Gene-level aggregate methods to evaluate genetic associations and gene-by-environment interactions in cross-sectional and longitudinal data

Jennifer Smith

University of Michigan

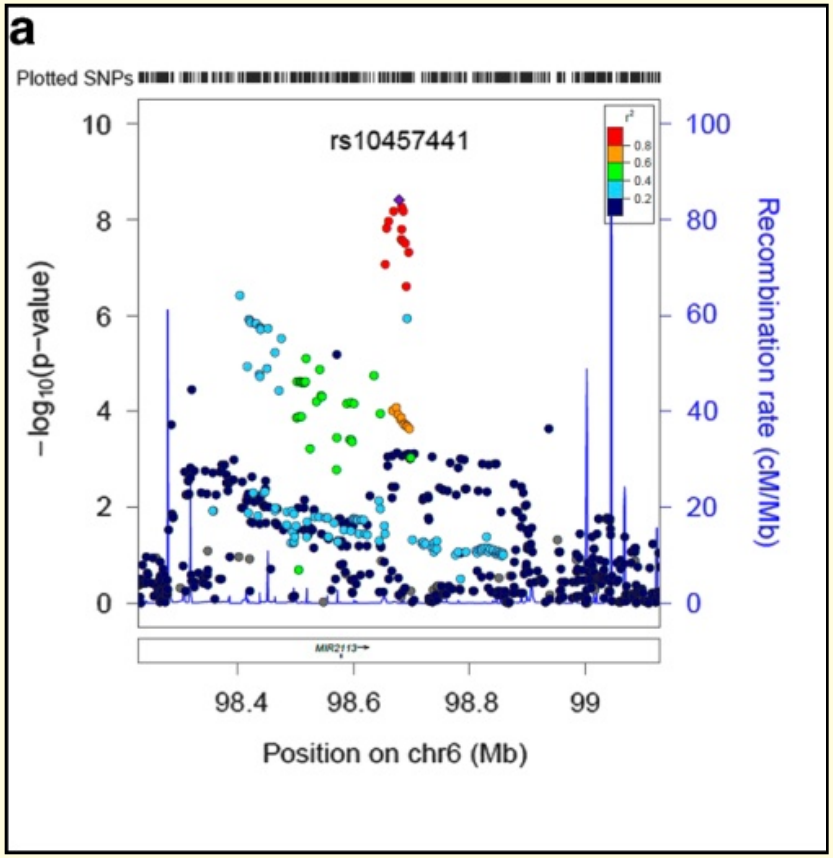
April 26, 2017

The genome is a big place with lots of variation!

Number of variants in HRS imputed data (2006-2012)

Minor Allele Frequency in HRS	Number of Variants
Common (≥ 5%)	28 million
1% to 5%	10 million
Rare (< 1%)	9 million
Total	**47 million**

Cognition GWAS results for *MIR2113 genomic* region



Davies G, et al. (2015) Mol Psych, 20:183-192.

Gene-level aggregate tests

- Test the **aggregate effect** of multiple SNPs/variants within a gene or genomic region on an outcome of interest

- **Burden tests**
  - Collapse variants to test for cumulative effects
  - Most powerful when all variants have **similar magnitude and directions of effect**
  - Can include any number of variants effectively without losing power

- **Non-burden tests**
  - Tests the distribution of the aggregated **association statistics** across variants
  - Optimal performance when variants have **differing magnitudes and directions of effect**
  - Can use both **common and rare variants**
  - Can lose power when variants > sample size
  - Examples: SKAT and LGEWIS

Sequence Kernel Association Test (SKAT)

$$\boldsymbol{Y}_i = \alpha_0 + \boldsymbol{\alpha_1}'\boldsymbol{X}_i + \boldsymbol{\beta}'\boldsymbol{G}_i + \varepsilon_i$$

where:

$\boldsymbol{Y}_i$ = outcome for subject i

α_0 = intercept term

$\boldsymbol{X}_i$ = vector of non-genetic covariates (e.g., age, sex, genetic PCs)

$\boldsymbol{G}_i$ = vector of genotypes for variants,

ε_i = error term; follows a distribution with mean 0 and variance σ^2

- Assume $\boldsymbol{\beta}_j$, j=1,…,p, follows an arbitrary distribution with mean 0 and variance $w_j\tau$
  - w_j = pre-specified weight for each variant
  - τ = variance component
- Testing H_0: $\boldsymbol{\beta}$ = 0 is equivalent testing H_0: τ = 0

