Statistical Methods for Detecting and Interpreting Rare Variant Quantitative Trait Associations

Dajiang J. Liu

Doctoral Candidate

Rice University &

Baylor College of Medicine

PASI 2011 Xalapa, Mexico

Overview

- Hypotheses of common disease etiologies

- Methods for mapping rare variant/complex trait associations
  - How to detect associations
  - How to interpret identified associations
  - How to replicate identified associations

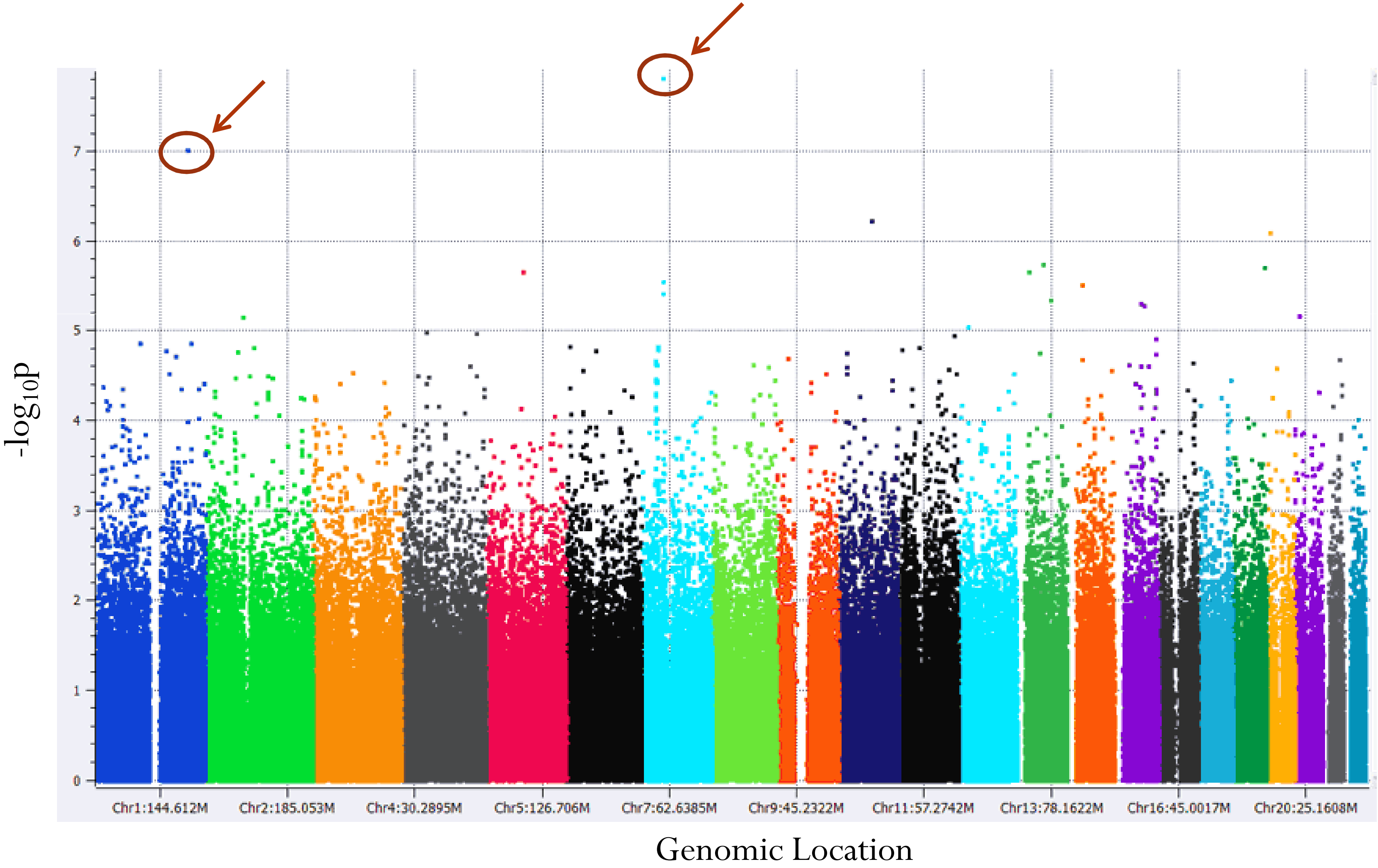
Statistical Genetics

- Gene mapping:
  - Aim:
    - Understand and characterize genetic architecture of complex traits
    - Find disease genes for complex human genetics using statistical approaches
  - Approach:
    - "Compares the inheritance pattern of a trait with the inheritance pattern of chromosomal regions"

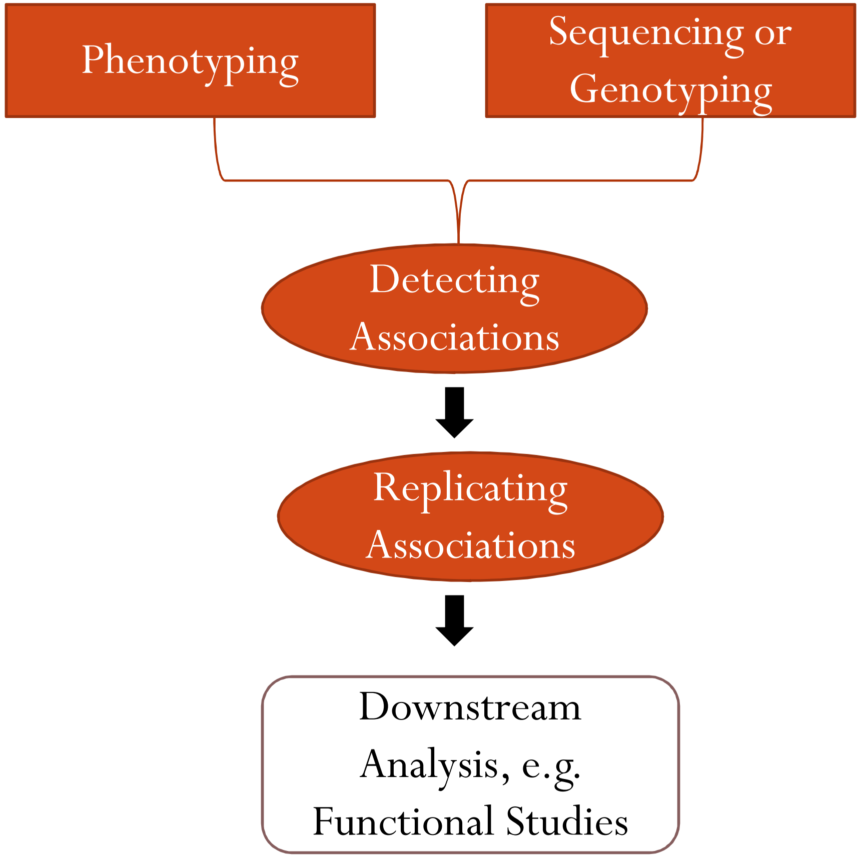
Statistical Genetics

- Association mapping
  - Linkage disequilibrium (LD) mapping
  - Genotype hundreds of thousands of genetic markers across the genome
  - Test the correlations between genetic markers with the phenotype of interest

Statistical Genetics



Association Analysis Pipeline



Phenotyping

Sequencing or Genotyping

Detecting Associations

Replicating Associations

Downstream Analysis, e.g. Functional Studies

Complex Trait Etiology Hypotheses

- Two parallel hypotheses
  - Common disease / common variants hypothesis
  - Common disease / rare variants hypothesis

Common Disease / Common Variants Hypothesis

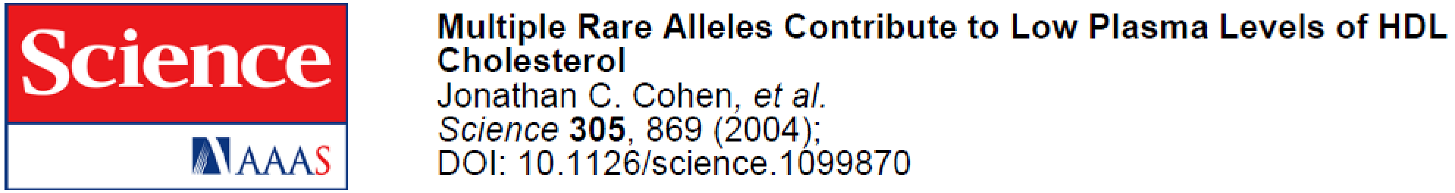
- Common diseases (traits) are caused by common variants with moderate effects
  - For binary traits, most identified variants have odds ratios (OR) <1.2
  - For quantitative trait, most indentified variants shift the mean trait value by $<0.05\sigma$
    - For human height trait of an American male,
      - $0.05\sigma=0.05\times2.8=0.14$ inches
- For most complex traits, the identified common variants only explain <10% of the heritability.

Common Disease / Rare Variants Hypothesis

- Common diseases are caused by multiple rare variants with larger genetic effects
  - Not large enough to cause familial aggregation
  - For binary trait, most rare variants have ORs of 2~4
    - Bodmer and Bonila *Nature Genetics 2008*
  - For quantitative trait, most variants shift QT mean by >0.1σ
    - Kryukov et al *PNAS 2009*

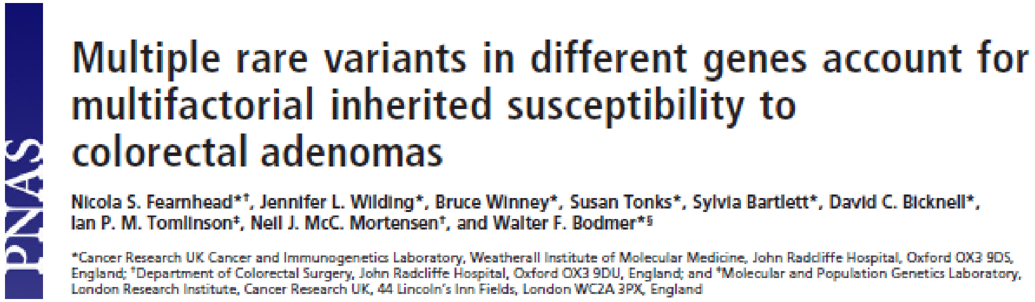
Common Disease Etiology Hypothesis II

- Examples for CD/RV
  - *ABCA1*, *APOA1*, and *LCAT*/low density lipoprotein (LDL)



Multiple Rare Alleles Contribute to Low Plasma Levels of HDL Cholesterol
Jonathan C. Cohen, *et al.*
Science **305**, 869 (2004);
DOI: 10.1126/science.1099870

