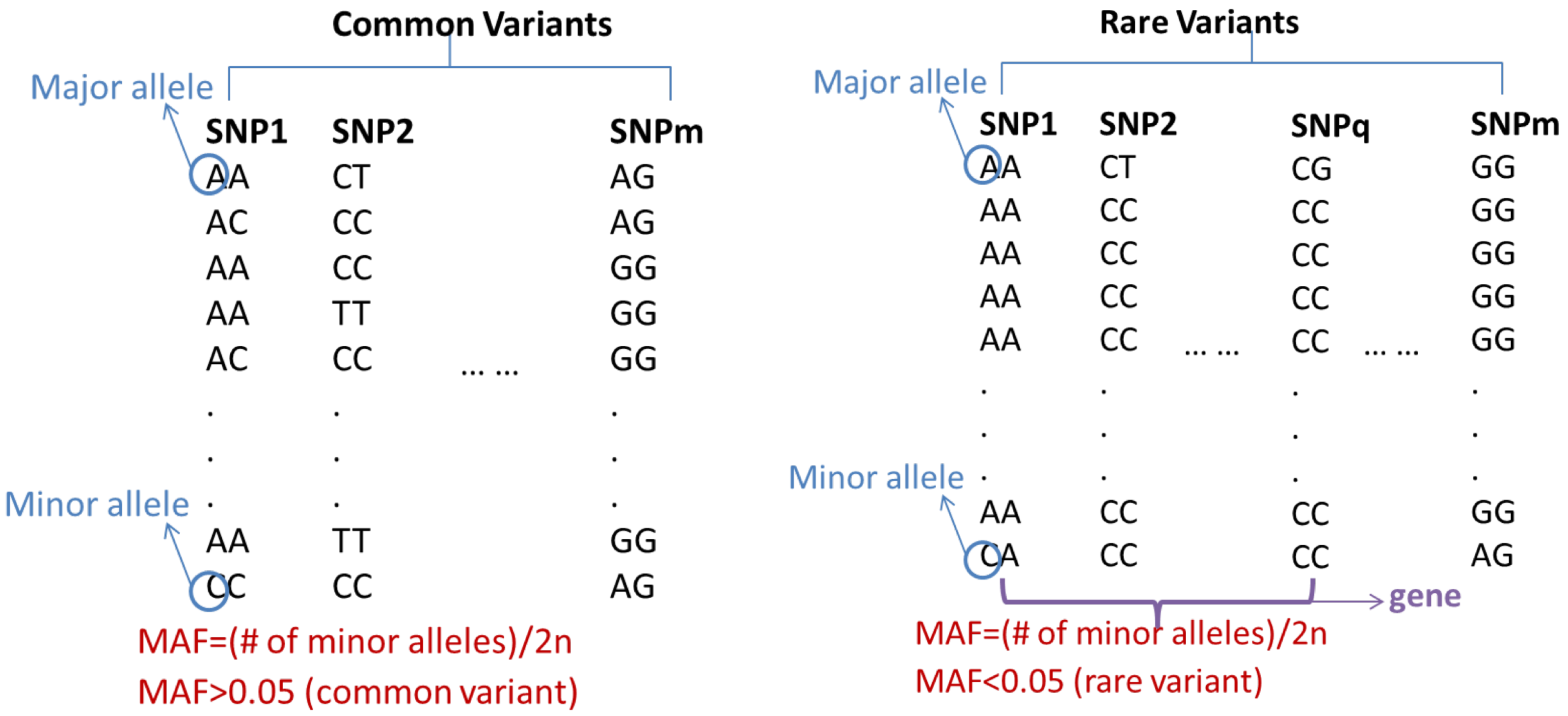
Genetic Rare-Variant Test

Qi Yan
Research Assistant Professor
Division of Pulmonary Medicine
Department of Pediatrics
Children's Hospital of Pittsburgh of UPMC
University of Pittsburgh

Common VS Rare

- **Genotypes:**
  - Common variants (e.g. MAF≥0.05): single marker test;
  - Rare variants (e.g. MAF<0.05): test at gene level (e.g. SKAT).



Common Variants

SNP1	SNP2		SNPm
AA	CT		AG
AC	CC		AG
AA	CC		GG
AA	TT		GG
AC	CC	... ...	GG
.	.		.
.	.		.
.	.		.
AA	TT		GG
CC	CC		AG

MAF=(# of minor alleles)/2n

MAF>0.05 (common variant)

Rare Variants

SNP1	SNP2		SNPq		SNPm
AA	CT		CG		GG
AA	CC		CC		GG
AA	CC		CC		GG
AA	CC		CC		GG
AA	CC	... ...	CC	... ...	GG
.	.		.		.
.	.		.		.
.	.		.		.
AA	CC		CC		GG
CA	CC		CC		AG

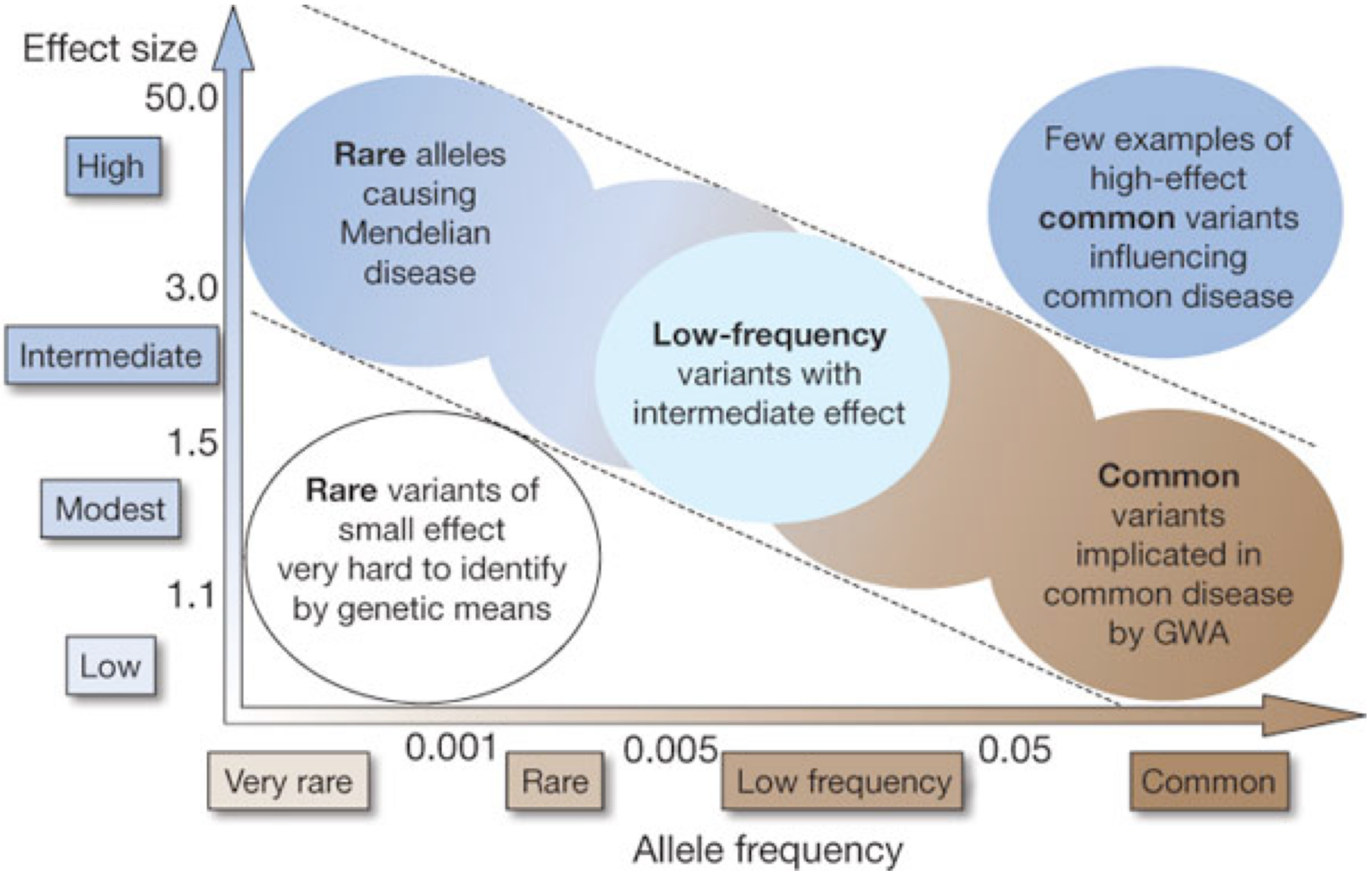
MAF=(# of minor alleles)/2n

MAF<0.05 (rare variant)