Wu MC, et al. (2011) AJHG, 89:82-93

Sequence Kernel Association Test (SKAT)

- H_0: $\tau = 0$
  - No association between the variants and the outcome
  - Assesses whether there is **any variance** in the set of β_js from the mean of 0
  - Can be **in any direction** (+/-)
- Variance component score test
- **Q statistic** is the sum of a mixture of χ^2 distributions based on genotype covariance matrix
- Null hypothesis is evaluated explicitly and used as a reference distribution to compute **a gene-level p-value** using bootstrapping

Wu MC, et al. (2011) AJHG, 89:82-93

SKAT variant weights

- Good choice of variant weights can improve power
- Functional vs. non-functional (e.g., PolyPhen score)
- Common vs. rare
  - Up-weighting rare variants is supported by evolutionary theory
  - Authors suggest weighting by the *Beta* function using minor allele frequency (MAF)

where:

w_j = weight for variant$_j$

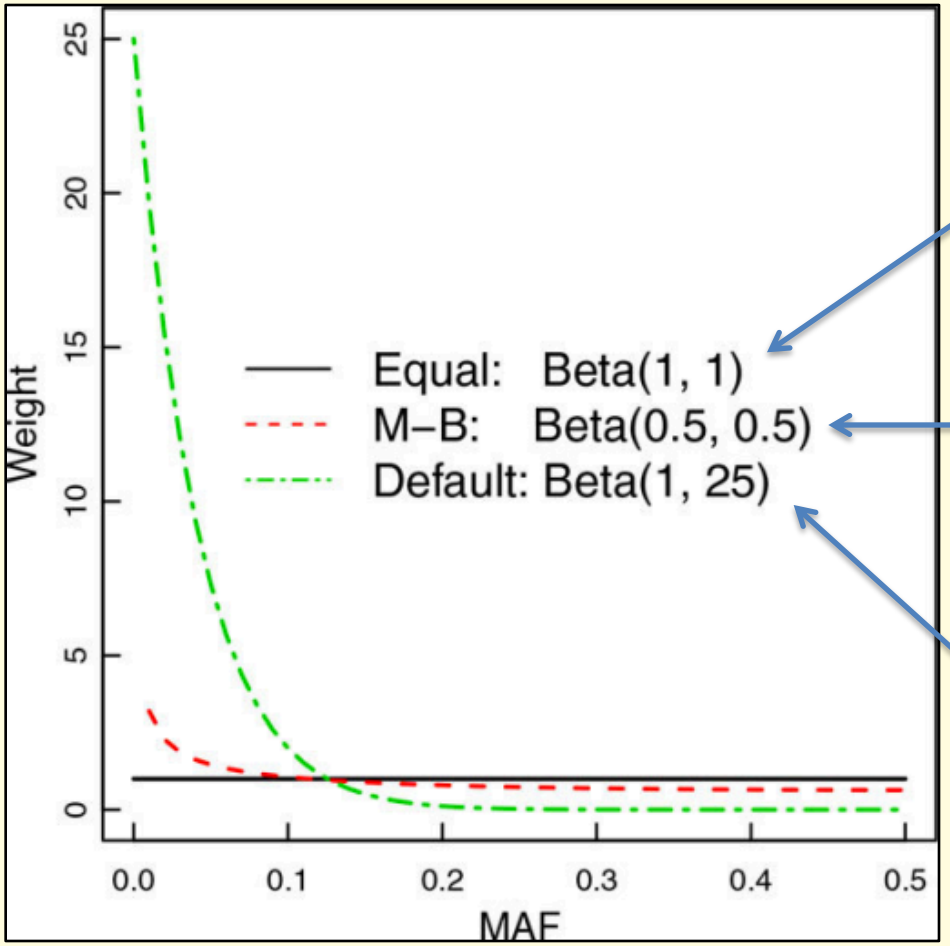
MAF$_j$ = minor allele frequency for variant$_j$

α_1 and α_2 = shape parameters for the beta function

$$\sqrt{w_j} = Beta(MAF_j; \alpha_1, \alpha_2)$$

Wu MC, et al. (2011) AJHG, 89:82-93

SKAT variant weights: Minor allele frequency



- All variants weighted equally
- Low power for rare variants

- Rare variants slightly up-weighted
- Can detect signals for both common and rare variants
- Lower power if both types of variants are not contributing

- Rare variants highly up-weighted
- Low (if any) power for common variants

Mallaney and Sung (2014) BMC Proc.