  - *AXIN1*, *CTNNB1*, *hMLH1*, and *hMSH2*/ colorectal adenomas



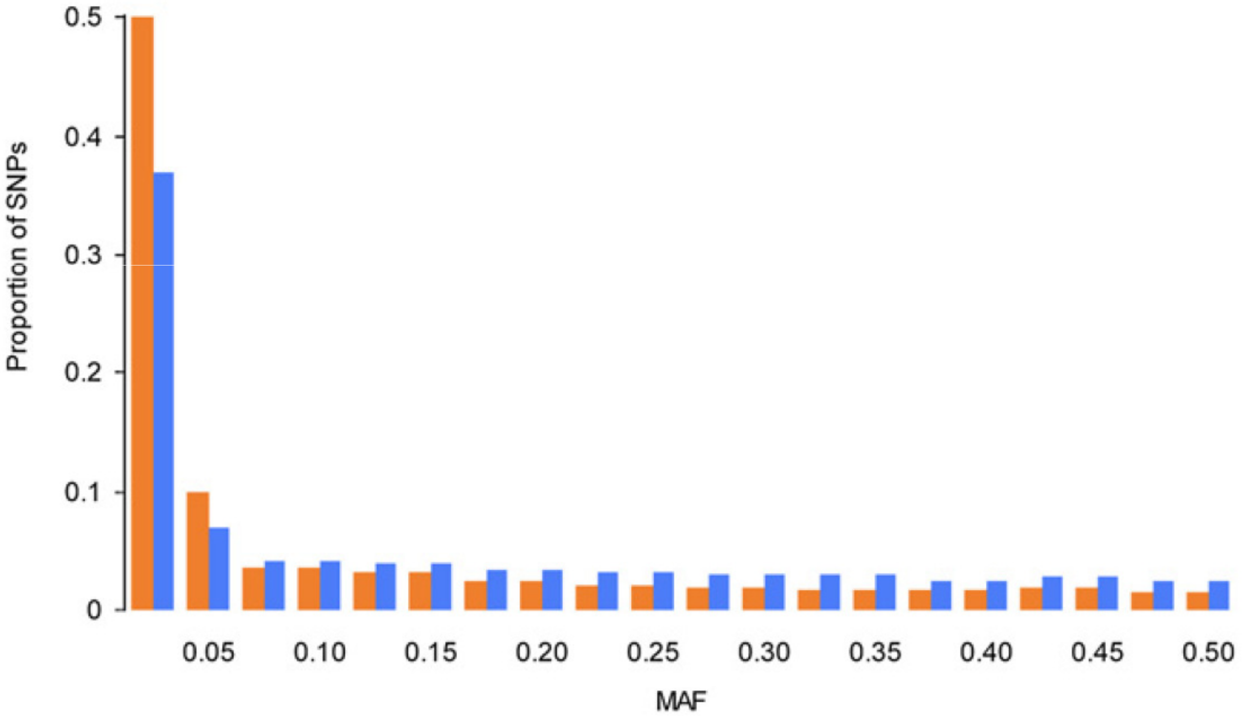
Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas

Nicola S. Feamhead[*,†], Jennifer L. Wilding[*], Bruce Winney[*], Susan Tonks[*], Sylvia Bartlett[*], David C. Bicknell[*], Ian P. M. Tomlinson[‡], Neil J. McC. Mortensen[†], and Walter F. Bodmer[*,§]

[*] Cancer Research UK Cancer and Immunogenetics Laboratory, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, England; [†] Department of Colorectal Surgery, John Radcliffe Hospital, Oxford OX3 9DU, England; and [‡] Molecular and Population Genetics Laboratory, London Research Institute, Cancer Research UK, 44 Lincoln's Inn Fields, London WC2A 3PX, England

Importance of Rare Variants

- Most genetic variants are "rare"

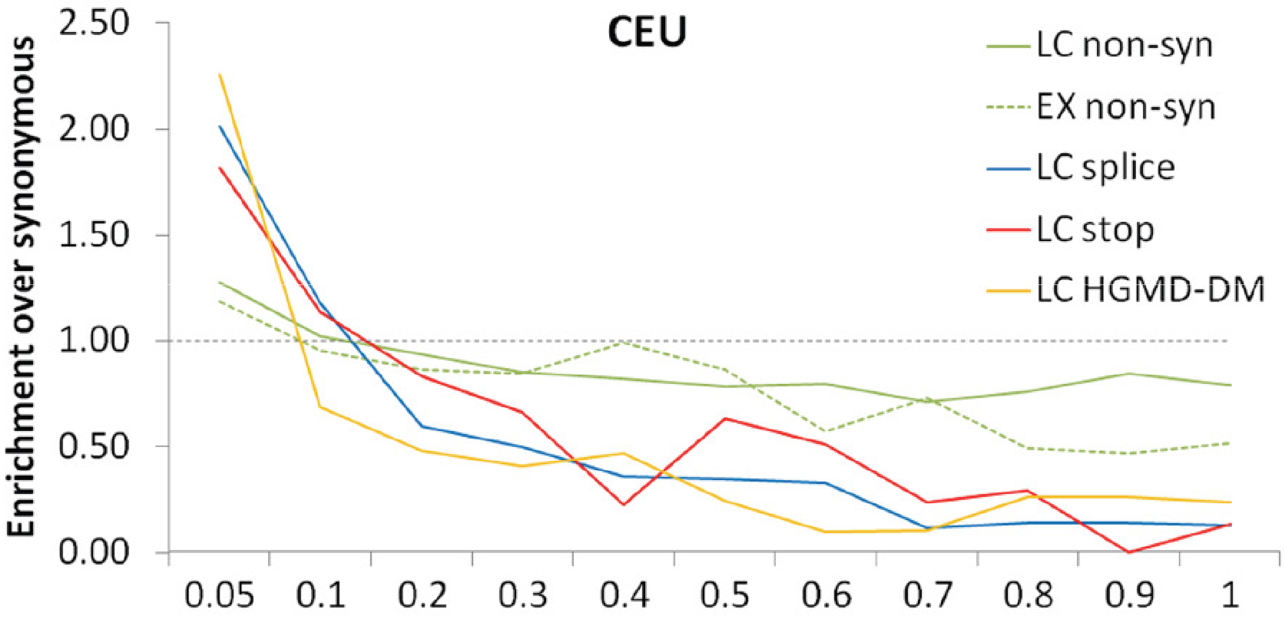


ENCODE III Site Frequency Spectrums

Gorlov et al *AJHG 2008*

Importance of Rare Variants

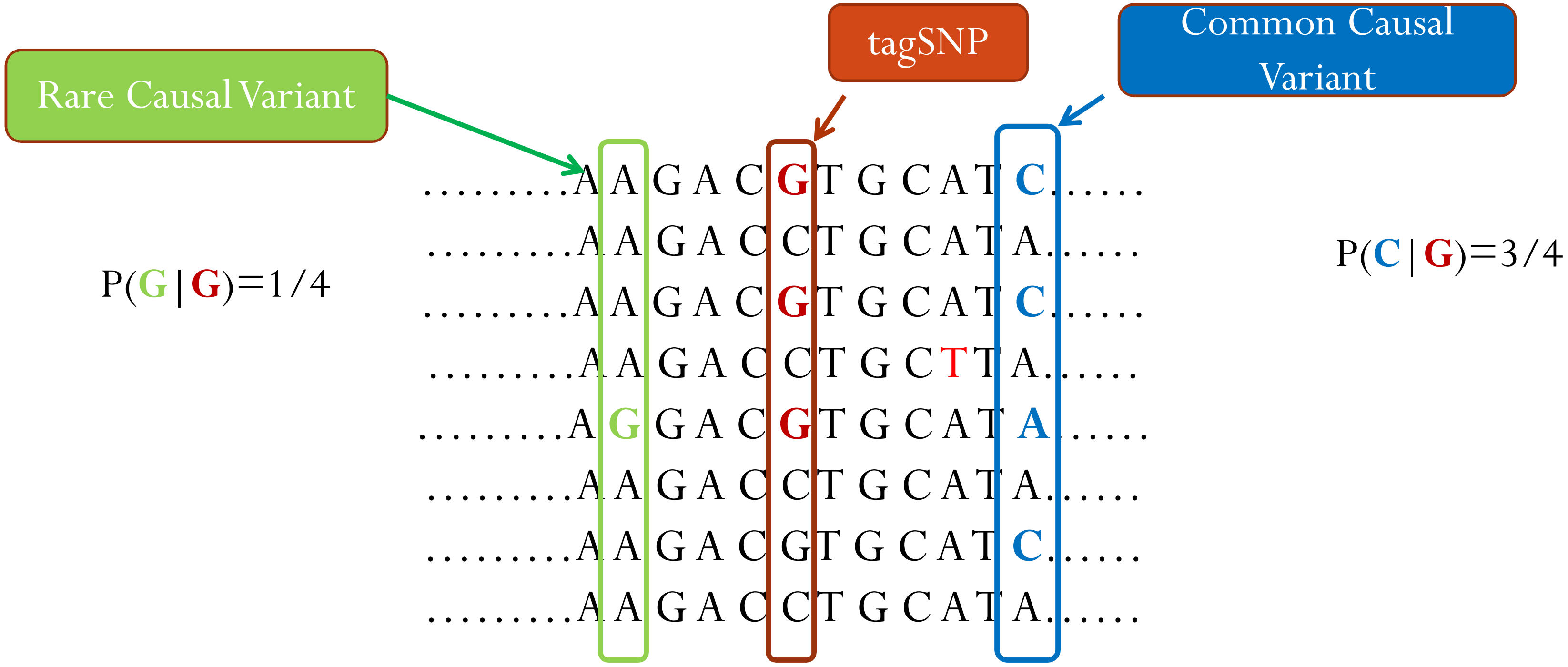
- Rare alleles are enriched with functional variants



- Most rare mis-sense mutations are functionally deleterious
  - Kryukov et al *AJHG 2007*

CD/CV and Indirect Association Mapping

- When CD/CV hypothesis holds
  - tagSNPs can be genotyped
  - Untyped common causal variants can be captured by tagSNPs
  - Test for the association between tagSNPs and phenotypes



Rare Causal Variant

tagSNP

Common Causal Variant

P(G|G)=1/4

P(C|G)=3/4

```
.........A A G A C G T G C A T C......
.........A A G A C C T G C A T A......
.........A A G A C G T G C A T C......
.........A A G A C C T G C T T A......
.........A G G A C G T G C A T A......
.........A A G A C C T G C A T A......
.........A A G A C G T G C A T C......
.........A A G A C C T G C A T A......
```

CD/RV and Direct Association Mapping

- When CD/RV hypothesis holds
  - Sequence entire genomic region
  - All genetic variants are uncovered
  - Variants are directly tested for their associations with the phenotype of interest

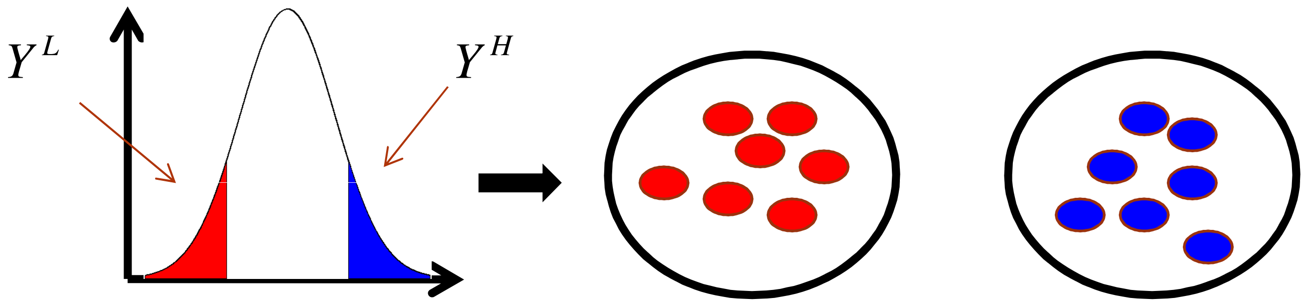
- Direct association mapping of rare variants is made possible by second generation sequencing and target enrichment technologies

Challenges in Sequence Based Genetic Studies

- High cost of sequencing
  - Especially when sequencing a large number of individuals at high coverage depth
- Non-negligible error rates
- Rare variants involved in complex traits are of
  - Moderate effect sizes
  - Low aggregated allele frequencies