- **Only subset of functional elements include common variants**
- **Rare variants are more numerous and thus will point to additional loci**

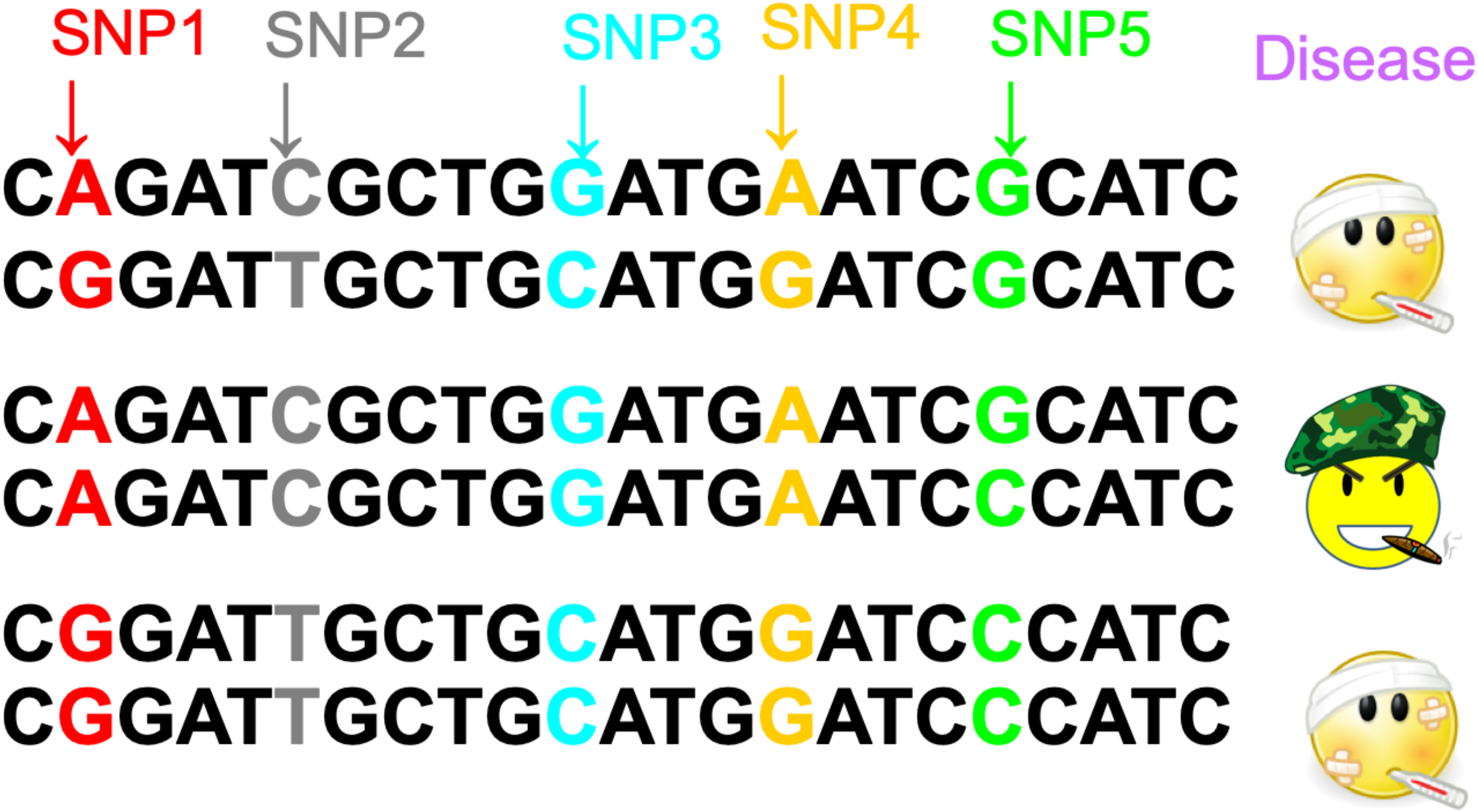
Common VS Rare

Genetic Spectrum of Complex Diseases



Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

GWAS



SNP1 SNP2 SNP3 SNP4 SNP5 Disease

CAGATCGCTGGATGAATCGCATC
CGGATTGCTGCATGGATCGCATC

CAGATCGCTGGATGAATCGCATC
CAGATCGCTGGATGAATCCCATC

CGGATTGCTGCATGGATCCCATC
CGGATTGCTGCATGGATCCCATC

Single Marker Test for Rare Variant

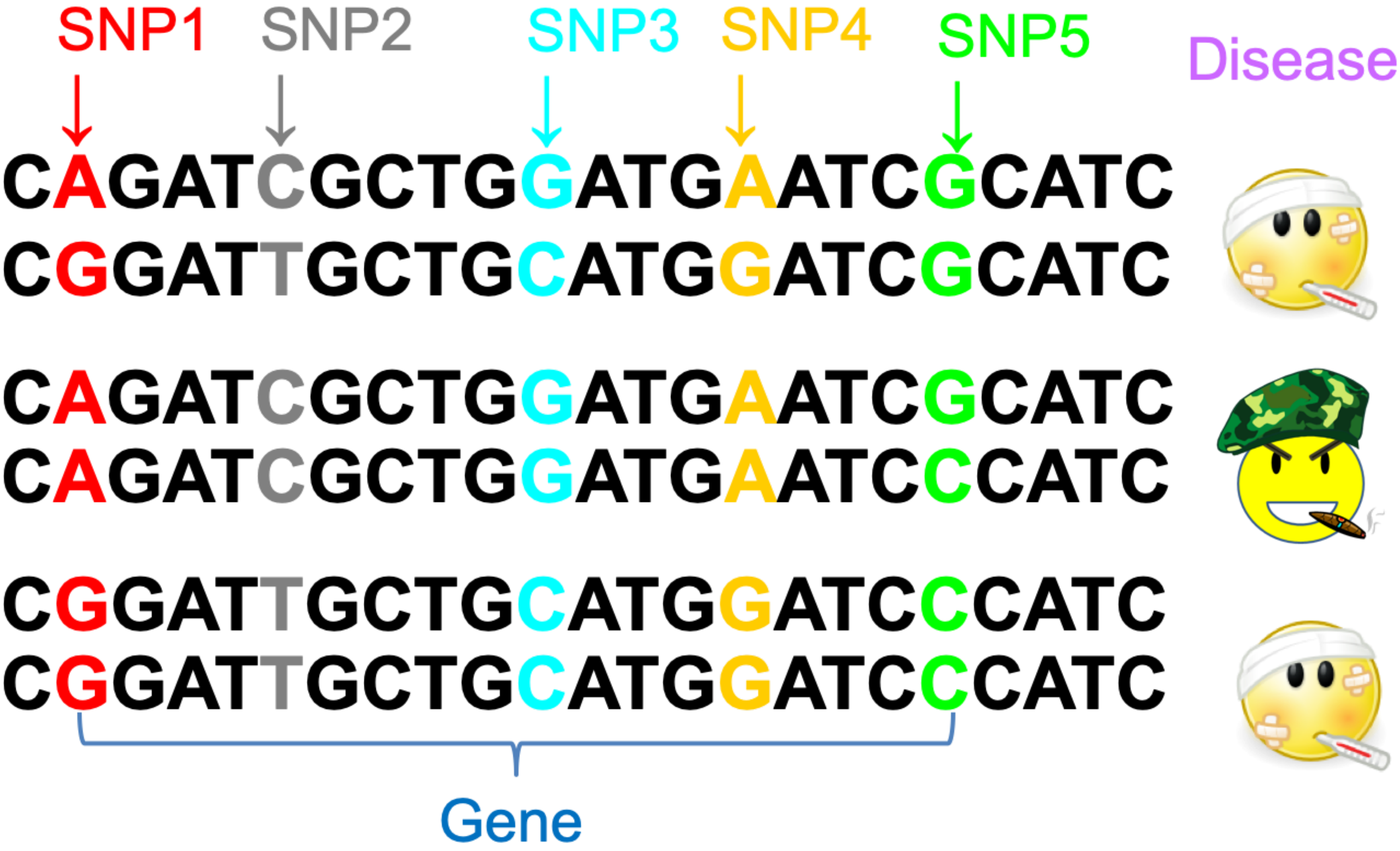
- Rare variants are hard to detect
- Rare variants have low frequency that makes single marker test less powerful
- Rare causal SNPs are hard to identify even with large effect size

