Hypothesis testing and variable selection for Studying Rare Variants in Sequencing Association Studies

Xihong Lin

Department of Biostatistics
Harvard School of Public Health

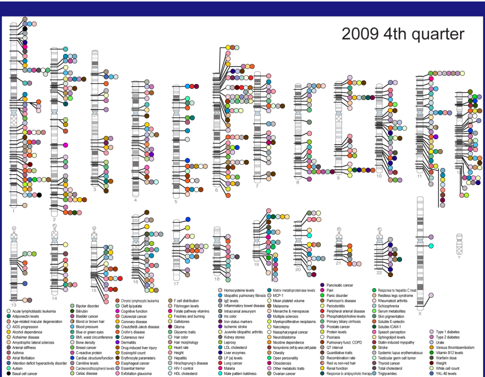
xlin@hsph.harvard.edu

Outline

- Goals and Challenges
- Sequencing Association Tests:
  - Collapsing Methods
  - SKAT
- Selection of Causal Variants
- Simulations studies and Analysis of Dallas Heart Study Data
- Discussions

Genome-Wide Association Studies (GWAS)

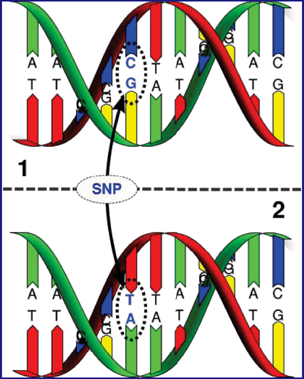
- GWAS have identified > 1200 common genetic variants (SNPs) associated with human diseases.



- Most currently used SNP arrays (Affymetrics and Illumina) genotype 500K-1M SNPs/sample, with an upcoming 5 million SNP array.

Single Nuclide Polymorphism (SNP)

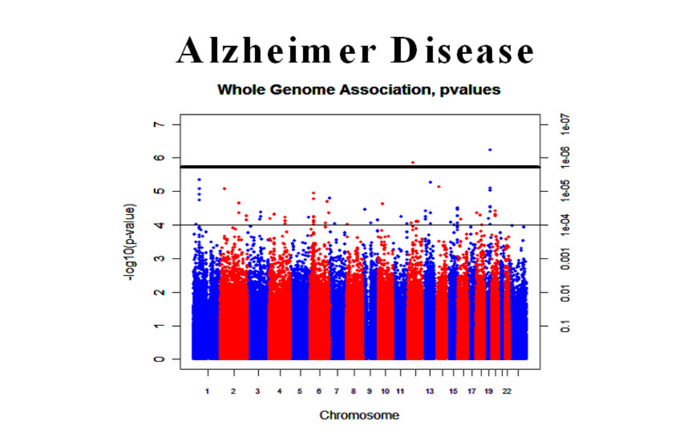
We share 99.9% of our DNA. Small variations (SNPs) at some locations make us different, about 1 in 1000 basepases (bps).



Common Approach in GWAS

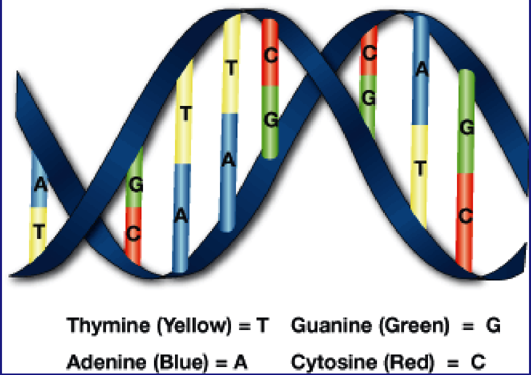
- Discovery phase:
  - Regress outcome (e.g. case/control) on each individual SNP (AA=0, AB=1, BB=2) (Minor Allele Frequency(MAF)=Pr(B)$>$ 0.05).
  - Rank p-values (Manhattan plot).
- Validation phase: Validate the top SNPs in independent samples.