Study Designs for Mapping Quantitative Traits

- To reduce sequencing and improve power
  - Sequence individuals with **extreme traits**



- For quantitative trait Y_i , to implement selective sampling
  - Set cutoffs Y^H and Y^L
  - Select N^H individuals with trait values $\geq Y^H$ and N^L individuals with trait values $\leq Y^L$

Study Designs for Mapping Quantitative Traits

- Combining and jointly analyzing publicly available cohort
  - ESP - Exome Sequencing Project
    - ESP2500 controls

NHLBI Grand Opportunity Exome Sequencing Project (ESP)

Methods for Mapping Rare Variants

- Methods for mapping common variants are underpowered

- Many methods have been developed for mapping rare variants
  - More powerful than
  - common variant analysis methods
  - Which analyzes variants one-by-one

- All methods are based upon omnibus test
  - Multiple rare variants in the gene region are jointly tested
  - To aggregate signal from multiple rare variants
  - Reduce the load of multiple testing

Methods for Mapping Rare Variants

- Challenges for rare variants tests
  - When multiple rare variants are jointly analyzed,
    - Presence of non-causal variants will reduce power

- Non-causal variants cannot be eliminated by bioinformatics tools
  - Low specificity and sensitivity for those tools
    - PolyPhen2, SIFT
  - Functionality does not imply causality

Methods for Mapping Rare Variants

- Strategies used to reduce the impact of non-causal variants
  - Weight or group variants
  - Variable selection based approaches
  - Random effects model based approaches

Methods for Mapping Rare Variants I: Fixed Effect Model

- Methods based upon grouping or weighting variants:
  - **Combined multivariate and collapsing (CMC)** Li and Leal *AJHG 2008*
  - **Weighted sum statistic (WSS)** Madsen and Browning *PLoS Genet 2009*
  - **Kernel based adaptive cluster (KBAC)** Liu and Leal *PLoS Genet 2010*
  - **Replication based test (RBT)** Ionita-Laza et al *PLoS Genet 2011*

Methods for Mapping Rare Variants I: Fixed Effect Model

- Variable selection based methods: select the best set of variants that explain the phenotype/genotype associations
  - **Variable threshold test (VT)** Price et al *AJHG 2010*
    - Motivated by population genetics
  - **RareCover method:** Bansal et al *PLoS Comp Bio 2010*
    - Greedy search algorithm
  - **Selective grouping method**: Zhang et al *PLoS ONE 2010*

  - **Comprehensive approach**: Hoffmann et al *PLoS ONE 2010*

Methods for Mapping Rare Variants II: Random Effects Model

- Genetic effects at different nucleotide sites are assumed to follow a (prior) distribution
- The null hypothesis is the (prior) distribution has zero variation

- **Goeman's empirical Bayesian score statistic (EBS):** Goeman et al *JRSSB 2004*
  - General testing framework for high dimensional data

- **Evolutionary Mixed Model for Pooled Association Testing (EMMPAT):** King et al *PLoS Genet 2011*
  - Incorporate evolutionary information from simulated data

- **C-alpha test:** Neale et al *PLoS Genet 2011*

Limitations of Existing Methods

- Most of the methods do not have a rigorous likelihood model which is crucial for
  - Making valid inferences
  - Estimating genetic parameters of interest

- Some methods do not allow controlling for covariates
  - E.g. WSS, RARECOVER, C-alpha, etc.

- Some methods are developed for mapping binary trait, and cannot analyze full quantitative trait information:
  - E.g. WSS, KBAC, RBT etc.

Limitations of Existing Methods

- Necessary to
  - Have a unifying framework which extends existing methods to quantitative trait analysis
  - Overcome (some of) the limitations
  - Make a comprehensive comparison of
    - Different rare variant tests, and
    - Their extensions in UNI-QTL framework

A Unifying Framework for Mapping Rare Variant Quantitative Trait Associations

UNI-QTL Framework

- Many existing fixed effect model based methods can be extended in a unifying likelihood framework for mapping rare variants in quantitative trait loci (UNI-QTL)
  - Liu, Banuelos and Leal *to be submitted, 2011*
- Joint model sampling ascertainment mechanisms and genotype-phenotype associations
- Allows efficient inferences and estimations of genetic parameters of interest

Notations

- Focus on quantitative trait mapping
  - Quantitative trait of interest or quantitative trait residuals after controlling for confounders: Y_i
  - Locus multi-site genotype:

$$\vec{X}_i = \left(x_i^1, x_i^2, \cdots, x_i^S\right)$$

  - Each element of the genotype vector is coded by an indicator:

$$x_i^s = \begin{cases} 1 & \text{if individual } i \text{ carries variants at site } s \\ 0 & \text{otherwise} \end{cases}$$

  - Define carrier frequencies

$$q^s = \Pr\left(x_i^s = 1\right); \; q = \sum_{s \in RV} q^s$$

UNI-QTL Framework

- Fixed effect models:

$$Y_i = \alpha_0 + \beta C\left(\vec{X}_i, Y_i\right) + \sum_j \alpha_j Z_{ij} + \varepsilon_i$$

- Existing methods can be incorporated through the coding function $C\left(\vec{X}_i, Y_i\right)$

UNI-QTL Framework

- To model sample ascertainment mechanisms, conditional likelihood is used:

$$\Pr\left(Y_i \middle| A_i = 1, \vec{X}_i; \beta, \vec{\alpha}\right) = \frac{\Pr\left(A_i = 1 \middle| Y_i, \vec{X}_i; \beta, \vec{\alpha}\right)\Pr\left(Y_i \middle| \vec{X}_i; \beta, \vec{\alpha}\right)}{\int \Pr\left(A_i = 1 \middle| y_i, \vec{X}_i; \beta, \vec{\alpha}\right)\Pr\left(y_i \middle| \vec{X}_i; \beta, \vec{\alpha}\right) dy_i}$$

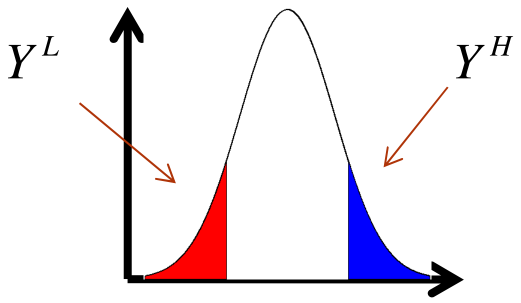
- A_i is the status of being sampled

UNI-QTL Framework

- For an extreme sampling study design that selects N^H individuals with trait values ≥ Y^H and N^L individuals with trait values ≤Y^L

$$\Pr\left(A_i = 1 \middle| Y_i, \vec{X}_i; \beta, \vec{\alpha}\right) = \Pr\left(A_i = 1 \middle| Y_i; \beta, \vec{\alpha}\right)$$

$$\propto \begin{cases} N^H / \Pr\left(Y_i \geq y^H\right) & \text{if } Y_i \geq y^H \\ 0 & \text{if } y^L < Y_i < y^H \\ N^L / \Pr\left(Y_i \leq y^L\right) & \text{if } Y_i \leq y^H \end{cases}$$



UNI-QTL Framework

- Extend the approach by Huang and Lin *AJHG 2007*
  - To the case of "unbalanced sampling"

$$N^H / \Pr\left(Y_i \geq y^H\right) \neq N^L / \Pr\left(Y_i \leq y^L\right)$$

- Unbalanced sampling frequently happens:
  - For example in Ahituv et al *AJHG 2007*
    - They sequenced:
      - 378 extremely obese individuals with BMI >95th percentile
      - 379 extremely lean individual with BMI <10th percentile

UNI-QTL Framework

- Association testing can be carried out by likelihood based score test
  - Numerically stable
    - Does not require maximization under the alternative hypothesis
  - Statistically efficient
    - Most powerful if the model is correctly specified

Extending Existing Rare Variant Tests

- Defining an auxiliary trait for each individual *i*,
  - If high extreme trait is of interest

$$Y_i^* = \begin{cases} 1 & Y_i \geq y^H \\ 0 & Y_i \leq y^L \end{cases}$$

  - On the other hand, if the lower extreme is of interest

$$Y_i^* = \begin{cases} 1 & Y_i \leq y^L \\ 0 & Y_i \geq y^H \end{cases}$$

- Compute the coding function using $\left\{\vec{X}_i, Y_i^*\right\}_i$

Extending Existing Rare Variant Tests

- Examples:
  - Collapsing coding (Li and Leal *AJHG 2008*):

$$C^{CMC}\left(\vec{X}_i, Y_i^*\right) = \delta\left(\sum\nolimits_{s\in RV} x_i^s > 0\right)$$

  - WSS coding (Madsen and Browning *PLoS Genet 2009*)

$$C^{WSS}\left(\vec{X}_i, Y_i^*\right) = \sum\nolimits_{s\in RV} w^s x_i^s$$

    - The weights are assigned based upon the allele frequency in one extremes
    - Lower frequency variants are assigned higher weights.

Extending Existing Rare Variant Tests

- Variable threshold test:
  - Define the coding function with respect to a (variable) frequency threshold

$$C_f^{VT}\left(\vec{X}_i, Y_i^*\right) = \delta\left(\sum_{s \in RV_f} x_i^s > 0\right)$$

  - The test statistic is defined by

$$T = \max_f T_f$$

Extending Existing Rare Variant Tests

- RARECOVER method
  - 1.) Set $RV = \Phi$, $U = \{1, 2, \cdots, S\}$, and

    $T_{old}^{RC} = 0 \quad T_{new}^{RC} = 0$
  - 2.) For each variant $u \in U \setminus RV$, calculate $\left\{C_u\left(\vec{X}_i\right) = \sum_{s \in RV + \{u\}} x_i^s\right\}_i$

    and the score statistic $T_u = S_{\hat{\theta}}\left(\left\{C_u\left(\vec{X}_i\right), Y_i\right\}_i\right)$.
  - 3.) Set $T_{old}^{RC} = T_{new}^{RC}$, and $T_{new}^{RC} = \max_u T_u$
  - 4.) Update $U = U \setminus \{u\}$
  - Repeat steps 2 to step 4 if $T_{new}^{RC} - T_{old}^{RC} > D$ and $U \neq \Phi$

  - The statistic for the dataset is given by $T^{RC} = T_{old}^{RC}$.

Extending Existing Rare Variant Tests

- KBAC (Liu and Leal *PLoS Genet 2010*)
  - Assign weights based upon the multi-site genotype;
  - Assume that there are *M* different multi-site genotypes, G_0, G_1, $\ldots G_M$

$$C^{KBAC}\left(\vec{X}_i, Y_i\right) = \sum_i K\left(\vec{X}_i = G_m\right)$$

  - Weights are assigned based upon the distribution of multi-site genotypes between samples from two extremes
  - Multi-site genotypes that are more enriched in one extreme is assigned higher weights.