Single Marker Test for Rare Variant

- Disease prevalence ~10%
- Type I error 5x10^{-6}
- To achieve 80% power
- Equal number of cases and controls

- Minor Allele Frequency (MAF) = 0.1, 0.01, 0.001
- Required sample size = 486, 3545, 34322,

Alternate Tests for Rare Variant



SNP1 SNP2 SNP3 SNP4 SNP5 Disease

CAGATCGCTGGATGAATCGCATC
CGGATTGCTGCATGGATCGCATC

CAGATCGCTGGATGAATCGCATC
CAGATCGCTGGATGAATCCCATC

CGGATTGCTGCATGGATCCCATC
CGGATTGCTGCATGGATCCCATC

Gene

Alternate Tests for Rare Variant

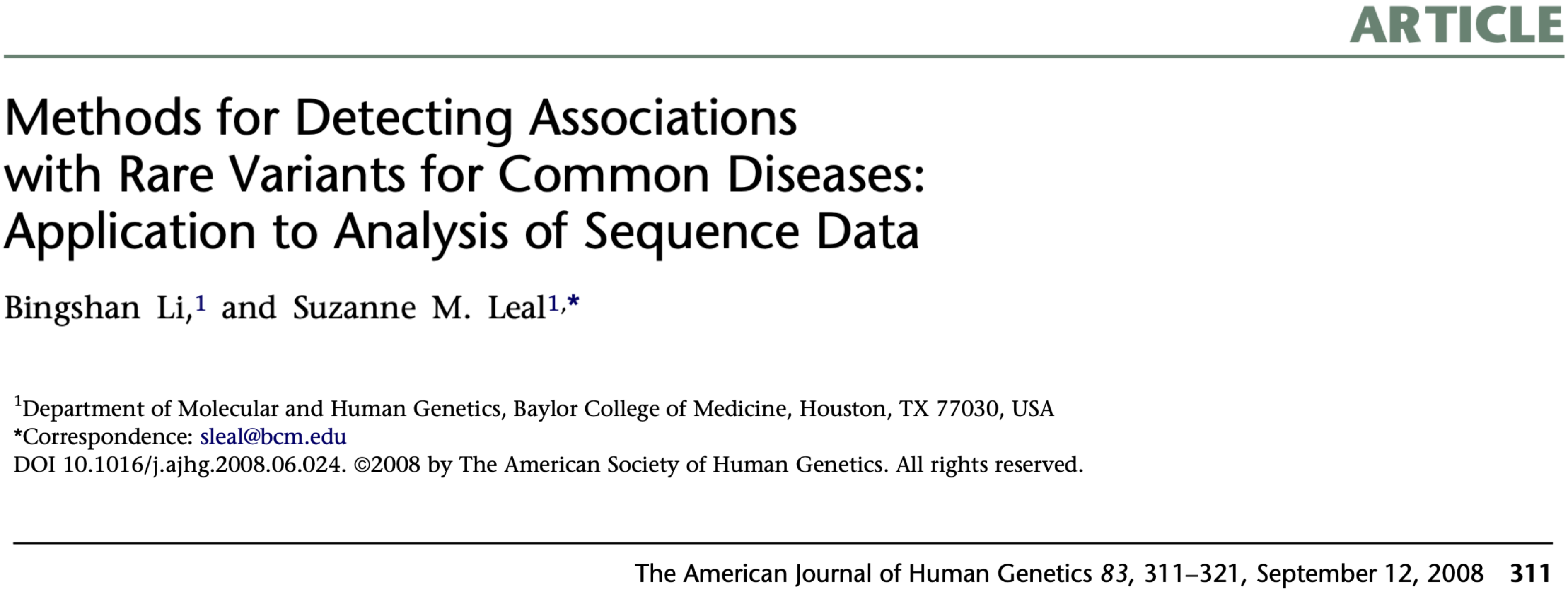
- **Burden Test**
- **Sequence Kernel Association Test (SKAT)**
- **Function Linear Model (FLM)**

- Gene-based tests
- How to handle potential high dimension of rare variants in a gene

Alternate Tests for Rare Variant

- **Burden Test**
- **Sequence Kernel Association Test (SKAT)**
- **Function Linear Model (FLM)**

Burden Test



ARTICLE

Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data

Bingshan Li,[1] and Suzanne M. Leal[1,*]

[1] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

*Correspondence: sleal@bcm.edu

DOI 10.1016/j.ajhg.2008.06.024. ©2008 by The American Society of Human Genetics. All rights reserved.

The American Journal of Human Genetics *83*, 311–321, September 12, 2008 **311**

Burden Test

Alternatives to Single Marker Test Collapsing Method (Burden Test)

- Group rare variants in the same gene/region

- Score each individual
  - Presence or absence of rare copy
  - Weight each variant

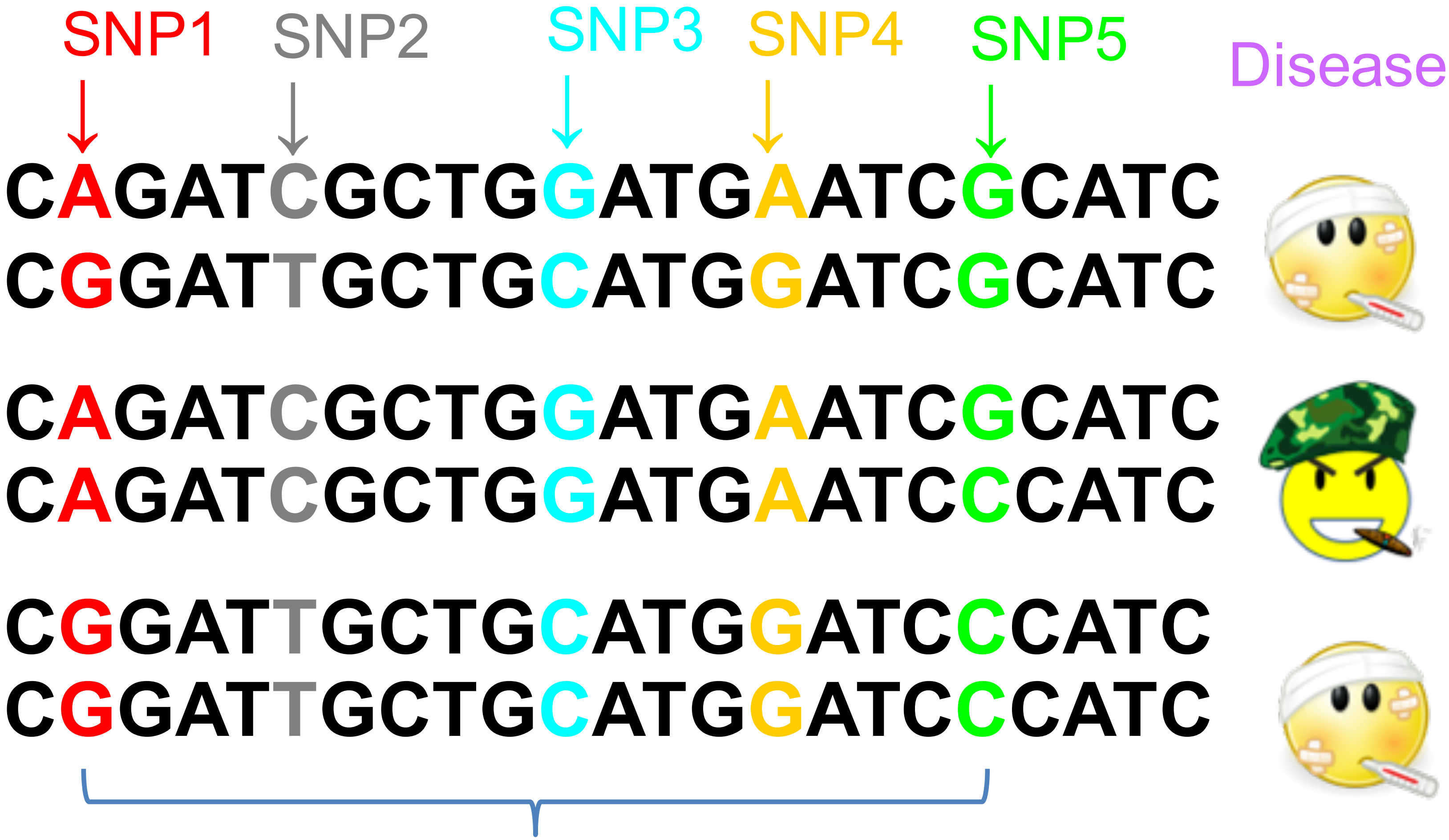
$$X_j = \begin{cases} 1 & \text{rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

