac{1}{m}$		$\gamma = 0.99$		$\gamma = 0.95$		$\gamma = 0.9$	
	beta	mixture	beta	mixture	beta	mixture	beta	mixture
all variants	1	1	0.9993095	1	1	1	1	1
high risk	1	1	0.9995661	1	1	1	0.9952272	1
moderate risk	1	1	1	1	0.9987782	1	0.9992063	1

FDR of the top gene using beta prior in the moderate risk dataset:

- $\gamma = 0.95$, FDR ≈ 0.007
- $\gamma = 0.9$, FDR ≈ 0.01

Top 20 genes with beta prior: high impact variants

gene.name	chr	sites	BFbeta	p.beta	BF(TO)	BF(Livepool)	BF(HSPH)	BF(IARC)
CAMTA2	17	48	807.97	2.53e-04	35.13	1.64	1.71	10.48
ADAMTSL4	1	52	397.20	5.41e-04	3.85	0.98	17.91	10.09
CACNA1G	17	44	283.65	7.77e-04	2.54	0.97	13.08	10.24
SCRIB	8	56	249.19	8.93e-04	4.35	1.70	2.19	24.98
SREBF2	22	43	247.25	9.01e-04	8.26	0.94	3.38	21.21
ERBB2	17	36	224.39	1.00e-03	7.70	0.89	5.06	2.10
PCDH7	4	22	212.20	1.06e-03	4.62		5.35	3.61
SAMD4B	19	21	139.04	1.68e-03	1.16			1.95
CDC42BPA	1	38	135.67	1.73e-03	1.03		1.15	346.24
PAMR1	11	22	127.98	1.84e-03	7.83		1.05	24.03
PP2D1	3	31	121.32	1.95e-03	2.45	2.21	11.00	3.03
WDR92	2	21	120.58	1.96e-03		1.65	1.08	7.29
CCDC60	12	31	116.36	2.04e-03	9.69	0.89	1414.61	1.04
ABL2	1	30	114.33	2.08e-03	5.43	3.52	1.66	4.98
KIF20A	5	24	113.20	2.10e-03	49.67		2.74	2.92
RBM14	11	21	110.41	2.16e-03		1.09	5.84	1.77
TERT	5	26	106.05	2.26e-03	28.90	0.89	1.39	13.90
AXDND1	1	37	90.50	2.68e-03	1.30	0.94	11.12	5.40
LRSAM1	9	37	86.92	2.81e-03	195.87	2.64	1.91	1.41
FN1	2	63	78.13	3.15e-03	2.51	1.07	5.26	3.70

Impact of protective variants

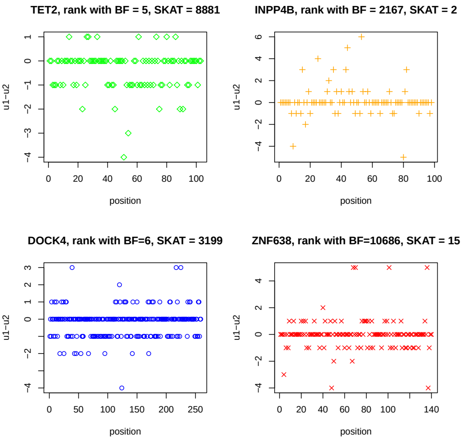


Figure: The Y-axis represents the difference in total minor allele counts between controls (u_1) and cases (u_2) at each single site (locus) of the region. If the genetic variant has a deleterious effect on the disease, then $u_2 > u_1$ and conversely if it has a protective effect, then $u_1 > u_2$.

Discussion

Discussion

- The use of empirical Bayes priors along with a Bayesian control of FDR offer a comprehensive framework to make genome-wide statistical inference about the important chromosomal regions associated with the disease of interest
- How to define the priors? asymptotic properties of BF or informative priors?
- logBF $\approx$ logLR $+$ log $\frac{\pi(\theta|H_1)}{\pi(\theta|H_0)}$ - term
- The regression framework might offer a good compromise (Zhou and Guan, JASA, 2018) but still not fully developed for discrete outcomes
- Future developments include the extension of the BF approach to account for variant-level covariates and family designs

Acknowledgments

- **Dr. Wei Xu (thesis co-supervisor)**
  Princess Margaret Cancer Centre and Dalla Lana School of Public Health, University of Toronto
- **Dr. Rayjean Hung (lung cancer study)**
  Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital and Dalla Lana School of Public Health, University of Toronto
- **Dr. Geoffrey Liu (lung cancer study)**
  Princess Margaret Cancer Centre and Dalla Lana School of Public Health, University of Toronto
- **Briollais Lab: Apostolos Dimitromanolakis and Fode Tounkara**

CIHR STAGE
STRATEGIC TRAINING FOR ADVANCED GENETIC EPIDEMIOLOGY

OICR
Ontario Institute for Cancer Research