Summary of Methods

Properties	Rare Variant Tests							
	CMC	ANRV	WSS	KBAC	VT	RARECOVER	RBT	C-alpha/SKAT
	Original/Extended	Original/Extended	Original/Extended	Original/Extended	Original/Extended	Original/Extended	Original/Extended	
Allow controlling for covariates?	Yes/Yes	Yes/Yes	No/Yes	Yes/Yes	Yes/Yes	No/Yes	No/Yes	No/Yes
Analyze full quantitative Trait Information?	Yes/Yes	Yes/Yes	No/Yes	Yes/Yes	Yes/Yes	No/Yes	No/Yes	No/Yes
Allow testing one-side hypothesis?	Yes/Yes	Yes/Yes	Yes/Yes	Yes/Yes	Yes/Yes	No/Yes	Yes/Yes	No/No
Allow analytic evaluation of statistical significance	Yes/Yes	Yes/Yes	No/No	No/No	No/No	No/No	No	No/Yes(??)

Comparisons of Rare Variant Tests

- Simulation Experiment
  - Using "realistic" population genetic and complex trait models

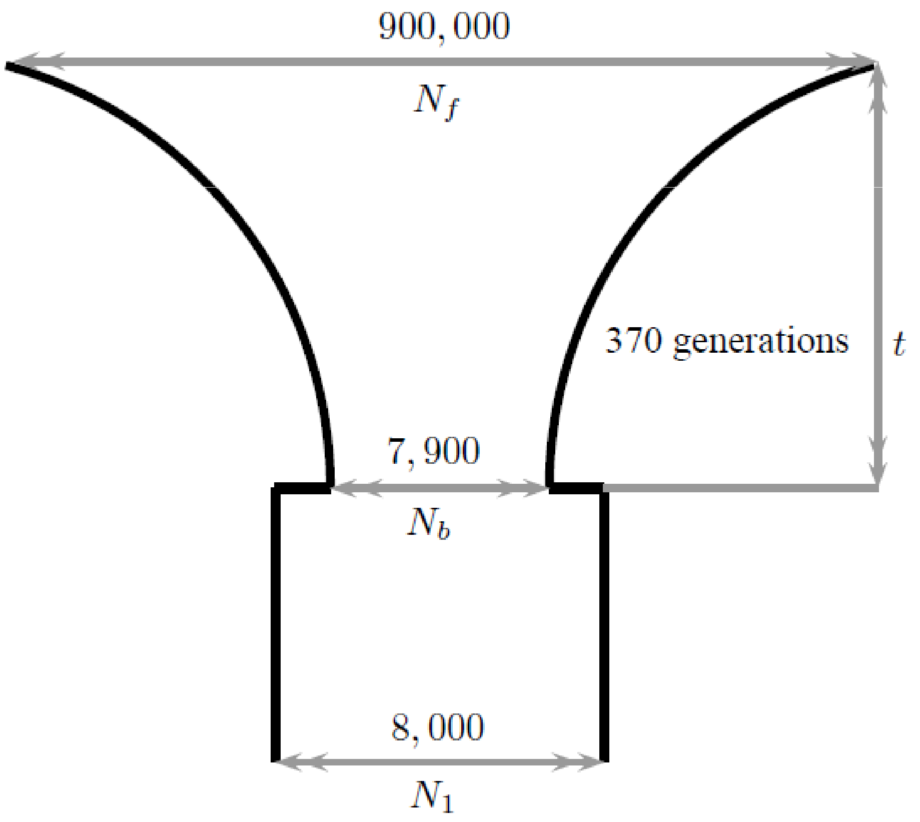
- Analysis of a sequence dataset from the Dallas Heart Study
  - *ANGPTL3*, *ANGPTL4*, *ANGPTL5* and *ANGPTL6* genes

Comparisons of Rare Variant Tests

- Eight tests are compared
  - Eight tests are generalized in the UNI-QTL framework
    - CMC-ProScore vs. CMC-UNIQTL
    - ANRV-ProScore vs. ANRV-UNIQTL
    - VT-ProScore vs. VT-UNIQTL
    - WSS-BINARY vs. WSS-UNIQTL
    - KBAC-BINARY vs. KBAC-UNIQTL
    - RARECOVER-BINARY vs. RARECOVER-UNIQTL
    - RBT-BINARY vs. RBT-UNIQTL
    - C-alpha vs. EBS

Population Genetic Model

- Demographic history of European population
  - Kryukov et al *PNAS 2009*



Simulation of Rare Variant Data

- Mutation rate
  - $\mu_S = 1.8 \times 10^{-8}$ per nucleotide site per generation
- Locus length
  - 1500 base pairs
  - Average gene coding region length
- Analyze only "non-synonymous" variants with minor allele frequency (<3%)
- Purifying selection is incorporated, and modeled as Gamma distribution

Simulation of Quantitative Traits

- Phenotypic model I:
  - Assuming genetic effects for causative variants is independent of their fitness:
    - Three different proportions of non-causal variants are used
      - 20%
      - 50%
      - 80%

Simulation of Quantitative Traits

- Phenotypic model II:
  - Relating genetic effects of variants with their fitness (selection coefficients)
    - Scenarios with different selection coefficient cutoffs are used
      - Variants with selection coefficients $>10^{-2}$ are causal
      - Variants with selection coefficients $>10^{-3}$ are causal
      - Variants with selection coefficients $>10^{-4}$ are causal

Simulation of Quantitative Traits

- Quantitative traits are simulated according to

$$Y_i \sim \mathrm{N}\left(\tilde{\alpha} + \sum\nolimits_{s \in \mathrm{CV}} \tilde{\beta} x_i^s, \tau^2\right)$$

- CV ~ the set of causal variants

- Parameters are chosen as follows:
  - Type I error evaluation:

$$\tilde{\alpha} = 0, \tilde{\beta} = 0$$

  - Power comparisons:
    - Two locus genetic effects are used:

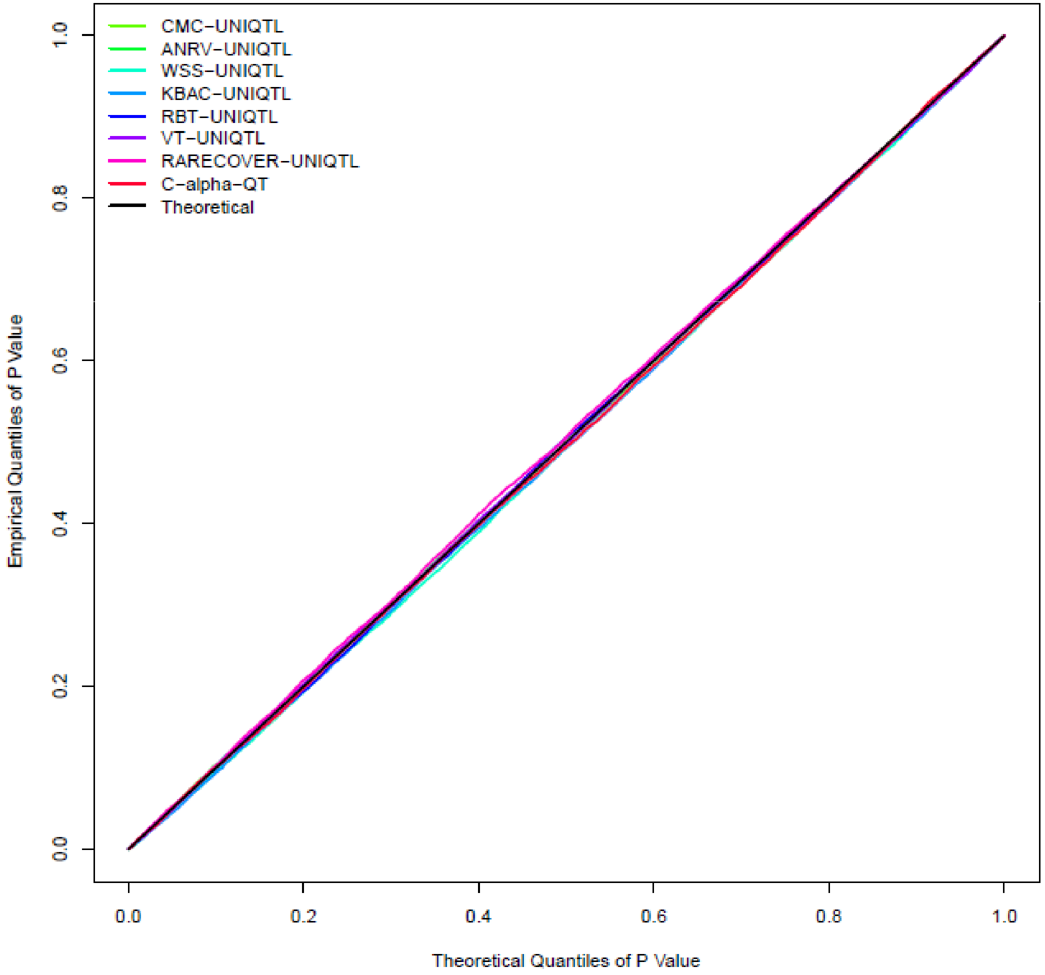
$$\tilde{\alpha} = 0, \tilde{\beta} = 0.25\tau \text{ or } 0.5\tau$$

Simulation of Quantitative Traits

- A cohort of 20000 individuals is used for selective sampling
- 2000 individuals from each extreme are selected and sequenced
- Two sided hypothesis is tested
  - $\alpha=0.05$
- Statistical significance for CMC-UNIQTL and ANRV-UNIQTL is evaluated analytically
  - Significance for all other tests were evaluated through permutations
- Analyze variants with MAF<1%