Common Approach in GWAS: Manhattan plot



Sequencing

Genotype all basepairs (bps) in the neighborhood of a gene, the whole exome, or the whole genome (3 billion bps).



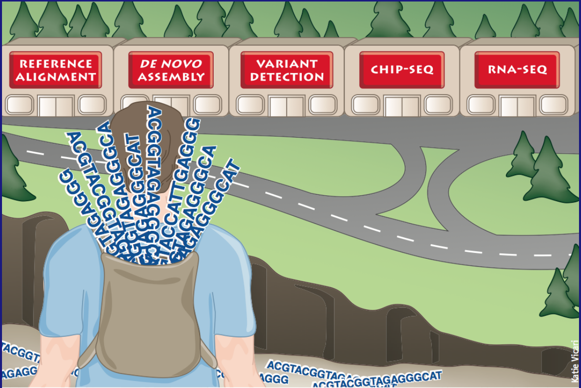
Next-Generation Sequencing Gap

”There is a growing gap between the generation of massively parallel sequencing output and the ability to process and analyze the resulting data.

Bridging this gap is essential, or the coveted $1,000 genome will come with a $20,000 analysis price tag.”

John McPherson, Nature Methods, 2009

Gap Between Sequencing-Generation and Data Analysis Capabilities



McPherson, 2009

Analysis of Next-Generation Sequencing Data

- NGS Platforms: Roche/454; Illumina/ Solexa; ABI SOLiD; Helicos.
- Data storage.
- Low-level analysis: base calling, alignment, assembly, SNP call.
- High-level analysis: (Re)sequencing association studies.

How many subjects are needed to observe a rare variant?

- Sample size required to observe a variant with MAF=p with at least θ chance

$$n > \frac{ln(1-\theta)}{2ln(1-p)}$$

- For $\theta = 99.9\%$, the required minimum sample size is

MAF	0.1	0.01	0.001	0.0001
Minimum n	33	344	3453	34537

(Re)sequencing Association Studies

- Strategy:
  - Identify all observed variants within a sequenced (sub)-region.
  - Region: gene, moving window, intron, exon, ...
  - Test the joint effect of rare/common variants while adjusting for covariates.

Regression Models

- Covariates $\mathbf{X}_i$: age, gender, population stratification.
- Observed rare and common variants in a region: $S_1, \cdots, S_p$
- Model: continuous trait (linear) and binary trait(logistic):

$$\mu_i \text{ or } logit(p_i) = \alpha_0 + \boldsymbol{\alpha}\mathbf{X}_i + \beta_1 S_{i1} + \cdots + \beta_p S_{ip}$$

- Let the data speak about the true unknown β's: some might be 0, - or +.
- "True" non-zero β's are "causal"

Understanding Collapsing Methods

- Suppose only rare variants (with MAF $<$ some threshould) are considered.
- If all β's are the same, the model becomes

$$\mathrm{logit}(p_i) = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X}_i + \beta N_i,$$

where $N_i = S_{i1} + \cdots + S_{ic}$=total number of rare variants in the region.

Understanding Collapsing Methods

- This means the collapsing method assumes (1) all the rare variants are causal and (2) they have the same effect (both in terms of direction and magnitude).
- The collapsing method is optimal if this assumption is true.
- If majority of rare variants have no effects or some are in different directions, the collapsing methods will have substantial power loss.