- Use individual score as a new “genotype”

- Test in a regression framework

Li and Leal (2008) *Am J Hum Genet* **83:**311-321

Burden Test



SNP1 SNP2 SNP3 SNP4 SNP5

Disease

CAGATCGCTGGATGAATCGCATC
CGGATTGCTGCATGGATCGCATC

CAGATCGCTGGATGAATCGCATC
CAGATCGCTGGATGAATCCCATC

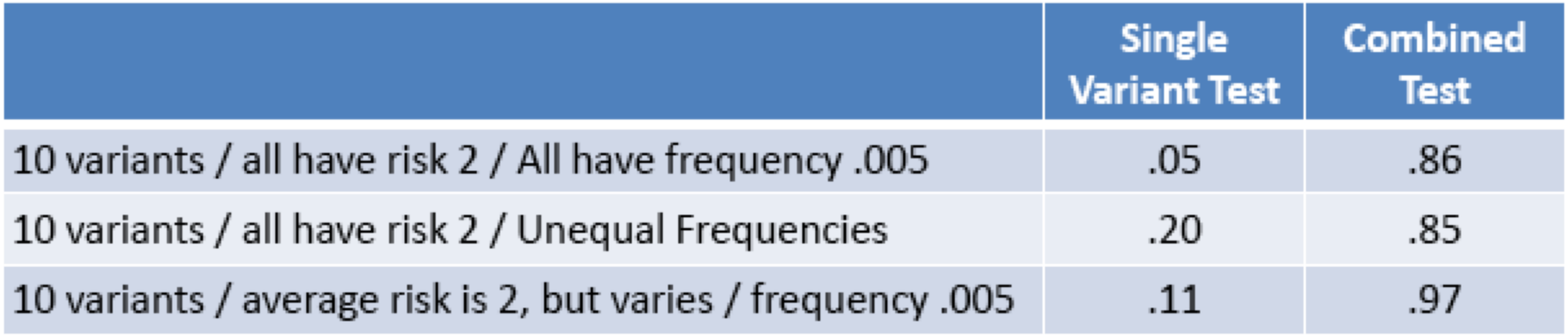
CGGATTGCTGCATGGATCCCATC
CGGATTGCTGCATGGATCCCATC

New "Genotype" = SNP1 + SNP2 + … + SNP5

New "Genotype" = W1*SNP1 + W2*SNP2 + … + W5*SNP5

Burden Test

Power of Burden Test

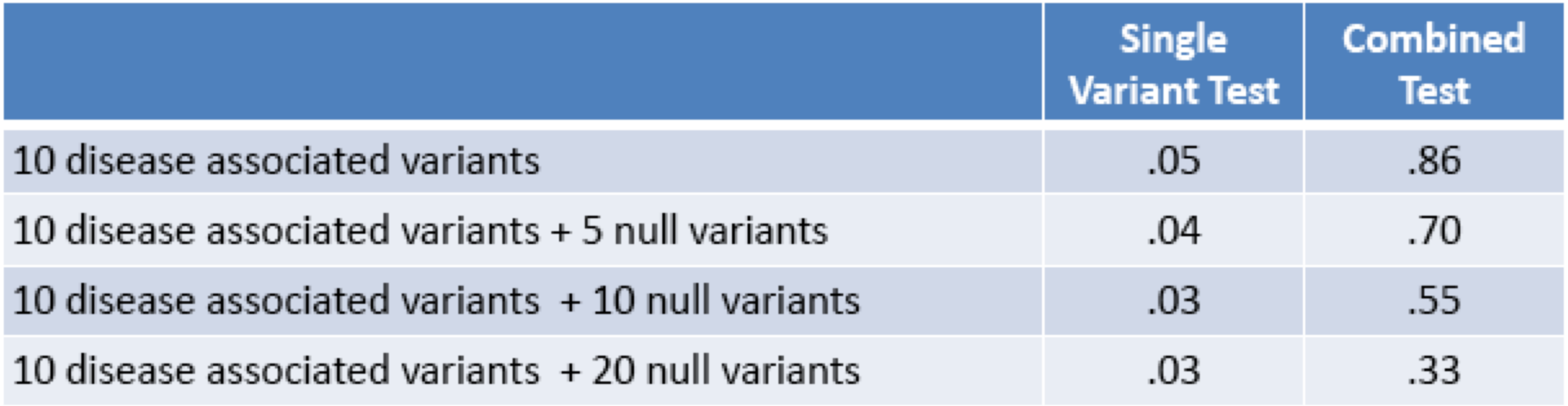


	Single Variant Test	Combined Test
10 variants / all have risk 2 / All have frequency .005	.05	.86
10 variants / all have risk 2 / Unequal Frequencies	.20	.85
10 variants / average risk is 2, but varies / frequency .005	.11	.97

- Power tabulated in collections of simulated data
- Combining variants can greatly increase power
- Appropriately combining variants is expected to be key feature of rare variant studies.

Burden Test

Impact of Null Variants

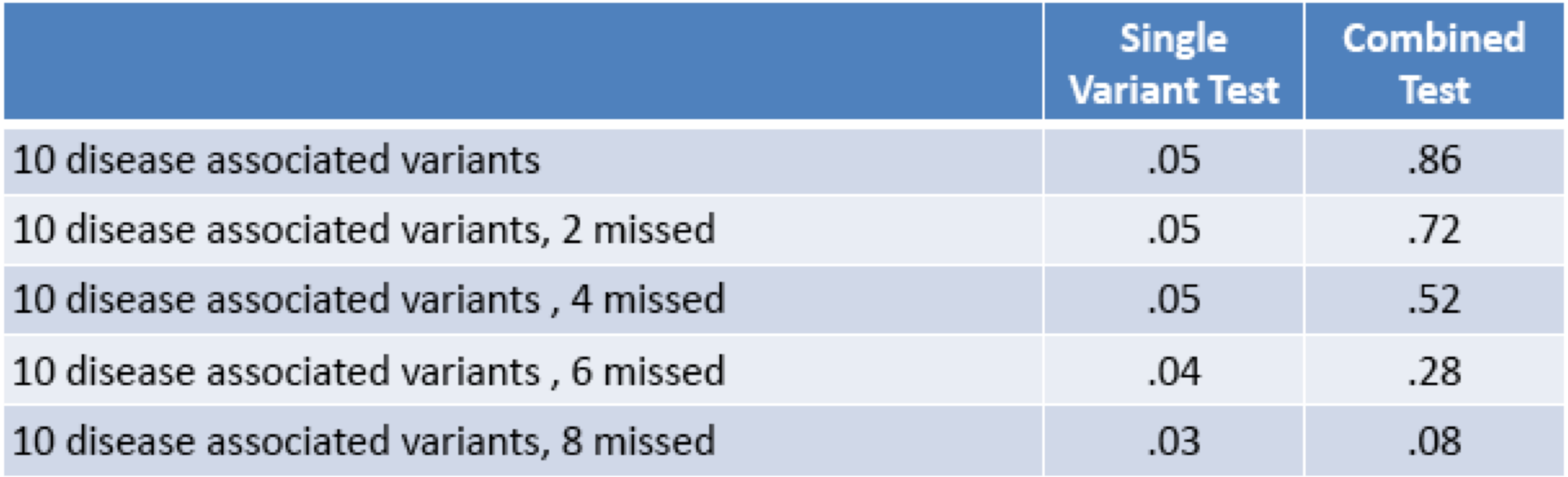


	Single Variant Test	Combined Test
10 disease associated variants	.05	.86
10 disease associated variants + 5 null variants	.04	.70
10 disease associated variants + 10 null variants	.03	.55
10 disease associated variants + 20 null variants	.03	.33

