Whole Exome Sequencing of Affected Sister Pairs with Early Onset Breast Cancer

BIRS Workshop:
New Statistical Methods for Family-based Sequencing Studies

Razvan Romanescu, Gesseca Gos, Irene Andrulis, Shelley B. Bull

Lunenfeld-Tanenbaum Research Institute

7 August 2018

Lunenfeld-Tanenbaum Research Institute
Sinai Health System

UNIVERSITY OF TORONTO
DALLA LANA SCHOOL OF PUBLIC HEALTH

Background – Breast Cancer (BRCA)

BRCA is a heterogeneous disease, mutations are rare.

Germline mutation identification for rare diseases benefits from starting with a homogeneous population of cases sharing the phenotype

In young women – BRCA more likely due to germline alterations affecting tumor susceptibility genes

Hypothesis:

Focusing on a genetically homogeneous cohort (young sister pairs with non-*BRCA* BC)

- enrich for the presence of rare intermediate-to-high risk variants
- enable discovery of novel variants

Whole Exome Sequencing Pilot Study - Design

Motivating study data:

Whole exome sequencing (WES) at 50X coverage

- affected sister pairs, at least one early-onset (≤45)
- recruited from high-risk families in Ontario Familial Breast Cancer Registry

Total of 21 families

- Family history of breast cancer
- screened negative for known mutations in high-penetrance genes BRCA1, BRCA2 and CHEK2*1100delC.

Objective: Identify novel rare variants for familial breast cancer => further validation studies.

Whole Exome Sequencing – Filtering Results

Preliminary Studies

Whole-exome sequencing of blood samples at 50X Coverage using Agilent SureSelect on HiSeq 2500

↓

Alignment to Human Reference Genome Hg19

↓

Variant Calling using HaplotypeCaller of the Genome Analysis Toolkit (GATK)

↓

Variant Quality Score Recalibration (GATK)

↓

Variant Annotation (wAnnovar)

↓

261,340 Variants Called

↓

256,337 Variants Passing Quality Control Filters
[Phred-scaled Q Score >= 30]
[Read Depth >= 10]

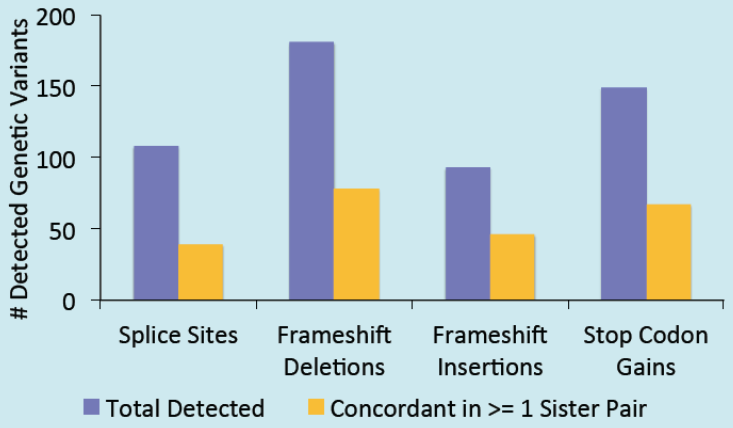
↓

Functional Variants:
1,440 Variants with Expected Functional Consequences (Protein Truncations)
[Stop codon gain mutations, Frameshift insertions or deletions, Splice site mutations]

↓

Rare Variants:
535 Varients Remaining After Removal of Those with Minor Allele Frequencies (MAF) > 0.1% in the Human Population (1000 Genomes Database; ESP6500 Database)

Rare Functional Variants



Concordant Variants of Interest

Gene Functional Category	# Concordant Rare Variants
DNA Repair	7
Cell Proliferation	3
Tumor Suppression	3
Cell Cycle Regulation	4
Other Genes Previously Associated with Cancer	8

Case-control mutation screening in a larger cohort

Statistical Methods to Identify Rare Variants ??

Consider methods with complementary strengths
Allelic and locus heterogeneity are important considerations.

Novel mutations may be family-specific or occurring in few families, with the possibility of extreme heterogeneity.

Methods for RV association analysis in affected sister/relative pairs

- exploit IBD sharing information
- susceptibility variants more often within regions shared IBD by ASPs compared to regions not shared IBD
  - powerful when multiple sibpairs carry shared RVs,
    e.g. multiple different mutations within the same gene
- less effective when families segregate different susceptibility genes

Propose extensions to consider multiple regions
(eg within a shared pathway, such as DNA repair)
more effective when there is locus heterogeneity

Affected Sister Pair Data - Notation

Assume i = 1, 2, … N families each with 2 affected sisters (l=1,2)

A genomic region with j = *1,2, …,R* RV loci (RV = MAF < 0.1%)
filtered on MAF reference information (e.g., 1000 Genomes)
and functional annotation (e.g. ANNOVAR).