Sequence Kernel Association Test (SKAT)

Main idea:

- Let the data speak.
- Allow majority of rare variants to have no effects
- Allow variants to have different directions and magnitudes
- Allow for epistatic effects
- Incorporate as much as prior knowledge as possible.
- Avoid thresholding
- ▸ Adjust for covariates

Sequence Kernel Association Test (SKAT)

- Recall logistic model:

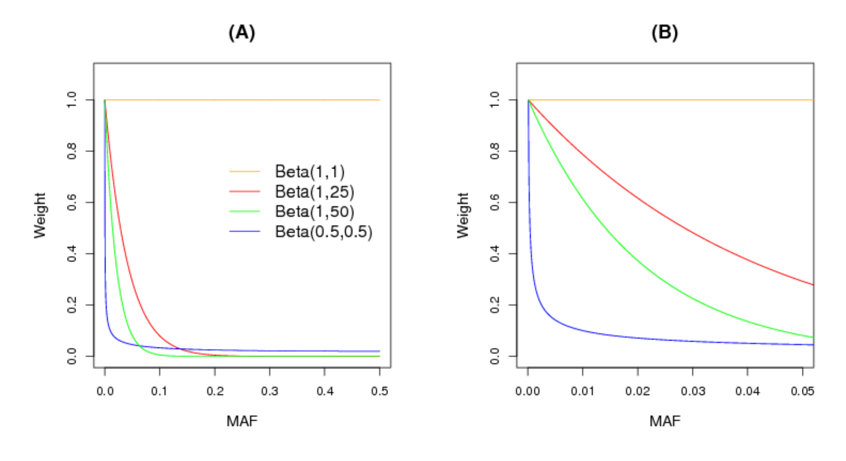
$$logit(\pi_i) = \alpha_0 + \boldsymbol{\alpha}\mathbf{X}_i + \beta_1 S_{i1} + \cdots + \beta_p S_{ip} \quad (1)$$

- No SNP-set (region) effect: $H_0 : \beta_1 = \cdots = \beta_p = 0$
- Standard LR test is a p-df test, little power.
- Assume $\beta_j \sim$ arbitrary distribution $F(0, w_j\tau)$, where w_j is a weight for variant j.
- $H_0 : \beta_1 = \cdots = \beta_p = 0 \Leftrightarrow H_0 : \tau = 0$ (score test for variance component in mixed models)

Choices of Weights in Sequence Kernel Association Test (SKAT)

- Upweight rarer variants.
- Assume weight w_j = decreasing function of MAF π_j
- Example: $w_j = Beta(\pi_j, a_1, a_2)$, where $Beta(\cdot)$=Beta function.
- An optimal choice of w_j is an indicator to indicate whether the j-th marker is a causal variant.

Beta weights



SKAT Statistic (Variance Component Score Test)

- SKAT =weighted sum of individual score statistics,

$$Q = \sum_{j=1}^{p} w_j U_j^2$$

where U_j is the score statistic for SNP j.

- Calculations of Q only requires fitting the null model

$$logit(p_i) = \alpha_0 + \alpha_1 \mathbf{X}_i$$

- P-value of Q can be calculated using a mixture of χ^2 distributions, which is easy to calculate using the Davies' method.

Computational Speed of SKAT

Assume 1000 subjects

Sequence Size	300Kb	3Mb	3Gb (whole genome)
Time	2.5s	25s	7h

on a 2.33 GHz Laptop with 6Gb memory.

General SKAT

- Kernel $K(\mathbf{S}_i, \mathbf{S}_{i'})$ measures genetic similarity in a region between subject i and i' using the p SNPs.
- Examples:
  - Linear kernel=linear effect=Model (1):

    $$K(\mathbf{S}_i, \mathbf{S}_{i'}) = w_1 S_{i1} S_{i'1} + \cdots w_p S_{ip} S_{i'p}$$

    i.e., $\mathbf{K} = \mathbf{SWS}^T$
  - IBS Kernel (SNP-SNP interactions)

    $$K(\mathbf{S}_i, \mathbf{S}_j) = \frac{\sum_{k=1}^{p} w_k IBS(S_{ik}, S_{jk})}{2p}$$