- Including non-disease variants reduces power
- Power loss is manageable, combined test remains preferable to single marker tests

Burden Test

Impact of Missing Disease Alleles



	Single Variant Test	Combined Test
10 disease associated variants	.05	.86
10 disease associated variants, 2 missed	.05	.72
10 disease associated variants , 4 missed	.05	.52
10 disease associated variants , 6 missed	.04	.28
10 disease associated variants, 8 missed	.03	.08

- Missing disease alleles reduces power

- Still better than single marker test

Burden Test

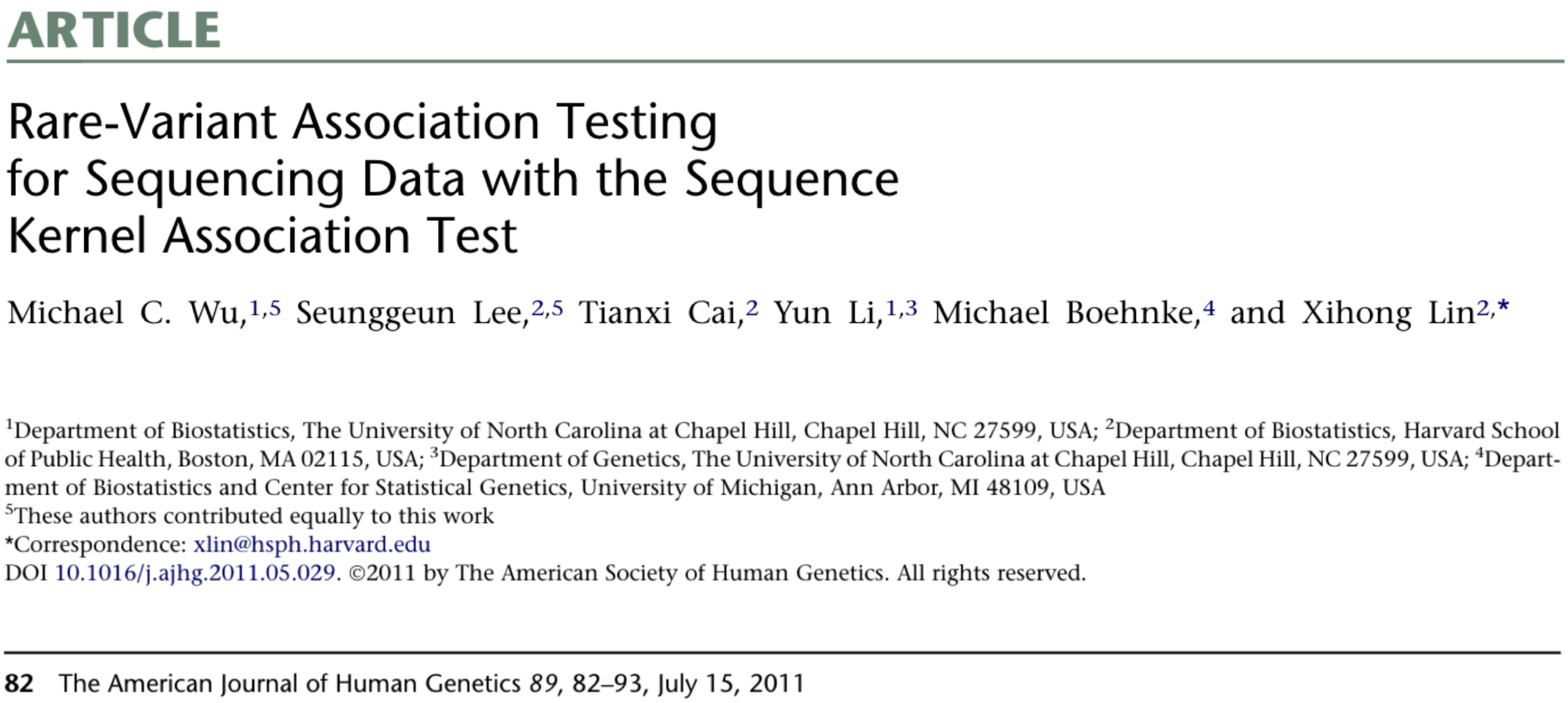
Challenges

- Assume all causal rare variants have the same effect direction
- It is hard to separate causal and null SNPs
  - Including all rare variants will dilute the true signals
- Assume the effect size of each rare variant the same

Alternate Tests for Rare Variant

- **Burden Test**
- **Sequence Kernel Association Test (SKAT)**
- **Function Linear Model (FLM)**

Sequence Kernel Association Test (SKAT):



ARTICLE

Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test

Michael C. Wu,[1,5] Seunggeun Lee,[2,5] Tianxi Cai,[2] Yun Li,[1,3] Michael Boehnke,[4] and Xihong Lin[2,*]

[1]Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [2]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA; [3]Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [4]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

[5]These authors contributed equally to this work

*Correspondence: xlin@hsph.harvard.edu

DOI 10.1016/j.ajhg.2011.05.029. ©2011 by The American Society of Human Genetics. All rights reserved.

Sequence Kernel Association Test (SKAT):

Let there be n subjects with q genetic variants. The n × 1 vector of the quantitative trait $\boldsymbol{y}$:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{G\gamma} + \boldsymbol{\varepsilon}$$

- $\mathbf{X}$ is an $n \times p$ covariate matrix,
- $\boldsymbol{\beta}$ is a $p \times 1$ vector containing parameters for the fixed effects (an intercept and $p - 1$ covariates),
- $\mathbf{G}$ is an $n \times q$ genotype matrix for the q rare genetic variants of interest,
- $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the random effects of the q genetic variants,
- $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector for the random error.

$$\boldsymbol{\gamma} \sim N(0, \tau\mathbf{W})$$

$$\boldsymbol{\varepsilon} \sim N\left(0, \sigma_{\mathrm{E}}^{2}\mathbf{I}\right)$$

where $\mathbf{W}$ is a predefined $q \times q$ diagonal weight matrix for each variant

Thus, the null hypothesis H_0: $\boldsymbol{\gamma} = 0$ is equivalent to H_0: $\tau = 0$, which can be tested with a variance component score test in the mixed model.

Sequence Kernel Association Test (SKAT):

Q: *What makes mixed model different from linear regression model?*
A: *random variables in addition to random error.*

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{G\gamma} + \boldsymbol{\varepsilon}$$ *"linear mixed model"*

$$\text{Var}(\mathbf{y}) = \tau\mathbf{GWG}' + \sigma_E^2\mathbf{I}$$

SKAT test statistic following a mixture of Chi-square distribution is:

$$\text{Q} = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\hat{\boldsymbol{\Sigma}}^{-1}\underbrace{\mathbf{GWG}'}\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

where the parameters are estimated under H_0 (i.e., H_0: $\tau = 0$)

- Called "**kernel**".
- Linear combination used here. Could be more flexible form.

Thus, under H_0: $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ *"linear regression model, no longer mixed model"*

$$\hat{\boldsymbol{\Sigma}} = \hat{\sigma}_E^2\mathbf{I}$$
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

- *The "full model" of SKAT is a linear mixed model*
- *The "null model" for the score test is a linear model*

Sequence Kernel Association Test (SKAT):