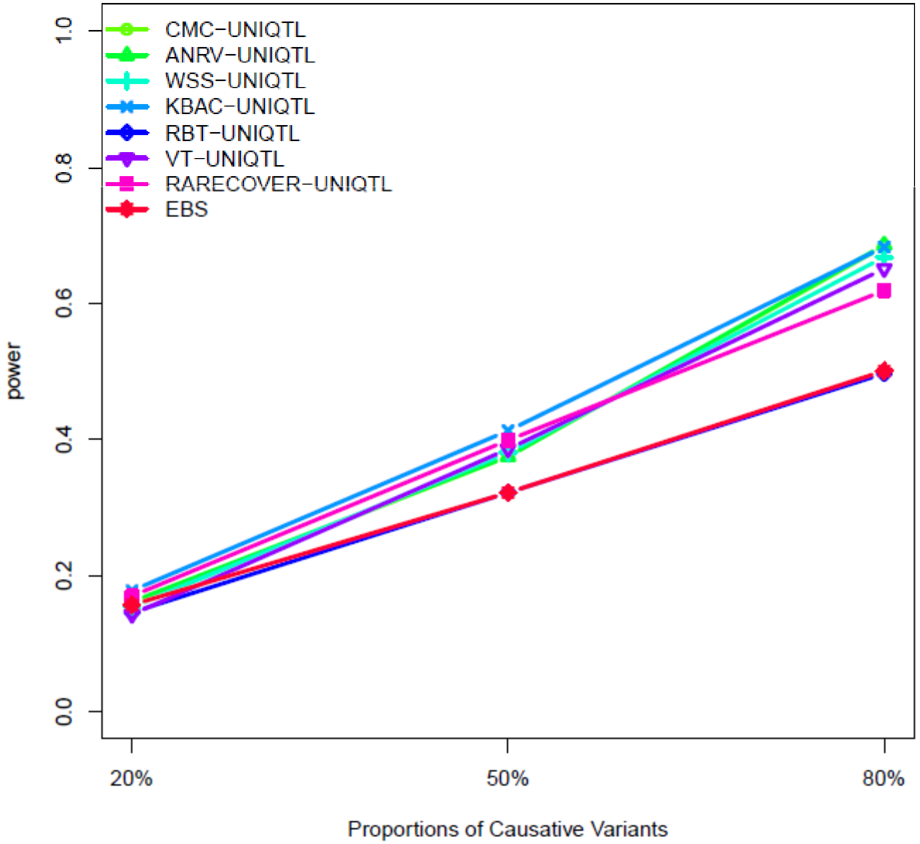
Type I error Evaluation

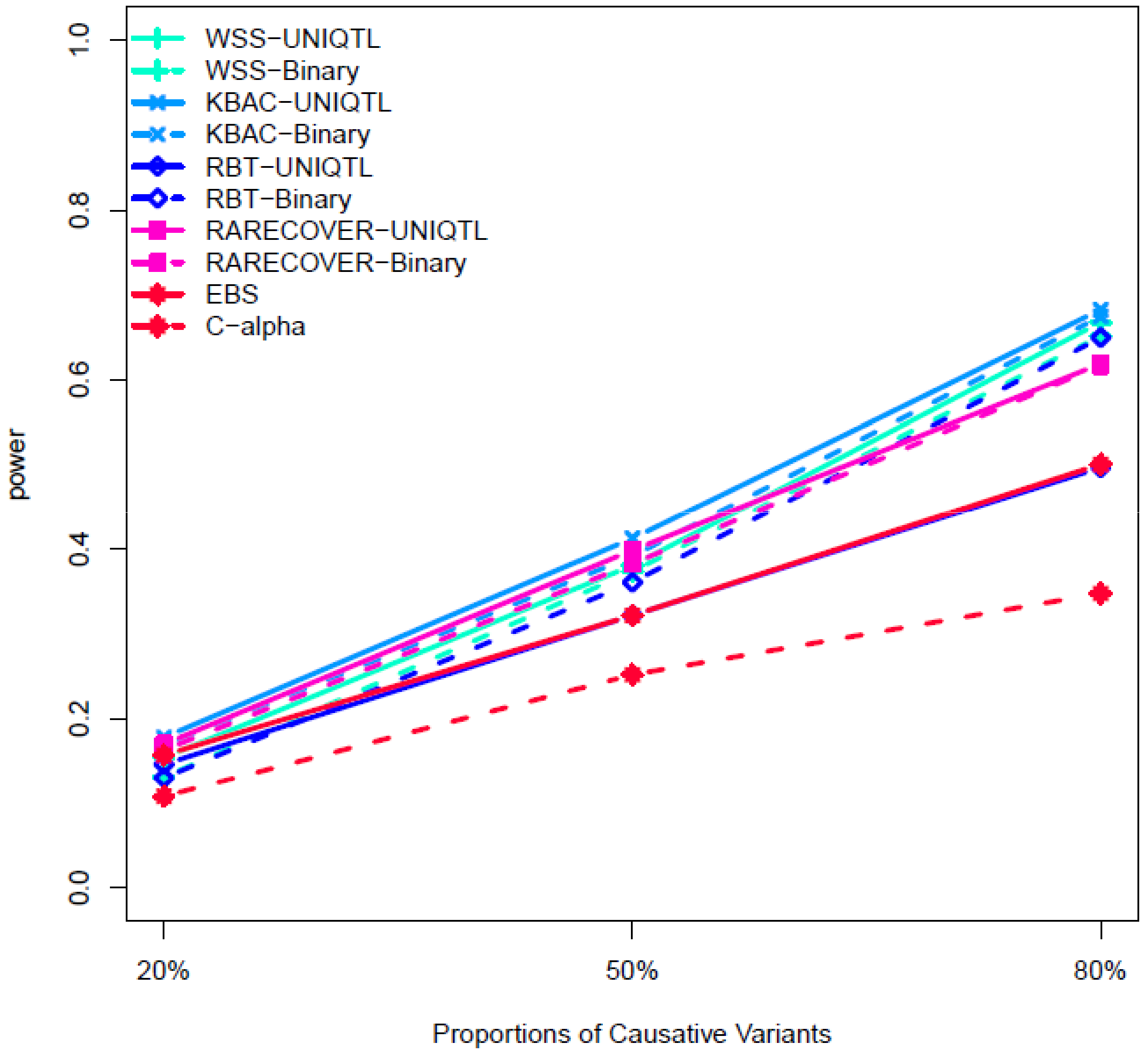
- QQ plot obtained using 5000 replicates



Power Comparisons I: Quantitative Traits

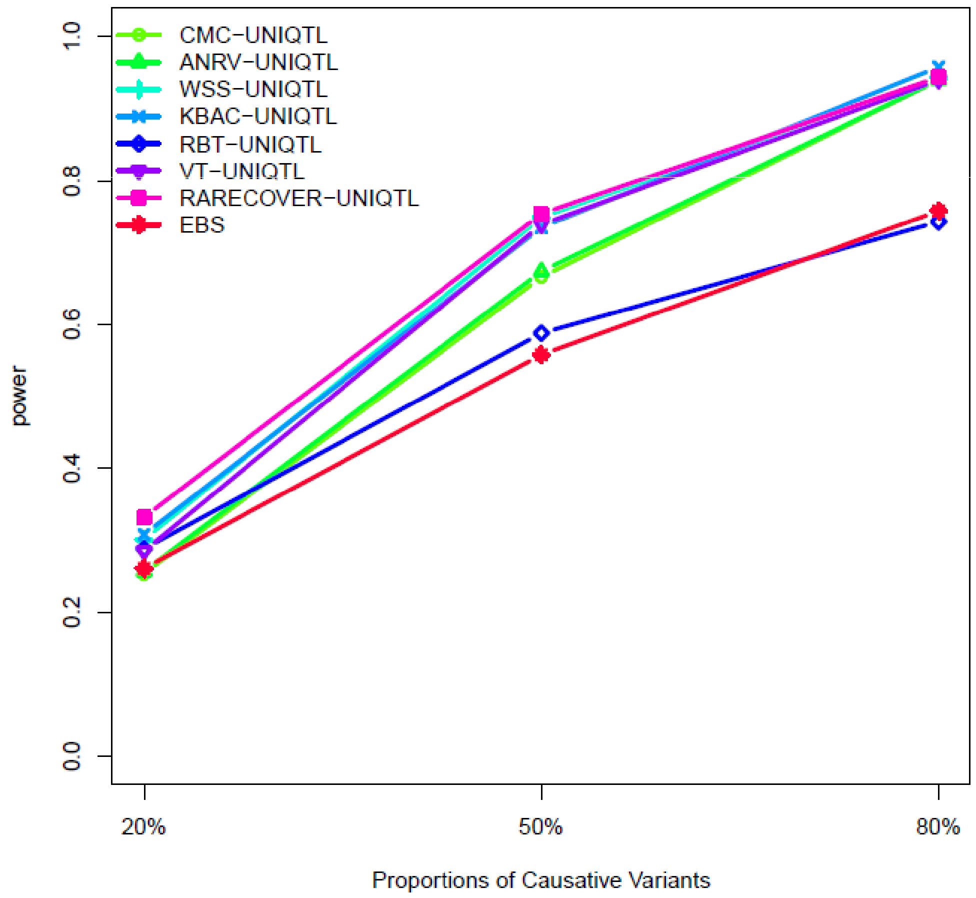
- Phenotypic model I which assumes independence between fitness and genetic effects
- $\tilde{\beta} = 0.25$

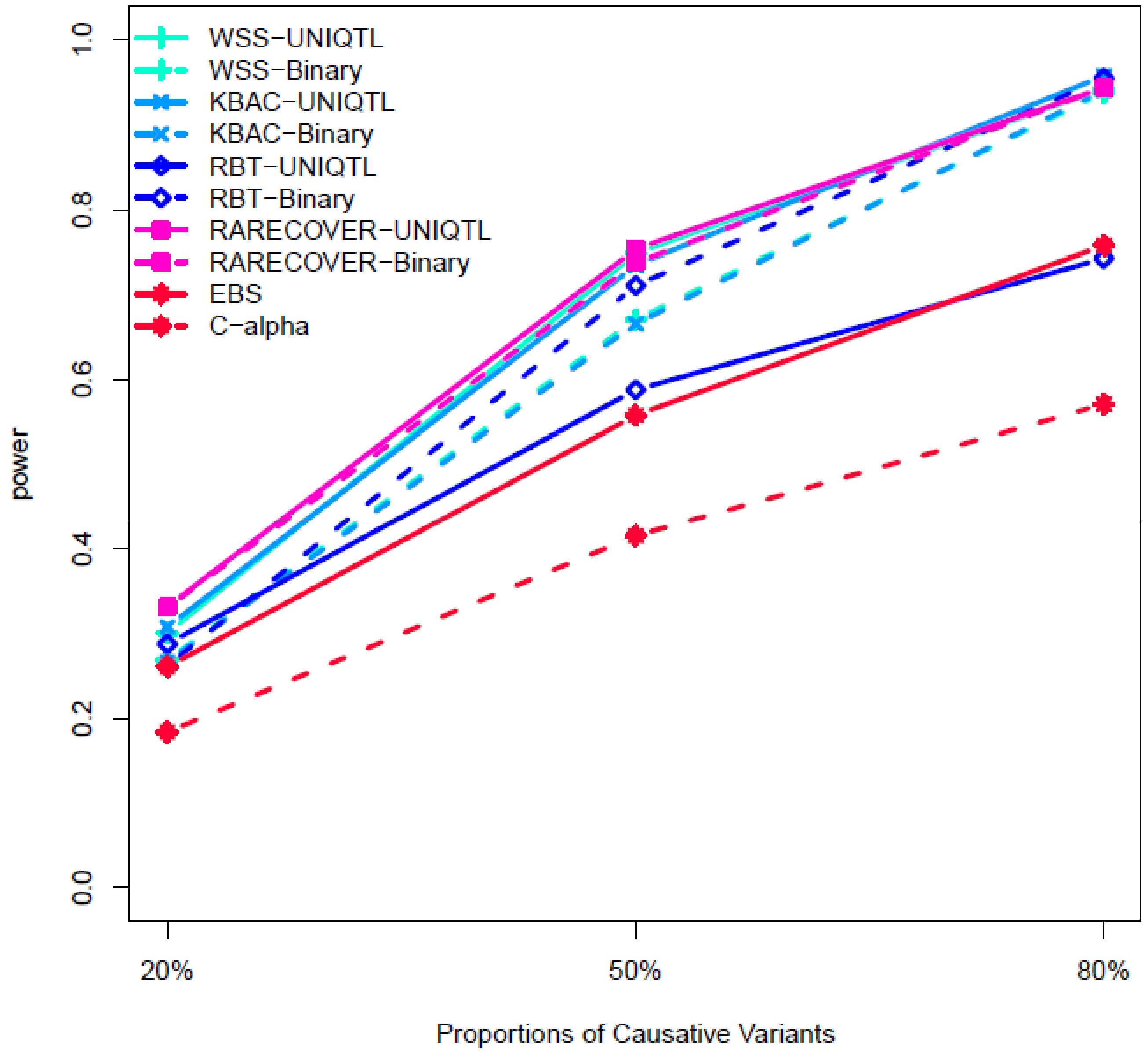



WSS−UNIQTL
WSS−Binary
KBAC−UNIQTL
KBAC−Binary
RBT−UNIQTL
RBT−Binary
RARECOVER−UNIQTL
RARECOVER−Binary
EBS
C−alpha
power
0.0
0.2
0.4
0.6
0.8
1.0
20%
50%
80%
Proportions of Causative Variants