General SKAT

- General logistic model $logit(\mathbf{p}) = \alpha\mathbf{X} + \mathbf{h}$, where $\mathbf{h} \sim$ arbitrary $F(0, \tau\mathbf{K})$.
- Example $h(\mathbf{S}) = \beta_1 S_1 + \cdots + \beta_p S_p$.
- Variance component test for the effect of a SNP set:

$$H_0 : h(\mathbf{S}) = 0 \Leftrightarrow H_0 : \tau = 0$$

- SKAT for a genetic region effect ($H_0 : \tau = 0$):

$$Q(\widehat{\beta}_0) = (\mathbf{y} - \widehat{\mathbf{p}}_0)'\mathbf{K}(\mathbf{y} - \widehat{\mathbf{p}}_0)$$

- P-values calculated using a mixture of χ^2 distributions with df often $<< p$. If complete LD, DF of SKAT=1.

Simulate Sequencing Data

- Generate sequencing data using a coalescent population genetic model.
- Most variants are rare: for example, for a 30Kb region:

# variants	MAF
626 true	
159 (25%)	$< 10^{-4}$
441 (71%)	$< 10^{-3}$
511 (88%)	$< 10^{-2}$

Simulation Set-up

- Simulation model for a given region:
  - Continuous Trait:

    $$Y_i = \alpha_0 + \mathbf{X}_i\boldsymbol{\alpha} + S_{i1}^{causal}\beta_1^{causal} + \cdots + S_{ic}^{causal}\beta_c^{causa}$$

    where $\mathbf{X}_i$ are covariates, $S_1^{causal}, \cdots S_c^{causal}$ are the genotypes for c rare causal variants and $\varepsilon_i \sim N(0,1)$
  - Binary trait (case-control):

    $$logit(\mu_i) = \alpha_0 + \mathbf{X}_i\boldsymbol{\alpha} + S_{i1}^{causal}\beta_1^{causal} + \cdots + S_{ic}^{causal}\beta_c^{cau}$$

- Note: Rare variants, including causal variants, are often not observed in finite samples.

Simulation Study: Methods Compared

- SKAT using all the variants **(SKAT)**
- Collapsing method **(C)**:

  binary indicator for any variants w/ MAF $<$3%
- Count/dosing method **(N)**:

  number of variants w/ MAF $<$3%

Size of SKAT for genome-wide type I error $\alpha = 10^{-6}$

Total Sample Size	Continuous Trait	Binary Trait
500	5.9×10^{-7}	1.0×10^{-8}
1000	8.0×10^{-7}	2.3×10^{-7}
2500	8.4×10^{-7}	5.6×10^{-7}
5000	8.8×10^{-7}	7.0×10^{-7}

Power

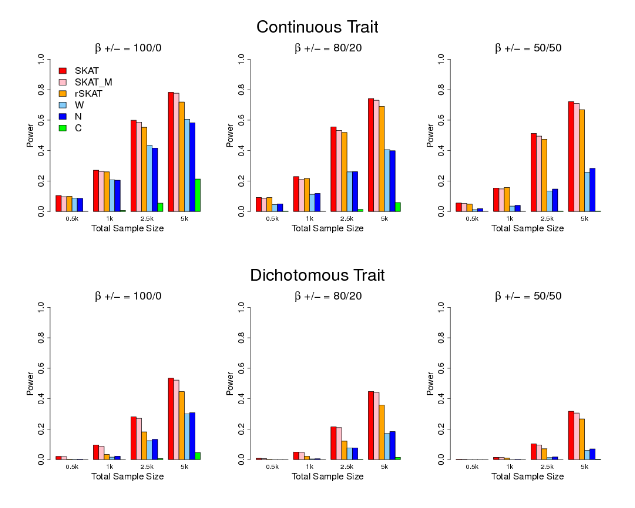
- 5% of variants with $MAF < 3\%$ are causal (15 randomly selected variants)
- In realized samples:

n	250	500	1000	2500	5000
$\bar{p}$	224	262	360	476	552
$\bar{m}$	3.1	4.9	7.1	10.5	12.8