Under null hypothesis, the variance of residual is

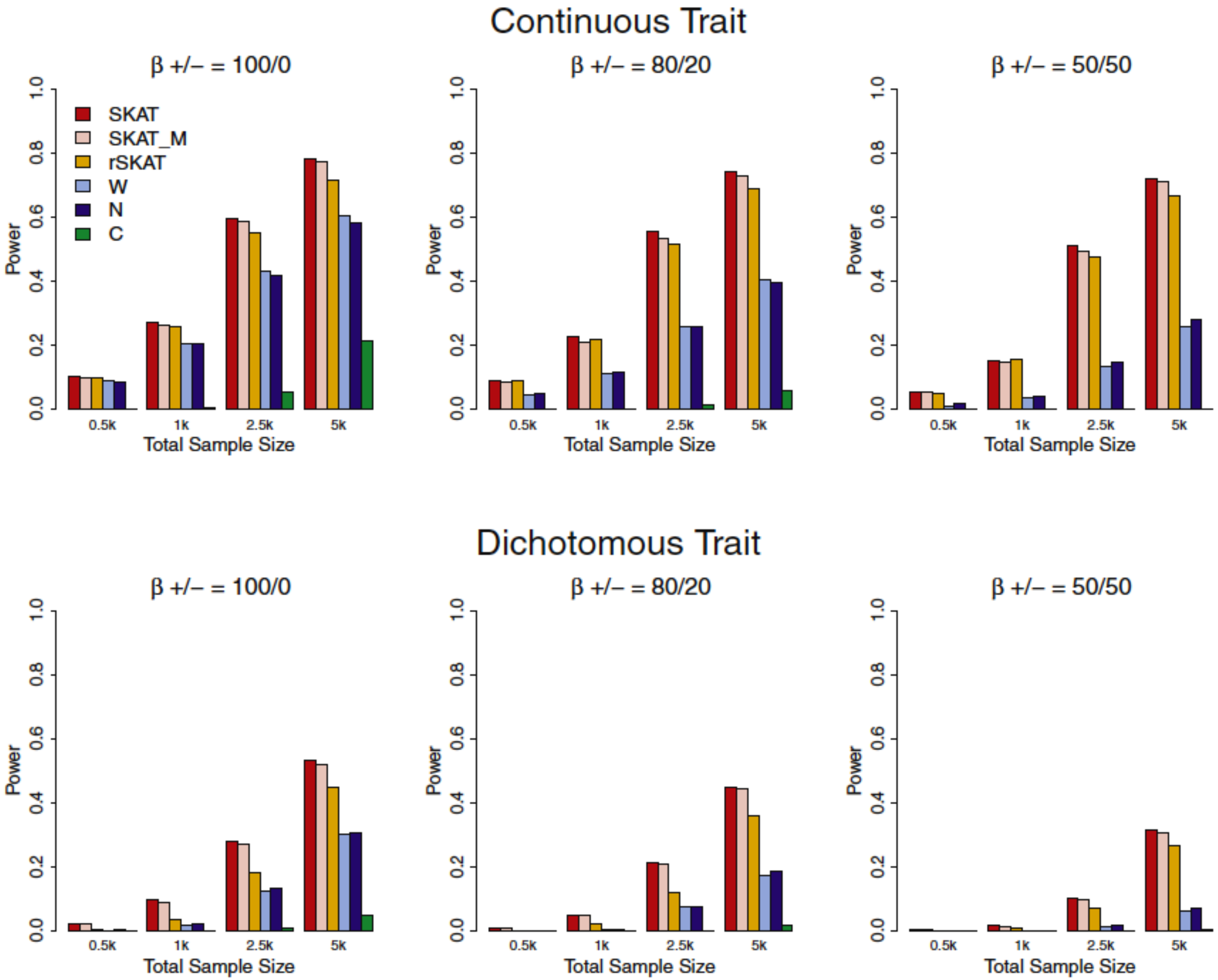
$$\mathrm{var}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)=\hat{\sigma}_E^2-\hat{\sigma}_E^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'=\mathbf{P}_0.$$

The statistic Q $=\hat{\sigma}_E^{-4}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\mathbf{GWG}'\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ is a quadratic form of $\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ and follows a mixture of chi-square distributions under H_0. Thus,

$$\mathrm{Q}\sim\sum_{i=1}^{q}\lambda_i\chi_{1,i}^2$$

where λ_i is the eigenvalues of the matrix $\hat{\sigma}_E^{-4}\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\mathbf{P_0}\mathbf{GW}^{\frac{1}{2}}$.

Sequence Kernel Association Test (SKAT):



Sequence Kernel Association Test (SKAT):

➢ Kernel Machine (KM) Regression for Linear Mixed Model:

With additional random effects (besides the genetic effects):

Let there be n subjects with q genetic variants. The $n \times 1$ vector of the quantitative trait $\boldsymbol{y}$ follows a linear mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\varepsilon}$$

- $\mathbf{X}$ is an $n \times p$ covariate matrix,
- $\boldsymbol{\beta}$ is a $p \times 1$ vector containing parameters for the fixed effects (an intercept and $p - 1$ covariates),
- $\mathbf{G}$ is an $n \times q$ genotype matrix for the q genetic variants of interest,
- $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the random effects of the q genetic variants,
- $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector for the random error,
- $\boldsymbol{u}$ is an $n \times 1$ vector for the random effects due to covariates (e.g., relatedness in families, multivariate traits or time for longitudinal data)

Sequence Kernel Association Test (SKAT):

➢ Kernel Machine (KM) Regression for Linear Mixed Model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\gamma} \sim N(0, \tau\mathbf{W})$$

$$\mathbf{u} \sim N(0, \mathbf{K})$$

$$\boldsymbol{\varepsilon} \sim N\left(0, \sigma_{\mathrm{E}}^2\mathbf{I}\right)$$

where **W** is a predefined $q \times q$ diagonal weight matrix for each variant, and **K** is an $n \times n$ covariance matrix

For a linear mixed model, we use the log-likelihood

$$l = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{y}) = \tau\mathbf{G}\mathbf{W}\mathbf{G}' + \mathbf{K} + \sigma_E^2\mathbf{I}$. In the log-likelihood, the first term $-\frac{1}{2}\log|\boldsymbol{\Sigma}|$ is fixed and independent of trait **y** when replacing **Σ** with its estimator.

Sequence Kernel Association Test (SKAT):

➢ Kernel Machine (KM) Regression for Linear Mixed Model:

Take the first derivative

$$\frac{dl}{d\tau} = -\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{GWG}') + \frac{1}{2}(\mathbf{y}-\mathbf{X\beta})'\mathbf{\Sigma}^{-1}\mathbf{GWG}'\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{X\beta}),$$

The first term is fixed and independent of **y**. We take twice the second term to be derived as our test statistic Q.

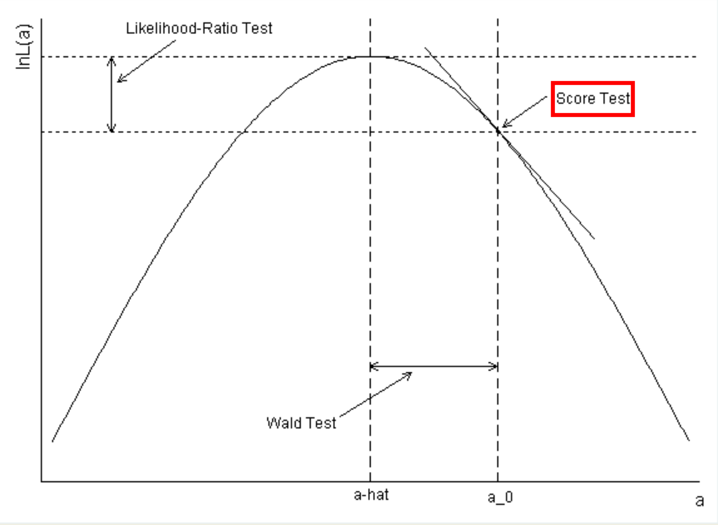
$$\mathrm{Q} = \left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\widehat{\mathbf{\Sigma}}^{-1}\mathbf{GWG}'\widehat{\mathbf{\Sigma}}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

where the parameters are estimated under H_0 (i.e., H_0: $\tau = 0$)

Thus, under H_0: $\quad \mathbf{y} = \mathbf{X\beta} + \mathbf{u} + \boldsymbol{\varepsilon}$

$$\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{K}} + \hat{\sigma}_E^2\mathbf{I}$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\mathbf{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{\Sigma}}^{-1}\mathbf{y}$$



Sequence Kernel Association Test (SKAT):

➢ Kernel Machine (KM) Regression for Linear Mixed Model:

Under null hypothesis, the variance of residual is

$$\mathrm{var}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)=\mathrm{var}\left(\mathbf{y}-\mathbf{X}\left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}\right)=\widehat{\boldsymbol{\Sigma}}-\mathbf{X}\left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'=\mathbf{P}_0.$$

The statistic Q is a quadratic form of $\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ and follows a mixture of chi-square distributions under H_0. Thus,

$$\mathrm{Q}\sim\sum_{i=1}^{q}\lambda_i\chi_{1,i}^2$$

where λ_i is the eigenvalues of the matrix $\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{P}_{\mathbf{0}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}^{\frac{1}{2}}$.