Power Comparisons I: Quantitative Traits

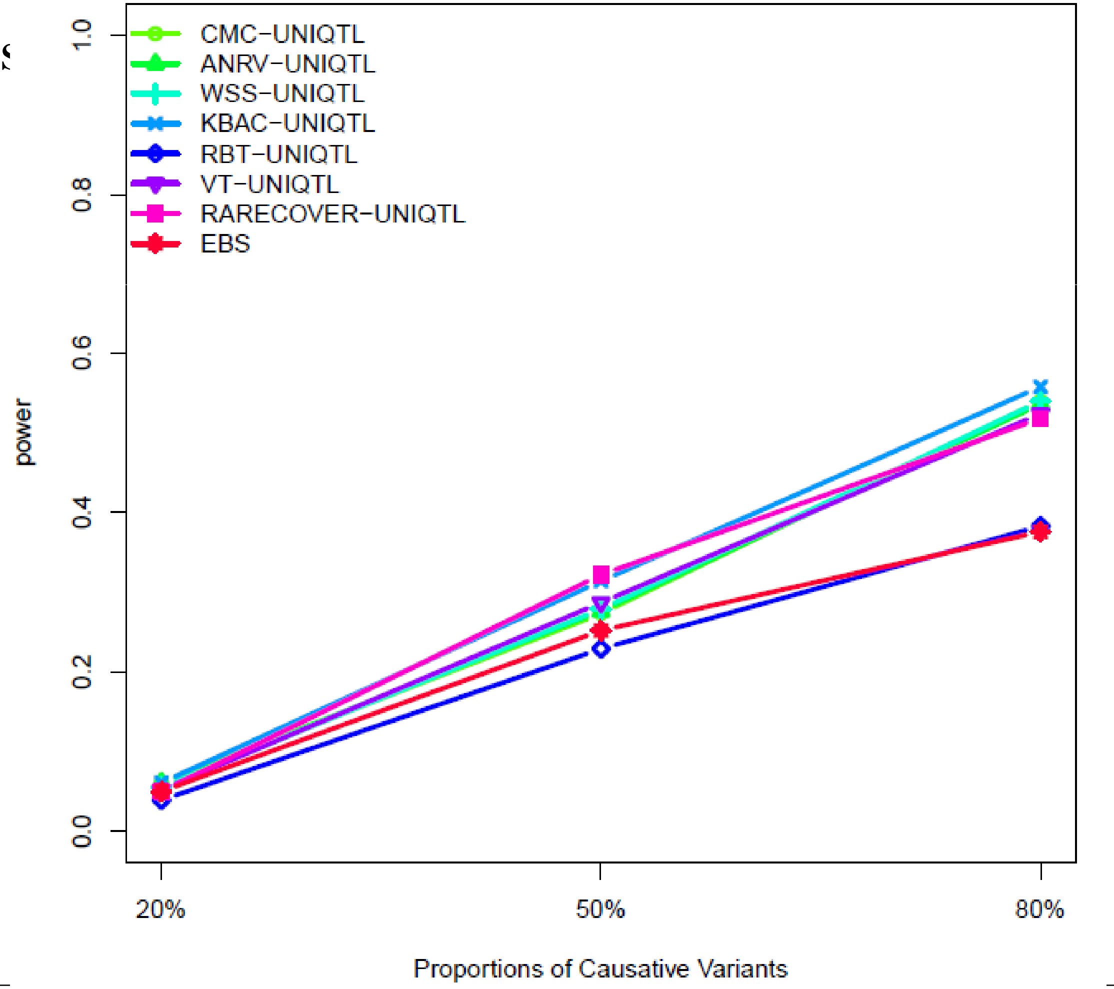
- Phenotypic model I which assumes independence between fitness and genetic effects
- $\tilde{\beta} = 0.5$

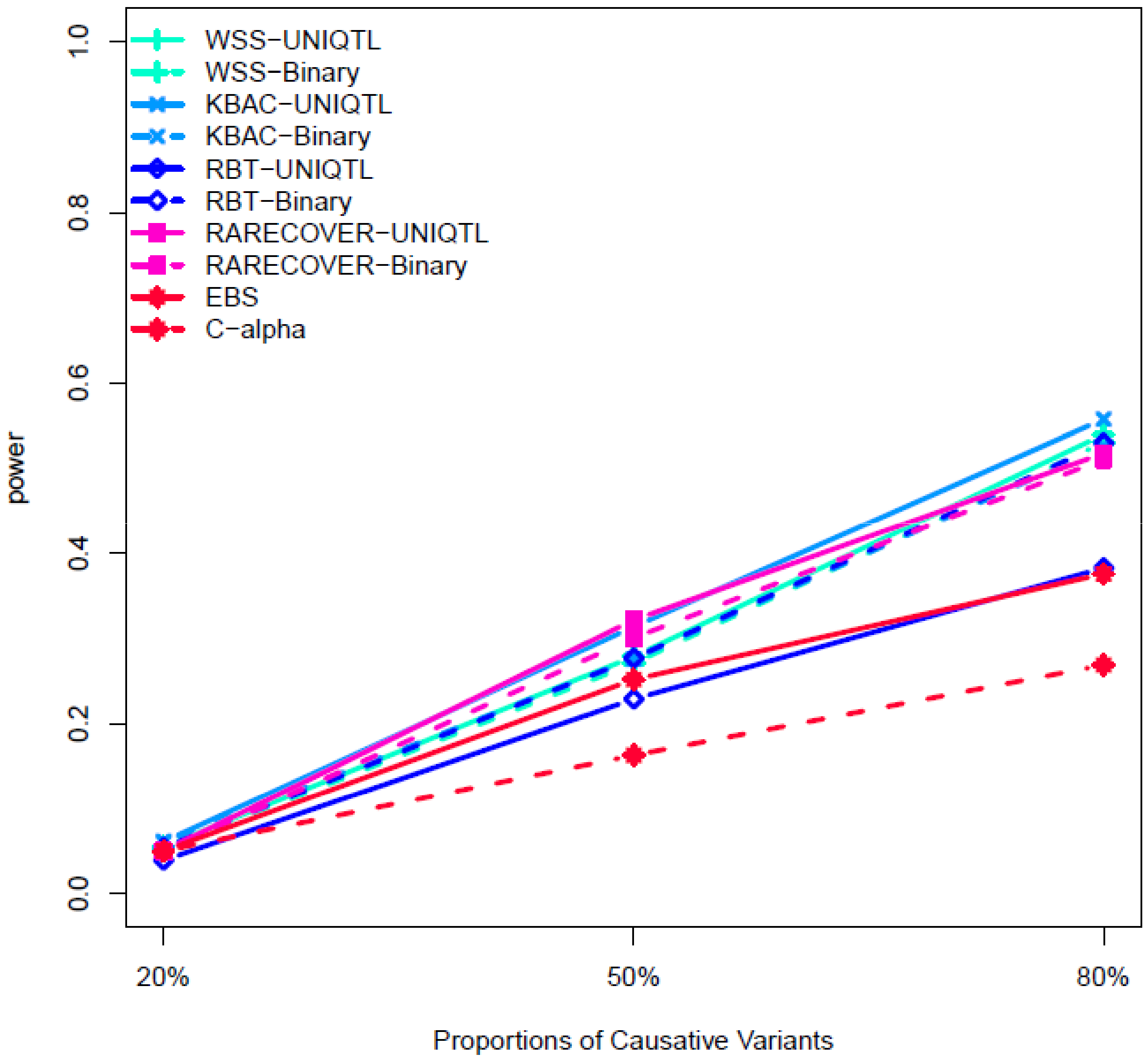



WSS−UNIQTL
WSS−Binary
KBAC−UNIQTL
KBAC−Binary
RBT−UNIQTL
RBT−Binary
RARECOVER−UNIQTL
RARECOVER−Binary
EBS
C−alpha
power
0.0
0.2
0.4
0.6
0.8
1.0
20%
50%
80%
Proportions of Causative Variants

Power Comparisons I: Quantitative Traits

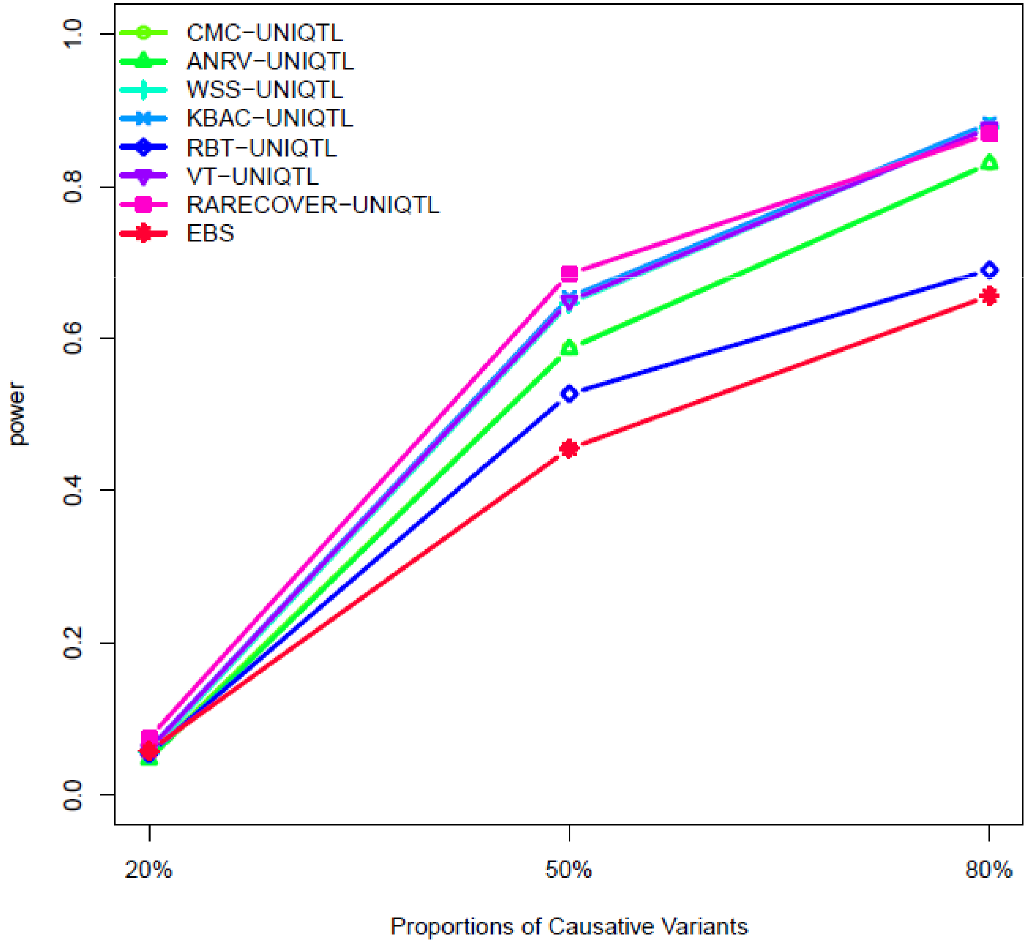
- Phenotypic model II which relates genetic effects of variants with their fitnes
- $\tilde{\beta} = 0.25$