Let X_{ilj} be the RV allele count (0,1,2) at locus j in family i, sister l

Q_{ij} is the sum of the RV allele counts (0 - 4) in sibpair i at locus j

Q_i is the sum of Q_{ij} over j = *1,2, …, R* RV loci

Z_i is the # of alleles shared IBD (0,1,2) in the genomic region
(assuming no recombination)

Statistical Inference: Single Region Test

Epstein et al (2015) model the dependence of $Q_i.$ on Z_i

$$E[Q_i.|Z_i] = 4\mu_0 + 2(\mu_1 - \mu_0)Z_i$$

$$Var[Q_i.|Z_i] = 4\sigma_0^2 + 2Z_i(2\sigma_1^2 - \sigma_0^2)$$

Means (μ_0, μ_1) & variances (σ_0^2, σ_1^2) of rare allele counts depend on the IBD sharing

μ_0 - mean of RV sum on parental haplotype NOT IBD

μ_1 – mean of RV sum on parental haplotype inherited IBD

Efficient score test H_0: ($\mu_1 = \mu_0$)

in inverse variance weighted regression

Burden type

Robust to population stratification

Does not require a linkage signal to detect association

More powerful than case-control design

Single Region Regression Test

Simplification:

Assume $\sigma_0^2 = \sigma_1^2$ (true under the null)

$$Q_i.|Z_i = \alpha + \beta Z_i$$

The test of $\beta = 0$ vs. H_a: $\beta \neq 0$ $\quad T_{reg} = \dfrac{\hat{\beta}}{SE(\hat{\beta})} \sim t\,(N-2)$

Weighted version:

$$Q_i.|Z_i = \alpha + \beta Z_i + (4 + 2Z_i)\varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$

Allows allele counts in a sibpair to depend on IBD

Multi-Region Regression Test

$$Q_{qi\cdot} = \sum_{j=1,\ldots,R_q} Q_{qij} \qquad q = 1,2,\ldots,p$$

Multi-variate regression: $Q_{qi\cdot}|Z_{qi} = \alpha + \beta Z_{qi}$

A sibpair has a RV allele count for each of p regions:

Count depends on IBD in the region

Assumption of a shared $\beta = \beta_q$

test statistic for $\beta = 0$ vs. $\beta \neq 0$

$$T_{reg} = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim N(0,1)$$

Allelic Parity Test (ignores IBD)

Define $$D_{i\cdot} = \sum_{j=1,..,R} D_{ij}$$

$$D_{ij} = \begin{cases} 0, & Q_{ij} = 0 \\ -1, & Q_{ij} = 1 \\ 2, & Q_{ij} = 2, \end{cases} \quad , \quad \text{for } j = 1,..,R, \quad i = 1,\dots,N$$

$$T_{a.p.} = \frac{\sum_{i=1,..,N} D_{i\cdot}}{SD(D_{i\cdot})\sqrt{N}} \sim Student\ t(df = N-1)$$

Discrete small sample distribution => Synthetic distribution

Multi-Region extension for p regions: $$D_{\cdot i\cdot} = \sum_{q=1,..,p} \sum_{j=1,..,R_q} D_{qij}$$

Simulation Study Design

Heterogeneity Models:

A family potentially segregates one rare mutation that increases susceptibility / reduces age at onset.

The mutation differs between families, but is in the same gene/region * *allelic heterogeneity*

The mutation is in a different gene/region in different families * *locus heterogeneity*

A mutation in any one of several independent regions (eg. that form a functional pathway) can increase risk of disease

Simulation Study Data Generation

Genetic Data:

'sim1000G' R package

- 594 European haplotypes from 1000 Genomes
- 10-20 kb region (one gene) - 100 RV loci with MAF < 0.1%
- 15% are potential "causal" mutations in carrier families
- one RV "carrier" haplotype assigned to each family
- Mendelian segregation to daughters

Age at Onset Data:

'FamEvent' R package

- proportional hazards model under dominant inheritance

$$h(t, X_{ilj}) = h_0(t - t_0)\, exp\, (\beta_j X_{ilj})$$

$$t_0 = 20$$

$$\beta_j = \log(8)$$

Ascertainment:

- Early age at onset criteria for sisters
- One sister age <40, another <50

10,000 datasets: 20, 100, 500 ASPs

Simulation Study Families – Single Region Test

Region A "carriers":

- "causal" mutation at a locus in Region A
- carrier penetrance function
- enriched for early age at onset

Region B "carriers":

- non-carrier in Region A
- "causal" mutation at a locus in Region B
- enriched for early age at onset

Region A&B "non-carriers":

- non-carrier in Regions A & B
- Environmental penetrance
- Sporadic disease with early age at onset

Simulation Study Families – Two Region Test

Three Regions (A, B, C) with "causal" RVs:
Regions A & B included in the test

In Family i, a sister can be one of:

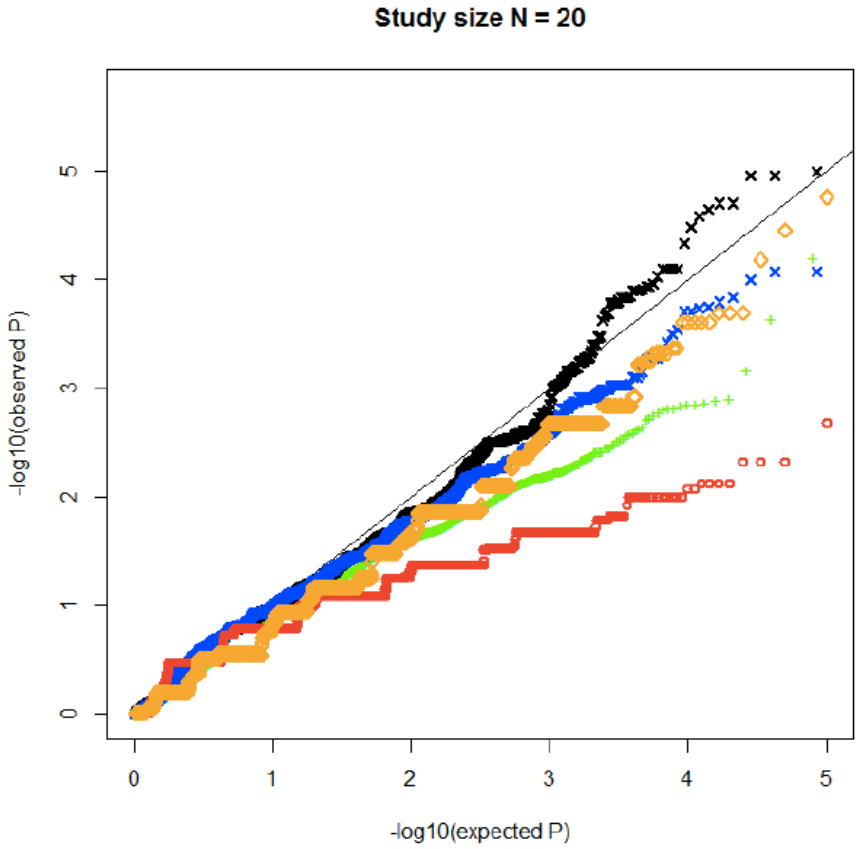
- carrier of a causal mutation in Region A
- carrier of a causal mutation in Region B
- carrier of a causal mutation in Region C
- non-carrier of any causal mutation in any region