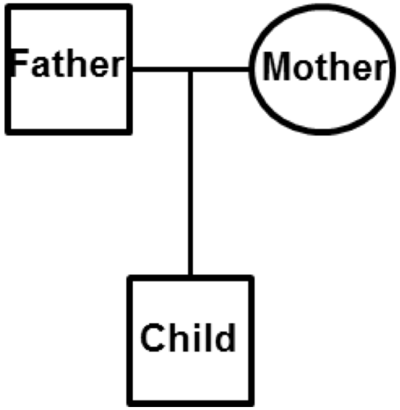
Sequence Kernel Association Test (SKAT):

- <u>Special case</u>: **Family Sequence Kernel Association Test (famSKAT) for Quantitative Traits for Family Data:**

The random variable for familial correlation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \underset{\downarrow}{\mathbf{u}} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\gamma} \sim N(0, \tau\mathbf{W}) \qquad \boldsymbol{\varepsilon} \sim N\left(0, \sigma_{\mathrm{E}}^{2}\mathbf{I}\right)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\delta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\delta} \sim N\left(0, \sigma_{\delta}^{2}\boldsymbol{\Phi}\right)$$



$$\boldsymbol{\Phi} = \begin{array}{cccc} \text{Father} & \text{Mother} & \text{Child} & \\ \left[\begin{matrix} 1 \\ 0 \\ 0.5 \end{matrix}\right. & \begin{matrix} 0 \\ 1 \\ 0.5 \end{matrix} & \left.\begin{matrix} 0.5 \\ 0.5 \\ 1 \end{matrix}\right] & \begin{matrix} \text{Father} \\ \text{Mother} \\ \text{Child} \end{matrix} \end{array}$$

<u>Under the null hypothesis ($\tau = 0$)</u>, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$

Sequence Kernel Association Test (SKAT):

- **<u>Special case</u>: Family Sequence Kernel Association Test (famSKAT) for Quantitative Traits for Family Data:**

We have test statistics:

$$\mathrm{Q} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

$$\widehat{\boldsymbol{\Sigma}} = \hat{\sigma}_{\delta}^{2}\boldsymbol{\Phi} + \hat{\sigma}_{E}^{2}\mathbf{I}$$

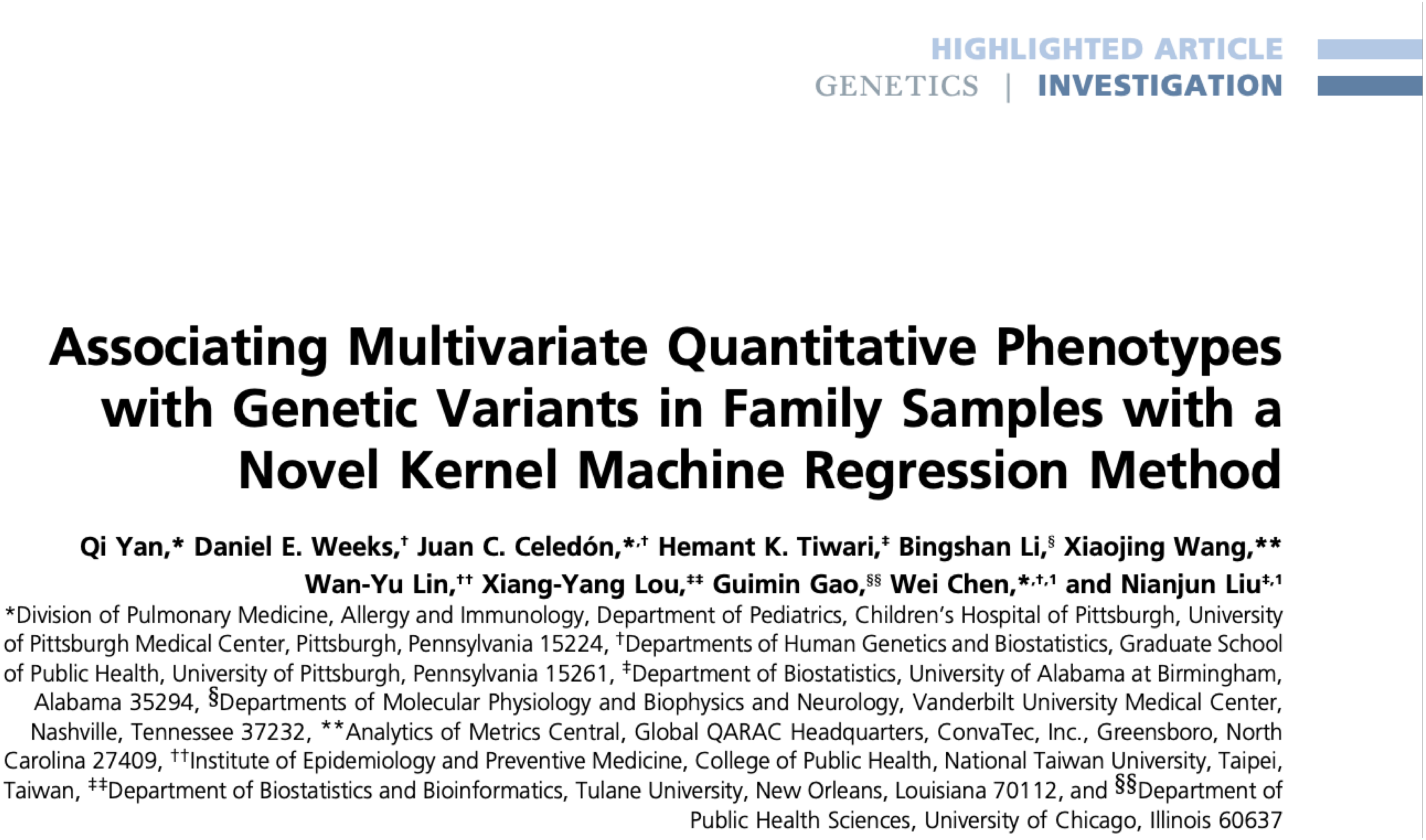
The statistic Q is a quadratic form of $\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ and follows a mixture of chi-square distributions

$$\mathrm{Q} \sim \sum_{i=1}^{q} \lambda_i \chi_{1,i}^{2}$$

where λ_i is the eigenvalues of the matrix $\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{P}_{0}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}^{\frac{1}{2}}$.

Sequence Kernel Association Test (SKAT):

- **<u>Special case</u>: Multivariate Family Kernel Machine (MF-KM) regression for Quantitative Traits for Family Data:**



HIGHLIGHTED ARTICLE
GENETICS | INVESTIGATION

Associating Multivariate Quantitative Phenotypes with Genetic Variants in Family Samples with a Novel Kernel Machine Regression Method

Qi Yan,[*] Daniel E. Weeks,[†] Juan C. Celedón,[*,†] Hemant K. Tiwari,[‡] Bingshan Li,[§] Xiaojing Wang,[**] Wan-Yu Lin,[††] Xiang-Yang Lou,[‡‡] Guimin Gao,[§§] Wei Chen,[*,†,1] and Nianjun Liu[‡,1]

[*]Division of Pulmonary Medicine, Allergy and Immunology, Department of Pediatrics, Children's Hospital of Pittsburgh, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania 15224, [†]Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pennsylvania 15261, [‡]Department of Biostatistics, University of Alabama at Birmingham, Alabama 35294, [§]Departments of Molecular Physiology and Biophysics and Neurology, Vanderbilt University Medical Center, Nashville, Tennessee 37232, [**]Analytics of Metrics Central, Global QARAC Headquarters, ConvaTec, Inc., Greensboro, North Carolina 27409, [††]Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan, [‡‡]Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, Louisiana 70112, and [§§]Department of Public Health Sciences, University of Chicago, Illinois 60637

Sequence Kernel Association Test (SKAT):

- **Special case: Multivariate Family Kernel Machine (MF-KM) regression for Quantitative Traits for Family Data:**

We consider a data set containing *m* individuals and two correlated phenotypes for illustration. The model with correlation among phenotypes and familial correlation is

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{G\gamma} + \mathbf{h} + \boldsymbol{\varepsilon}$$

where **y** is a vector of continuous trait (i.e., **y** = $(y_{11}, y_{12}, y_{21}, y_{22}, \ldots, y_{m1}, y_{m2})$ where *m* is the number of samples). **h** is the random effect of correlated phenotypes corresponding to the polygenic contribution, and **ε** is the random effect of correlated phenotypes corresponding to the random environmental contribution.