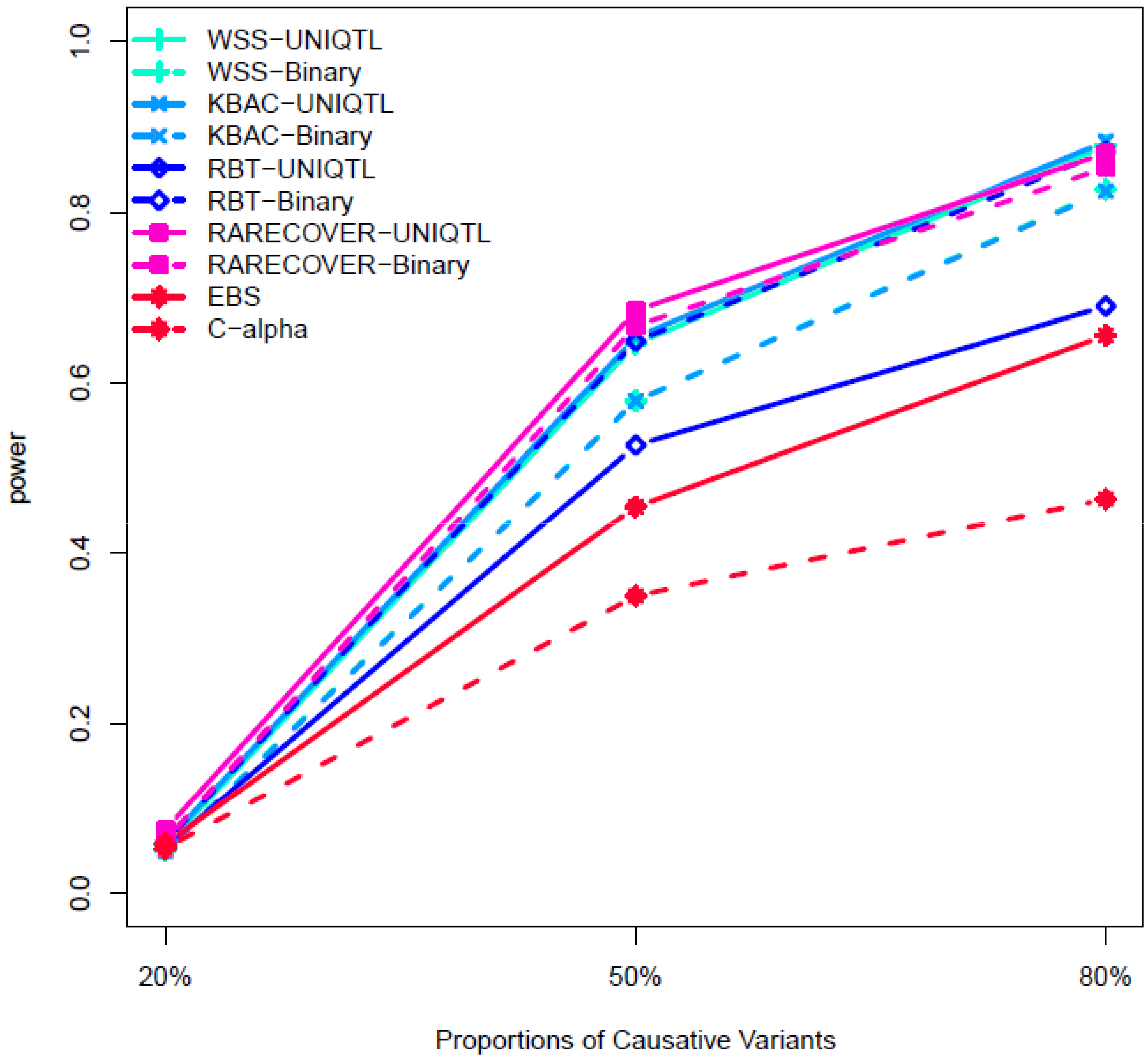



WSS−UNIQTL
WSS−Binary
KBAC−UNIQTL
KBAC−Binary
RBT−UNIQTL
RBT−Binary
RARECOVER−UNIQTL
RARECOVER−Binary
EBS
C−alpha
power
0.0
0.2
0.4
0.6
0.8
1.0
20%
50%
80%
Proportions of Causative Variants

Power Comparisons I: Quantitative Traits

- Phenotypic model II which relates genetic effects with their fitness
- $\tilde{\beta} = 0.5$




WSS−UNIQTL
WSS−Binary
KBAC−UNIQTL
KBAC−Binary
RBT−UNIQTL
RBT−Binary
RARECOVER−UNIQTL
RARECOVER−Binary
EBS
C−alpha
power
0.0
0.2
0.4
0.6
0.8
1.0
20%
50%
80%
Proportions of Causative Variants

Results

- Extended tests consistently outperform the original tests
  - Due to analyzing full quantitative trait
  - Due to the use of a likelihood based method which jointly models
    - Sampling mechanisms
    - Genotype-quantitative trait associations

- There does not exist a uniformly most powerful test

- The extended VT, WSS and KBAC and original RBT test perform well under a wide variety of scenarios

- The difference in power between different tests are small.

Analysis of Dallas Heart Study Dataset

- Dallas Heart Study is a population based study which consists of 3551 participants

- Nine phenotypes were measured:
  - Body mass index (BMI)
  - High density lipoprotein (HDL)
  - Low density lipoprotein (LDL)
  - Very low density lipoprotein (VLDL)
  - Triglyceride (TG)
  - Systolic blood pressure (SysBP)
  - Diastolic blood pressure (DiasBP)
  - Glucose level (Gluc)

Analysis of Dallas Heart Study Dataset

- Re-sequencing dataset of *ANGPTL3*, *ANGPTL4*, *ANGPTL5*, and *ANGPTL6*
- Exon and intron-exon boundaries are sequenced
- A total of 384 variant nucleotide sites are uncovered
- Most of the variants are rare, with MAF<1%

Analysis of Dallas Heart Study Dataset

- Within each race/sex stratus
  - Samples are quantile normalized

- For each phenotype,
  - Individuals with trait values >75th percentile and <25th percentile are used

- Non-synonymous variants with MAF<3% are analyzed

Results:

GENE	Trait	CMC-UNIQTL[a]	ANRV-UNIQTL[a]	WSS[b,c] (UNIQTL| Binary)	KBAC[b,c] (UNIQTL| Binary)	RBT[b,c] (UNIQTL| Binary)	VT-UNIQTL[b]	RARECOVER[b] (UNIQTL| Binary)	EBS|C-alpha[b,c]
ANGPTL3	VLDL	0.064	0.054	0.17 | 0.042	0.048 | 0.02	0.522 | 0.174	0.102	0.568 | 0.176	0.043 | 0.036
ANGPTL4	TG	0.007	0.001	0.018 | 0.006	0.001 | 0.006	0.004 | 0.014	0.006	0.004 | 0.004	0.004 | 0.008
ANGPTL4	VLDL	0.017	0.005	0.04 | 0.013	0.024 | 0.016	0.01 | 0.068	0.018	0.024 | 0.062	0.012 | 0.022
ANGPTL5	BMI	0.004	0.017	0.002 | 0.004	0.002 | 0.004	0.26 | 0.01	0.022	0.016 | 0.086	0.32 | 0.252
ANGPTL5	HDL	0.038	0.031	0.102 | 0.18	0.053 | 0.178	0.028 | 0.314	0.136	0.044 | 0.238	0.032 | 0.158
ANGPTL6	BMI	0.023	0.018	0.006 | 0.125	0.042 | 0.206	0.162 | 0.106	0.036	0.154 | 0.138	0.35 | 0.644

How to Interpret Identified Associations

A Framework to Interpret Identified Associations

- Important to interpret identified associations
  - Estimate genetic parameters of interest
  - Quantify the proportion of missing heritabilities

- Estimated genetic parameters are important for
  - Making risk predictions
  - Designing replication studies

- Based upon Liu and Leal 2011 in preparation

Quantitative Trait Models

- Quantitative trait is assumed to follow

$$Y_i \sim \mathrm{N}\left(\tilde{\alpha} + \sum\nolimits_{s \in \mathrm{CV}} \tilde{\beta}^s x_i^s, \tau^2\right)$$

- CV is the set of causative variants
  - Unknown in real applications

- Total causative variants carrier frequency

$$q^{CV} = \sum\nolimits_{s \in \mathrm{CV}} q^s$$

Genetic Parameters of Interest

- Two parameters are of interest
  - (Causative) variants genetic effects: $\left\{\tilde{\beta}^s\right\}_{s=1,\cdots,S}$
  - Locus Genetic Variance

$$\sigma^2 = \sum\nolimits_{s\in \mathrm{CV}} \left(\tilde{\beta}^s\right)^2 q^s\left(1-q^s\right)$$

- Challenges:
  - Two quantities cannot be directly estimated
    - The set of causal variants are unknown
    - Rare variants can not be analyzed individually
      - Not powerful
      - Numerically unstable

Locus Average Effect I

- Instead of estimating each variant individually, locus average effect is defined, i.e.

$$\beta_{LAE} = \mathrm{E}\left(Y_i \middle| \sum_s x_i^s > 0\right) - \mathrm{E}\left(Y_i \middle| \sum_s x_i^s = 0\right)$$

- Mean quantitative trait difference between carriers and non-carriers

- β_{LAE} Can be efficiently estimated using the UNI-QTL model with CMC coding

Locus Average Effect II

- Define locus average effect induced genetic variance

$$\sigma_{LAE}^2 = (\beta_{LAE})^2 q(1-q)$$

- σ_{LAE}^2 can also be efficiently estimated using the UNI-QTL model

- **Theorem:** $\sigma_{LAE}^2 \leq \sigma^2$ with equality hold when all locus genetic variants are causal.

- Therefore, although locus genetic effects cannot be directly estimated, its **lower bound** can be efficiently estimated

Locus Average Effect III

- Variants involved in complex traits usually have moderate effect sizes

- If an upper bound for causative variant effects can be assumed, i.e.

$$\left|\tilde{\beta}^{s}\right| \leq \tilde{\beta}^{\max}, \text{ for all } s \in \text{CV}$$

- An upper bound for the locus genetic variance can also be efficiently estimated as a function of $\tilde{\beta}^{\max}$, i.e.

$$\hat{\sigma}^{2}_{\max} = \tilde{\beta}^{\max}\hat{\beta}_{LAE}q\left(1 - \frac{\hat{\beta}_{LAE}\hat{q}}{\tilde{\beta}^{\max}}\right)$$

Estimating Locus Average Effect

- If the genetic parameters are estimated using the same sample where the association was identified
  - The naïve estimates $\hat{\beta}_{LAE,naive}, \hat{\sigma}^2_{LAE,naive}$ can be seriously inflated
  - Winner's curse
    - "The winner of a bid tends to overpay, and is thus cursed"

- The bias due to winner's curse can be large for poorly powered genetic studies

Estimating Locus Average Effect

- In order to reduce the bias for winner's curse
    - A bootstrap-sample-split algorithm (BSS) is developed
        - Extend the method in Sun and Bull *Genetic Epi 2006*
- The bias due to winner's curse can be estimated and corrected with the BSS procedure.
- BSS algorithm is generic
    - can be applied to associations identified by any rare variant test ***T***

BSS Algorithm I

- For a sample with N^H individuals having trait values ≥ Y^H and N^L individuals having trait values ≤Y^L, and significance level α

- **Step 1: Obtain the naïve estimators** $\hat{\beta}_{LAE,naive}, \hat{q}_{naive}$

Repeat step 2-4 K times, for each k,

- **Step 2: Obtain a bootstrap sample B_k, and the residual sample is denoted by C_k**
  - B_k also consists of with N^H individuals having trait values ≥ Y^H and N^L individuals having trait values ≤Y^L