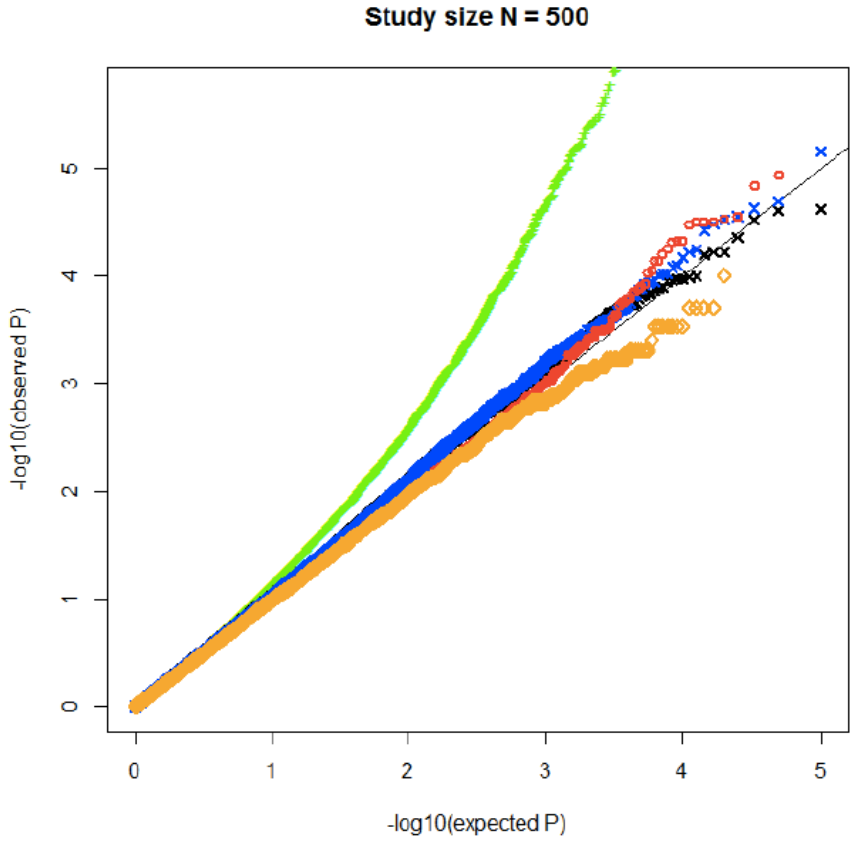
Simulation Study Results - Single Region Test

Type I error

100,000 Datasets

green – Epstein, blue – regression test,
black – weighted regression test,
red – allelic parity test (asymptotic),
orange – allelic parity test (synthetic distribution)