$\bar{\mathbf{p}} =$ Average # of total observed variants ($p_0 = 626$)

$\bar{\mathbf{m}} =$ Average # of observed causal rare variants ($m_0 = 15$)

Power simulations $\alpha = 10^{-6}$ (GW $-$ level) (SKAT vs Collapsing Methods)



SKAT Extension - Correlated β

- Motivation: When βs are positively correlated and most $\beta \neq 0$, collapsing methods can be more powerful than SKAT.
- Goal: Extend SKAT to accommodate this case.
- Idea: Assume the working correlation matrix of β as compound symmetric.

$$\mathbf{R}(\rho) = (1 - \rho)\mathbf{I} + \rho\mathbf{J}\mathbf{J}'$$

- New kernel matrix

$$K_\rho = \mathbf{S}\mathbf{W}^{1/2}\mathbf{R}(\rho)\mathbf{W}^{1/2}\mathbf{S}.$$

- $\rho = 0$: SKAT with linear weighted kernel.
- $\rho = 1$: Weighted count/dosing method (W).

SKAT Extension - Optimal correlation test

- If ρ is known, test statistics

$$Q_\rho = (\mathbf{y} - \widehat{\mathbf{p}}_0)'\mathbf{K}_\rho(\mathbf{y} - \widehat{\mathbf{p}}_0).$$

- Q_ρ follows a mixture of chisq distribution under the null, and p-values can be easily obtained.
- In practice, however, we do not know which ρ maximizes power.
- Test Stat=Smallest p-value from different ρ's

$$T = \inf_{0\leq\rho\leq1} P_\rho,$$

where P_ρ is the p-value of Q_ρ.

SKAT Extension - Optimal correlation test

- Calculate T using a simple grid search.

$$T = min_b P_{\rho_b}, \quad 0 = \rho_1 < \ldots < \rho_B = 1$$

- Null distribution of T uses the fact that Q_ρ is asymptotically the same as

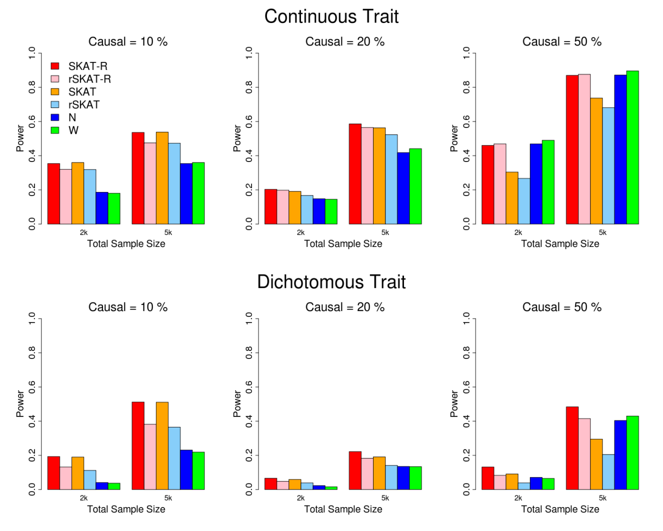
$$(1-\rho)A + \gamma(\rho)\eta, \tag{1}$$

where $\eta \sim \chi_1^2$ and A approximately follows a mixture of chisq, and $Corr(A, \eta) = 0$.