SKAT kernels

The more general SKAT model:

$$\boldsymbol{Y}_i = \alpha_0 + \boldsymbol{\alpha_1}'\boldsymbol{X}_i + f(\boldsymbol{G}_i) + \varepsilon_i$$

Where $f(\boldsymbol{G}_i)$ can specify different kernels:

- **Linear**
- Identity-by-state (IBS)
- Epistatic (variant*variant) interactions
- 2-way interaction (product kernel consisting of both variant main effects and variant*variant interactions)

Optimal Unified Sequence Kernel Association Test (SKAT-O)

- Optimally combines **non-burden test (SKAT)** with **burden test**
- Test statistic (Q_ρ):

$$Q_\rho = (1 - \rho)Q_S + \rho Q_B$$

where

  - Q_S is the test statistic from SKAT
  - Q_B is the test statistic from a burden test
  - ρ is a parameter calculated to optimize the test statistic

- Output includes the **Q statistic, p-value, and ρ**
- Allows the user to evaluate both types of effects simultaneously, and boosts power by performing the optimal tests
- "SKAT" package in R, includes a simulation function to calculate power

Lee S, et al. (2012) AJHG, 91:224-237

Example: HRS memory performance and decline

Use gene-based methods to assess whether **common** or **rare variants** in **41 genes** previously shown to be associated with cognition and related phenotypes (like AD):

1) Are associated with **episodic memory performance and decline** in non-Hispanic whites (N=9,914) and African Americans (N=2,223) from the Health and Retirement Study

- **Memory score** = total number of words correctly recalled immediately plus 5 minutes later (range: 0-20), assessed at multiple time points
- A **growth curve model** with random effects was used to extract features of memory performance and decline over time
- **Gene regions** included all variants in the gene +/- 5kb

Variant weights matter

- Whites (N=9,914)
- Memory decline (slope) = sex + college education + 4 genetic PCs + performance (age 65 intercept) + variants (1000G imputed data)
- SKAT
  - Beta(1,1): equal weights for all variants
  - Beta(1,25): rare variants up-weighted

Gene Name	Chromosome	Start Position	Gene Size (kb)	Number of Variants	P-value Beta (1,1)	P-value Beta (1,25)
XIRP1	3	39224701	9	101	**0.0228**	0.5635
CELF1	11	47487496	100	532	0.7159	**0.0498**
MS4A6A	11	59939081	13	165	0.1993	**0.0013**
DPH6	15	35509546	329	2132	**0.0197**	**0.0413**
PVRL2	19	45349432	43	401	**0.0043**	**0.0003**
TOMM40	19	45393826	13	116	**1.52E-08**	**1.55E-10**
APOE	19	45409011	4	30	**2.38E-11**	**2.09E-09**
APOC1	19	45417504	5	88	**1.64E-14**	**2.43E-11**

Results for genes with SKAT p<0.05 for at least one weighting scheme are shown

Inclusion of burden test matters

- Same analysis

- SKAT/SKAT-O
  - Beta(1,1): equal weights for all variants
  - Beta(1,25): rare variants up-weighted
  - ρ = 0 if SKAT, 1 if burden score

Gene Name	Chr	Gene Size (kb)	Number of Variants	Beta (1,1)			Beta (1,25)		
				SKAT	SKAT-O		SKAT	SKAT-O	
				P-value	P-value	ρ	P-value	P-value	ρ
XIRP1	3	9	101	**0.02**	**0.04**	0	0.56	**0.05**	1
CELF1	11	100	532	0.72	0.22	1	**0.05**	**0.03**	1
MS4A6A	11	13	165	0.20	0.23	0	**0.001**	**0.001**	0.04
DPH6	15	329	2132	**0.02**	**0.02**	0.01	**0.04**	0.07	0
PVRL2	19	43	401	**0.004**	**$1x10^{-06}$**	1	**$3x10^{-04}$**	**$5x10^{-04}$**	0
TOMM40	19	13	116	**$2x10^{-08}$**	**$1x10^{-08}$**	0	**$2x10^{-10}$**	**$2x10^{-10}$**	0
APOE	19	4	30	**$2x10^{-11}$**	**$2x10^{-10}$**	0	**$2x10^{-09}$**	**$1x10^{-08}$**	0
APOC1	19	5	88	**$2x10^{-14}$**	**$4x10^{-16}$**	0.25	**$2x10^{-11}$**	**$2x10^{-10}$**	0.01