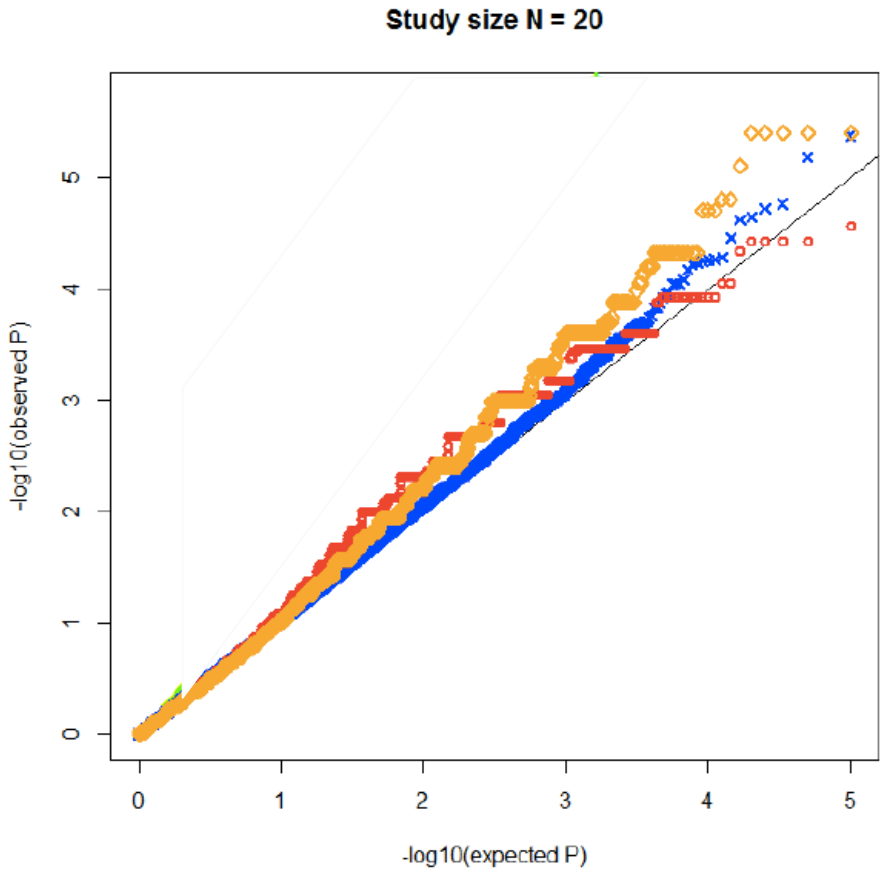


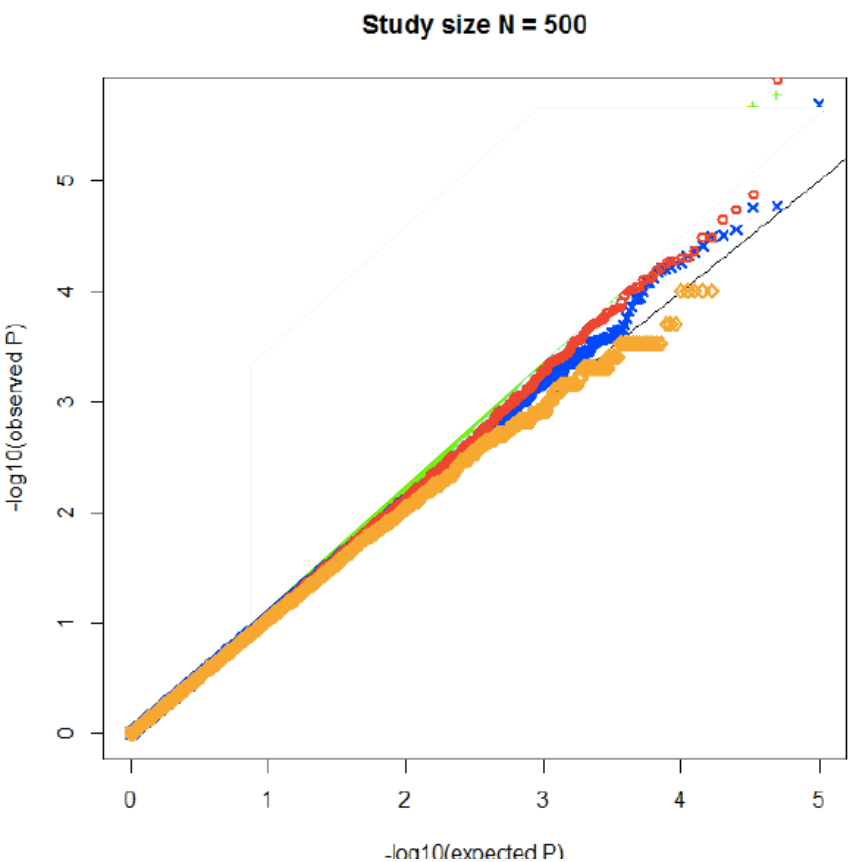
Simulation Study Results - Two Region Test

Type I error

100,000 Datasets

blue – regression test,
red – allelic parity test (asymptotic),
orange – allelic parity test (synthetic distribution)





Simulation Study Results - Power

N	20	100	500
	$\alpha = 0.05$		
regression	0.05	0.124	0.346
regresssion (w)	0.048	0.143	0.393
a.p.	0.033	0.229	0.633
a.p.synthetic	0.045	0.285	0.685
Epstein	0.087	0.216	0.419

10,000 Datasets

Single Region

N	20	100	500
	$\alpha = 0.05$		
regression	0.053	0.07	0.152
a.p.	0.06	0.166	0.493
a.p.synthetic	0.082	0.193	0.521

Multi Region

Summary & Discussion

Preliminary results – should be cautious

- Asymptotic assumptions in small samples?
- Robustness to non-normality in simplified linear regression
- Why does Epstein's model lose T1E control in small samples?

Impact of simulation design

- How plausible is the extreme heterogeneity hypothesis?
- Role of background risk due to common variants?

Applications

- How to specify RV sets in a region?
- How to choose regions for multi-region analysis?

Extension to WGS

- How to choose families? How many to re-sequence ?
- How to use pedigree data?

Design for population-based validation/replication?

Acknowledgements

Razvan Romanescu, Post-doctoral Fellow, Bull Lab
Gesseca Gos, Post-doctoral Fellow, Andrulis Lab
Michela Panarella, MSc Biostatistics, U of Toronto

Irene Andrulis, Lunenfeld-Tanenbaum Research Institute
Gord Glendon, Ontario Family Breast Cancer Registry, Cancer Care Ontario

Laurent Briollais and Colleagues, Lunenfeld-Tanenbaum Research Institute

CIHR IRSC
Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

Canadian Breast Cancer Foundation

OICR
Ontario Institute for Cancer Research

NSERC CRSNG

CIHR STAGE
Strategic Training for Advanced Genetic Epidemiology

McLAUGHLIN CENTRE
Advancing Genomic Medicine through Research and Education

Thanks

References:

Choi, Kopciuk, He, Briollais (2017). FamEvent: Family Age-at-Onset Data Simulation and Penetrance Estimation.
R package version 1.3.

Dimitromanolakis et al (2017). sim1000G: Easy to use multi marker genetic marker simulator in R for unrelated individuals or complex families

Epstein MP, et al. (2015). A Statistical Approach for Rare-Variant Association Testing in Affected Sibships.
Am. J. Hum. Genet. 96, 543-554.