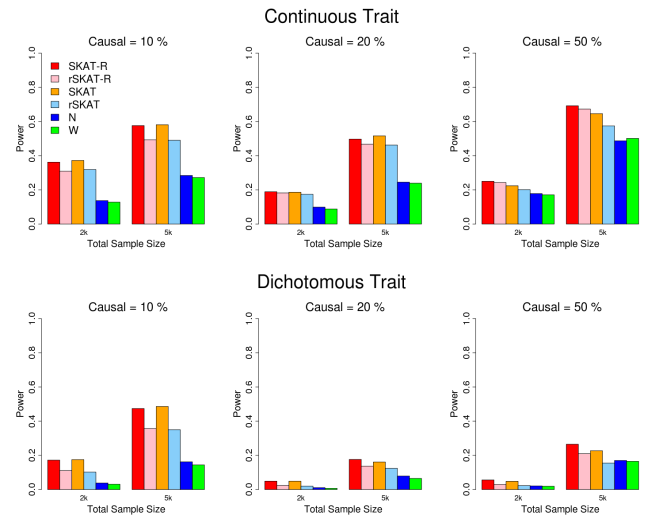
Simulation

- Power simulation on 5kb randomly selected regions.
- Percentages of causal variants = 10%, 20%, or 50%.
- $(\beta_j > 0)$% among causal variants = 100% or 80%.
- SKAT, Collasping (N, W) and the optimal correlation SKAT **(SKAT-R)**.

Power Simulations: All βs are positive, and $\alpha = 10^{-6}$



20% of βs are negative, and $\alpha = 10^{-6}$



Analysis of the Dallas Heart Study Data

- 93 variants in ANGPTL3, ANGPTL4, and ANGPTL5 and 50% are singletons.
- 3476 subjects
- Three ethnicity groups: Black, Hispanic, or White.
- logTG: log of serum triglyceride

Analysis Results of the Dallas Heart Study

	Continuous TG Level	Binary TG Level
SKAT-R	1.8×10^{-5}	1.1×10^{-4}
SKAT	9.5×10^{-5}	1.3×10^{-4}
C	1.9×10^{-3}	3.2×10^{-2}
N	7.2×10^{-5}	2.2×10^{-3}

Selection of Causal Rare Variants

- Problem of Interest: For a top hit region, e.g., a gene , how to select a subset of variants that are likely to be causal and pushed for validation?
- Penalized likelihood has been used to select possible causal variants for common variants, but with limited power for uncommon/rare variants.
- We focus on selecting candidate causal uncommon variants, with *MAF* of 1-5%.
- For very rare variants, e.g. $MAF < 1\%$, very large sample sizes are needed for variable selection.

Weighted Penalized Likelihood for Selecting Causal Rare Variants

- Regression models: continuous trait (linear) and binary trait(logistic):

$$\mu_i \text{ or } logit(p_i) = \alpha_0 + \boldsymbol{\alpha}\mathbf{X}_i + \beta_1 S_{i1} + \cdots + \beta_p S_{ip}$$

- Interested in selecting a subset of S_j that are likely to be associated with D.
- Idea: Incorporate the prior knowledge that rarer variants are more likely to be causal and have a larger effect in variable selection procedures.

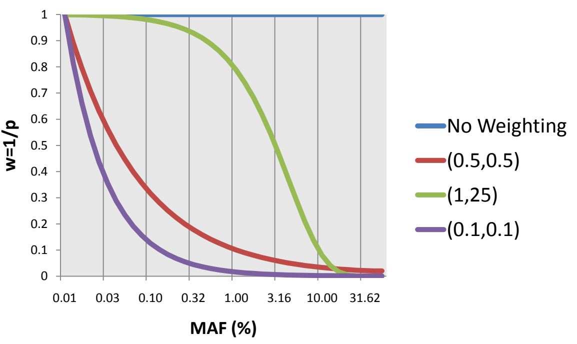
Weighted Penalized Likelihood for Selecting Causal Rare Variants

- Weighted Penalized Likelihood:

$$\sum_{i=1}^{n} \ell(Y_i, \beta) + \lambda \sum_{j=1}^{p} w_j^{-1} |\beta_j|$$

  where $w_j = Beta(MAF_j, a_1, a_2)$.
- Rarer variants have less penalty for β_j and are more likely to be selected.
- This is equivalent to assuming β_j follows a Laplace distribution with variance $(w_j \lambda^{-1})$, parallel to SKAT.