BSS Algorithm II

- **Step 3: Analyze the bootstrap sample *B_k* with test *T* and CMC-UNIQTL, and denote the p-value by** $P_{B_k}^{T}$ **and** $P_{B_k}^{CMC-UNIQTL}$

- **Step 4: Obtain estimates using sample *B_k* and *C_k*, the estimates are denoted by** $\left\{\hat{\beta}_{LAE,B_k}, \hat{q}_{B_k}\right\}$and $\left\{\hat{\beta}_{LAE,C_k}, \hat{q}_{C_k}\right\}$

BSS Algorithm III

- The bias due to winner's curse is given by

$$\hat{\mu}_\beta = \frac{\sum_k \left(\hat{\beta}_{LAE,B_k} - \hat{\beta}_{LAE,C_k}\right)\delta\left(P_{B_k}^T < \alpha, P_{B_k}^{CMC-UNIQTL} < \alpha\right)}{\sum_k \delta\left(P_{B_k}^T < \alpha, P_{B_k}^{CMC-UNIQTL} < \alpha\right)}$$

$$\hat{\mu}_q = \frac{\sum_k \left(\hat{q}_{B_k} - \hat{q}_{C_k}\right)\delta\left(P_{B_k}^T < \alpha, P_{B_k}^{CMC-UNIQTL} < \alpha\right)}{\sum_k \delta\left(P_{B_k}^T < \alpha, P_{B_k}^{CMC-UNIQTL} < \alpha\right)}$$

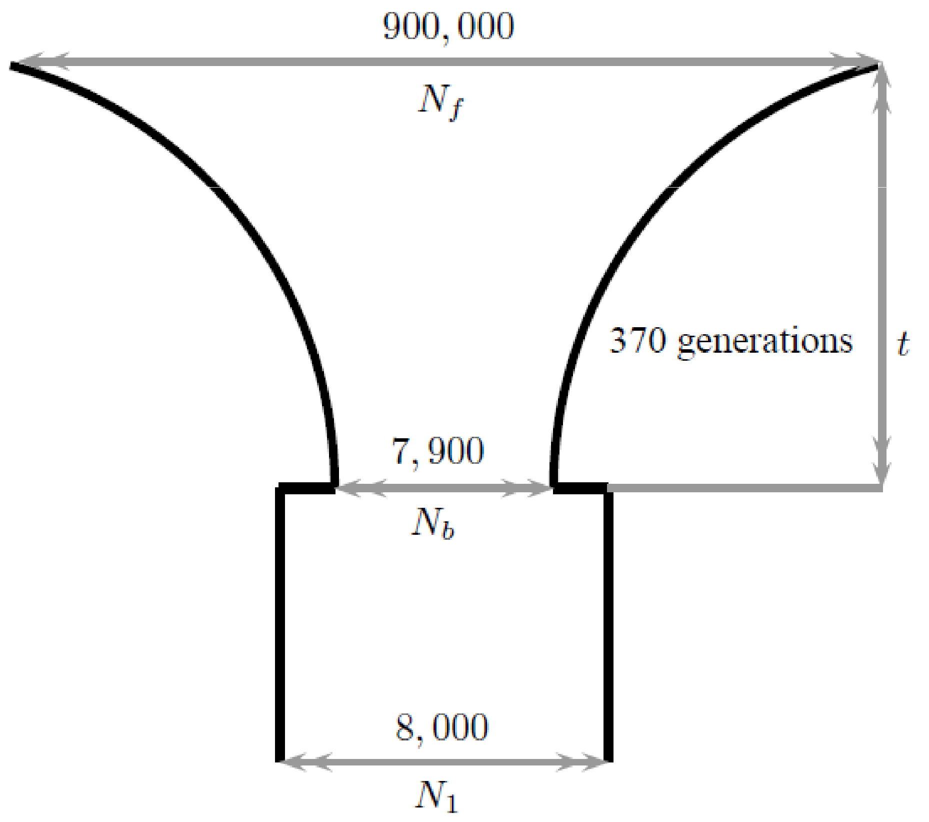
- The corrected estimator is given by

$$\hat{\beta}_{LAE,BSS} = \hat{\beta}_{LAE,naive} - \hat{\mu}_\beta$$

Simulation Experiment

Population Genetic Model

- Demographic history of European population
  - Kryukov et al *PNAS 2009*



Simulation of Rare Variant Data

- Mutation rate
  - μ_S=1.8×10^{-8} per nucleotide site per generation
- Locus length
  - 1500 base pairs
  - Average gene coding region length
- Analyze only "non-synonymous" variants with minor allele frequency (<3%)
- Purifying selection is incorporated, and modeled as Gamma distribution

Simulation of Quantitative Traits

- Phenotypic model I:
  - Assuming genetic effects for causative variants is independent of their fitness:
    - Two different proportions of causal variants are used
      - 50%
      - 80%

Simulation of Quantitative Traits

- Phenotypic model II:
  - Relating genetic effects of variants with their fitness (selection coefficients)
    - Scenarios with different selection coefficient cutoffs are used
      - Variants with selection coefficients $>10^{-3}$ are causal
      - Variants with selection coefficients $>10^{-4}$ are causal

Simulation of Quantitative Traits

- Quantitative traits are simulated according to

$$Y_i \sim \mathrm{N}\left(\tilde{\alpha} + \sum\nolimits_{s \in \mathrm{CV}} \tilde{\beta} x_i^s, \tau^2\right)$$

- Parameters are chosen as follows:

$$\tilde{\alpha} = 0, \tilde{\beta} = 0.25\tau \text{ or } 0.5\tau$$

Simulation of Quantitative Traits

- A cohort of 20000 individuals is used for selective sampling

- For small scale candidate gene studies,
  - 500 individuals from each extreme are selected and sequenced
  - $\alpha=0.05$

- For large scale whole exome studies
  - 1250 individuals from each extreme are selected and sequenced
  - $\alpha=2.5\times10^{-6}$

- Hypothesis is carried-out using the original WSS test

Three Estimators

- Naïve estimator $\hat{\beta}_{LAE,naive}$
  - Obtained using the same sample where the association is identified
  - No correction for winner's curse

- BSS-corrected estimator $\hat{\beta}_{LAE,BSS}$
  - Obtained using the same sample where the association identified

- Independent estimator $\hat{\beta}_{LAE,S2}$
  - Obtained using an independent stage 2 sample of equivalent sizes

Results of Simulation Experiment

$\tilde{\beta}$	Percentage of Causative Variants	β_{LAE}	Power	Bias for $\hat{\beta}_{LAE,naive}$	Bias for $\hat{\beta}_{LAE,BSS}$	Bias for $\hat{\beta}_{LAE,S2}$
Small scale Candidate Gene Study						
0.25	0.5	0.182	0.336	0.067	0.022	0.002
0.25	0.8	0.232	0.546	0.049	0.020	0.005
0.5	0.5	0.324	0.566	0.039	0.024	0.000
0.5	0.8	0.450	0.817	0.018	0.014	0.004
Large Scale Whole-exome Study						
0.25	0.5	0.201	0.044	0.055	0.014	0.005
0.25	0.8	0.234	0.188	0.045	0.019	-0.005
0.5	0.5	0.353	0.314	0.019	-0.001	-0.004
0.5	0.8	0.444	0.747	0.011	-0.009	0.000

Results of Simulation Experiment

$\tilde{\beta}$	Selection Coefficient for Causal Variants	β_{LAE}	Power	Bias for $\hat{\beta}_{LAE,naive}$	Bias for $\hat{\beta}_{LAE,BSS}$	Bias for $\hat{\beta}_{LAE,S2}$
Small scale Candidate Gene Study						
0.25	$>10^{-3}$	0.153	0.227	0.106	0.043	0.005
0.5	$>10^{-3}$	0.274	0.481	0.060	0.027	0.002
0.25	$>10^{-4}$	0.207	0.413	0.067	0.024	0.000
0.5	$>10^{-4}$	0.384	0.746	0.025	0.019	0.000
Large Scale Whole-exome Study						
0.25	$>10^{-3}$	0.195	0.031	0.068	0.021	-0.010
0.5	$>10^{-3}$	0.340	0.259	0.023	0.003	0.000
0.25	$>10^{-4}$	0.222	0.117	0.041	0.010	-0.002
0.5	$>10^{-4}$	0.394	0.586	0.021	-0.002	0.000

Conclusions:

- The naïve estimator can be seriously biased
  - If estimation is carried out using the same sample where the association was identified

- BSS algorithm can greatly reduce the bias due to winner's curse
  - Will not completely remove the bias for greatly underpowered studies

- Locus average effect β_{LAE} can be consistently estimated

Analysis of Dallas Heart Study Dataset

- Analyze three different populations separately

- Within each ethnic population
  - Quantile normalize the quantitative trait

- Variants with MAF<3% are analyzed
  - For each trait, samples with trait values in the upper and lower quartiles are used

Results

Associations		P Value	$\hat{\beta}_{LAE,naive}$	$\hat{\sigma}^2_{LAE,naive}$ (×10^{-2})	$\hat{\beta}_{LAE,BSS}$	$\hat{\sigma}^2_{LAE,BSS}$ (×10^{-2})
European Americans						
ANGPTL4	TG	0.017	-0.529	1.068	-0.437	0.703
ANGPTL4	VLDL	0.032	-0.467	0.892	-0.314	0.384
ANGPTL5	TCL	0.008	0.295	0.117	-0.023	0.001
ANGPTL5	LDL	0.01	1.772	1.263	1.065	0.304
African Americans						
ANGPTL3	TG	0.036	-0.237	0.102	-0.118	0.026
ANGPTL3	VLDL	0.023	-0.239	0.103	-0.148	0.040
Hispanic Americans						
ANGPTL6	TG	0.018	0.316	0.410	-0.049	0.008
ANGPTL6	VLDL	0.033	0.250	0.282	-0.195	0.140

Results

Associations		$\hat{\beta}_{LAE,BSS}$	$\hat{\sigma}^2_{LAE,BSS}$ (×10⁻²)	$\hat{\sigma}^2_{\max}$ (×10⁻²)		
				$\tilde{\beta}_{\max} = 0.75$	$\tilde{\beta}_{\max} = 1$	$\tilde{\beta}_{\max} = 1.25$
European Americans						
ANGPTL4	TG	-0.437	0.703	1.283	1.701	2.119
ANGPTL4	VLDL	-0.314	0.384	0.974	1.294	1.613
ANGPTL5	TCL	-0.023	0.001	0.023	0.030	0.038
ANGPTL5	LDL	1.065	0.304	NA	0.285	0.357
African Americans						
ANGPTL3	TG	-0.118	0.026	0.169	0.226	0.282
ANGPTL3	VLDL	-0.148	0.040	0.209	0.279	0.348
Hispanic Americans						
ANGPTL6	TG	-0.049	0.008	0.129	0.172	0.215
ANGPTL6	VLDL	-0.195	0.140	0.566	0.753	0.940

Acknowledgement

- Leal Lab
  - Prof. Suzanne M. Leal
  - Rosa Banuelos
- Drs. Helen Hobbs and Jonathan Cohen for sharing sequence data from *ANGPTL3*, *ANGPTL4*, *ANGPTL5* and *ANGPTL6* genes
- Dr. Shamil Sunyaev for sharing simulated genetic data
- Keck Fellowship in Pharmacogenomics