Results for genes with SKAT p<0.05 for at least one weighting scheme are shown

iSKAT and iSKAT-O for gene-environment interactions

$$\boldsymbol{Y}_i = \alpha_0 + \boldsymbol{\alpha_1}'\boldsymbol{X}_i + \boldsymbol{\alpha_2}'\boldsymbol{G}_i + \alpha_3 E_i + \boldsymbol{\beta}'\boldsymbol{S}_i + \varepsilon_i$$

where:

$\boldsymbol{S}_i$ = vector of GxE interaction terms

- Same idea as SKAT-O, except that the **distribution of the betas from the GxE interaction terms** are evaluated (H_0: $\boldsymbol{\beta} = 0$)
- Null hypothesis is that the interaction terms are all equal to zero
- Caveat: No gene-level joint tests have yet been developed to examine the variant main effects and GxE interaction effects simultaneously

Example: HRS memory performance and decline

Use gene-based methods to assess whether **common** or **rare variants** in **41 genes** known to be associated with cognition and related phenotypes:

1) Are associated with **episodic memory performance and decline** in non-Hispanic whites (N=9,914) and African Americans (N=2,223) from the Health and Retirement Study, and

2) Evaluate whether **interactions** between these genes and **educational attainment** are also associated with memory performance and decline

iSKAT/iSKAT-O Results

SKAT/SKAT-O (marginal variant effects)

Memory decline (slope) = sex + college education + 4 genetic PCs + age 65 intercept + **variants**

Gene Name	Chr	Gene Size (kb)	Number of Variants	Beta (1,1)			Beta (1,25)		
				SKAT	SKAT-O		SKAT	SKAT-O	
				P-value	P-value	ρ	P-value	P-value	ρ
MS4A6A	11	13	165	0.20	0.23	0	**0.001**	**0.001**	0.04

iSKAT/iSKAT-O (GxE interaction effects)

Memory decline (slope) = sex + college education + 4 genetic PCs + age 65 intercept + variants + **variants*college education**

Gene Name	Chr	Gene Size (kb)	Number of Variants	Beta (1,1)			Beta (1,25)		
				iSKAT	iSKAT-O		iSKAT	iSKAT-O	
				P-value	P-value	ρ	P-value	P-value	ρ
MS4A6A	11	13	165	0.22	0.15	1	**0.03**	**0.007**	1

Longitudinal gene-environment-wide interaction studies (LGEWIS) for repeated measures data

- New methods developed that can incorporate longitudinal measures
  - **Marginal variant effects** (optimized test)
  - **Gene-environment interactions**
- Use generalized estimating equations (GEE) instead of variance components
- Models the association with the outcome, not the longitudinal trajectory (change over time)
- Advantages
  - Increased power over using baseline or average measures
  - Allows for non-linear environmental effects
  - Robust when number variants >> sample size
  - Robust to model misspecification

He Z, et al., JASA (accepted); He, et al. (submitted)

Conclusions

- Advantages of gene-level aggregate methods
  - May be better suited for **multi-ethnic studies** than single variant analysis
  - **Increased power** due to limited number of tests, especially for rare variants
  - Allows **deeper interrogation** of the variation in genes identified through GWAS
  - Provide a platform for **integrating social science with genetics**
- Non-burden tests
  - **Less assumptions** than burden scores
  - **Flexible** in specifying genetic models
  - **Evolving** to incorporate more complicated data types

Thank you!

HRS Collaborators
Jessica Faul
Colter Mitchell
Wei Zhao
Erin Ware
Sharon Kardia
David Weir
MJ Kho
Jake Hillman

Methods Collaborators
Bhramar Mukherjee
Seunggeun Lee
Zihuai He
Yi-An Ko

Funding Sources
R03 AG048806, R03 046389
U01 AG009740, RC2 AG036495, RC4 AG039029

NIH National Institute on Aging
Turning Discovery Into Health