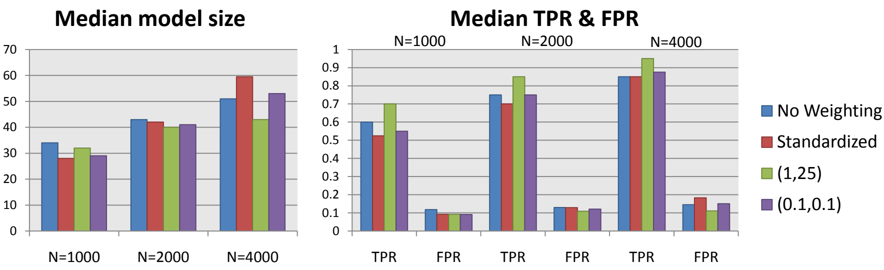
$Beta(MAF; a_1, a_2)$



Simulating Study

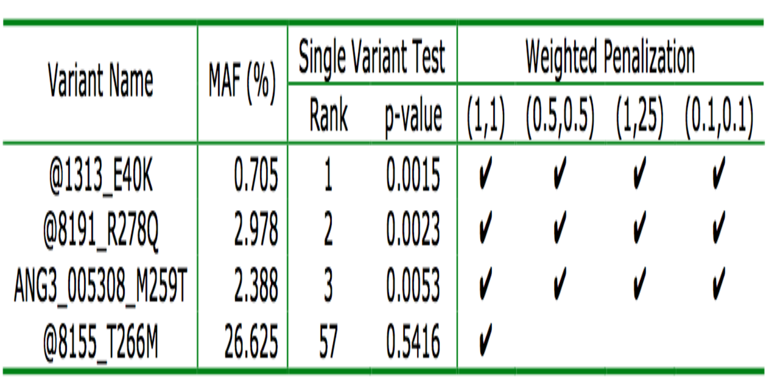
- Simulated sequence data using FREGENE (Chadeau-Hyam *et al*., 2008)
- For each dataset:
  - Considered a 30kb-long region ($\sim$200 observed variants)
  - Simulated 20 causal variants with *MAF* of $1-5\%$
  - Set $|\beta_j| = -\frac{\log 5}{4} \log_{10} MAF$ for causal variants.
- 500 such datasets were simulated for each scenario.

Simulation Results for Binary Traits



- Beta(1,25) gives smaller model size, higher TPR & lower FPR.

Analysis results of the Dallas Heart Study” TG level



Variant Name	MAF (%)	Single Variant Test		Weighted Penalization			
		Rank	p-value	(1,1)	(0.5,0.5)	(1,25)	(0.1,0.1)
@1313_E40K	0.705	1	0.0015	✓	✓	✓	✓
@8191_R278Q	2.978	2	0.0023	✓	✓	✓	✓
ANG3_005308_M259T	2.388	3	0.0053	✓	✓	✓	✓
@8155_T266M	26.625	57	0.5416	✓			

Discussions

- Power and sample size calculations for designing sequencing studies have been derived analytically.
- SKAT provides an attractive approach for sequencing association studies for rare variant effects.
- If the percentage of causal variants is high with the same direction, collapsing methods can have higher power than SKAT.
- The optimal correlation SKAT test (SKAT-R) accounts for correlation among β and outperforms both collapsing methods and SKAT in all cases.
- ▸ Weighted penalized likelihood provides an attractive way to select causal rare variants.

Acknowledgement

- Seunggeun Lee, Harvard (SKAT, SKAT-R)
- Mike Wu, UNC (SKAT)
- Lin Li, Harvard (Causal Variant Selection)
- Tianxi Cai, Harvard (SKAT)
- Yun Li, UNC (SKAT)
- Mike Boehnke, U Michigan (SKAT)

Grants: R37 (NCI MERIT) and PO1
SKAT paper: AJHG, in press.