$$\mathbf{h} \sim N\left(0, \quad \mathbf{\Phi} \otimes \begin{pmatrix} \sigma_{G1}^2 & \sigma_{G12} \\ \sigma_{G12} & \sigma_{G2}^2 \end{pmatrix}\right) \qquad \boldsymbol{\varepsilon} \sim N\left(0, \quad \mathbf{I} \otimes \begin{pmatrix} \sigma_{E1}^2 & \sigma_{E12} \\ \sigma_{E12} & \sigma_{E2}^2 \end{pmatrix}\right)$$

Sequence Kernel Association Test (SKAT):

- **Special case: Multivariate Family Kernel Machine (MF-KM) regression for Quantitative Traits for Family Data:**

Under the null hypothesis ($\tau = 0$), $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\varepsilon}$

$$\mathrm{var}(\mathbf{y}) = \boldsymbol{\Phi} \otimes \begin{pmatrix} \sigma_{G1}^2 & \sigma_{G12} \\ \sigma_{G12} & \sigma_{G2}^2 \end{pmatrix} + \mathbf{I} \otimes \begin{pmatrix} \sigma_{E1}^2 & \sigma_{E12} \\ \sigma_{E12} & \sigma_{E2}^2 \end{pmatrix} = \boldsymbol{\Sigma}$$

where $\boldsymbol{\Phi}$ is twice the $m \times m$ kinship matrix obtained from familial relationship and $\otimes$ is the kronecker product. σ_{G1}^2, σ_{G2}^2, σ_{G12}, σ_{E1}^2, σ_{E2}^2 and σ_{E12} represent the polygenic variances of the first and second traits, the polygenic covariance between the two traits, the environmental variances of the first and second traits, and the environmental covariance between the two traits.

Sequence Kernel Association Test (SKAT):

- **Special case: Multivariate Family Kernel Machine (MF-KM) regression for Quantitative Traits for Family Data:**

We have test statistics:

$$\mathrm{Q} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{GWG}'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Phi}\otimes\begin{pmatrix} \hat{\sigma}_{G1}^{2} & \hat{\sigma}_{G12} \\ \hat{\sigma}_{G12} & \hat{\sigma}_{G2}^{2} \end{pmatrix} + \mathbf{I}\otimes\begin{pmatrix} \hat{\sigma}_{E1}^{2} & \hat{\sigma}_{E12} \\ \hat{\sigma}_{E12} & \hat{\sigma}_{E2}^{2} \end{pmatrix}$$

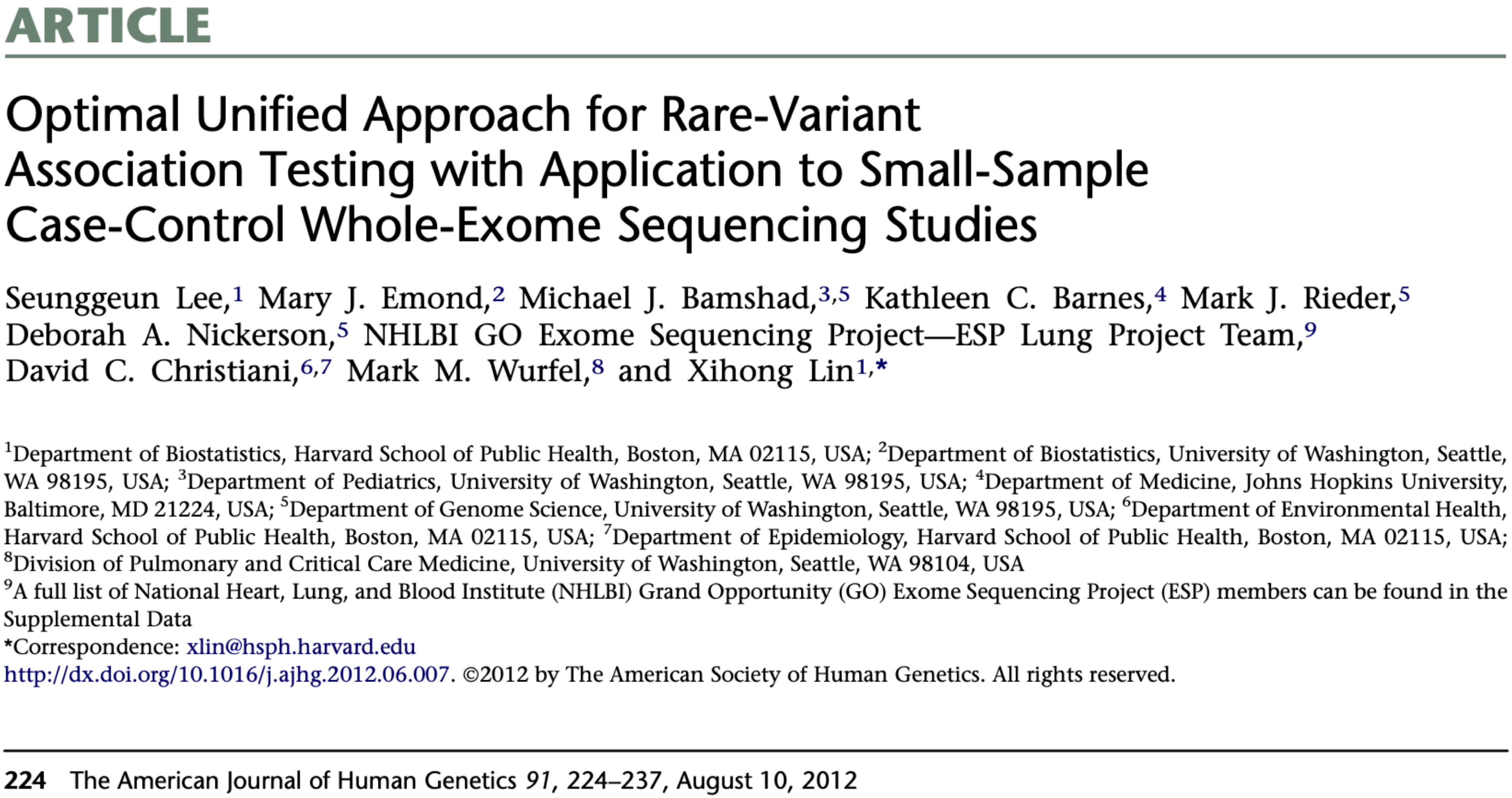
The statistic Q is a quadratic form of $\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ and follows a mixture of chi-square distributions

$$\mathrm{Q} \sim \sum_{i=1}^{q} \lambda_i \chi_{1,i}^{2}$$

where λ_i is the eigenvalues of the matrix $\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{P}_{\mathbf{0}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}^{\frac{1}{2}}$.

Sequence Kernel Association Test (SKAT-O):

- Balance between burden and SKAT



ARTICLE

Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies

Seunggeun Lee,[1] Mary J. Emond,[2] Michael J. Bamshad,[3,5] Kathleen C. Barnes,[4] Mark J. Rieder,[5] Deborah A. Nickerson,[5] NHLBI GO Exome Sequencing Project—ESP Lung Project Team,[9] David C. Christiani,[6,7] Mark M. Wurfel,[8] and Xihong Lin[1,*]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA; [2]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [3]Department of Pediatrics, University of Washington, Seattle, WA 98195, USA; [4]Department of Medicine, Johns Hopkins University, Baltimore, MD 21224, USA; [5]Department of Genome Science, University of Washington, Seattle, WA 98195, USA; [6]Department of Environmental Health, Harvard School of Public Health, Boston, MA 02115, USA; [7]Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA; [8]Division of Pulmonary and Critical Care Medicine, University of Washington, Seattle, WA 98104, USA

[9]A full list of National Heart, Lung, and Blood Institute (NHLBI) Grand Opportunity (GO) Exome Sequencing Project (ESP) members can be found in the Supplemental Data

*Correspondence: xlin@hsph.harvard.edu

http://dx.doi.org/10.1016/j.ajhg.2012.06.007. ©2012 by The American Society of Human Genetics. All rights reserved.

Sequence Kernel Association Test (SKAT-O):

- Balance between burden and SKAT

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{G\gamma} + \boldsymbol{\varepsilon}$$

We still test *H₀*: $\tau = 0$, assume $\boldsymbol{\gamma} \sim N\left(0,\ \ \tau \mathbf{W}^{\frac{1}{2}} \mathbf{R}_{\rho} \mathbf{W}^{\frac{1}{2}}\right)$ instead of $\boldsymbol{\gamma} \sim N(0,\ \ \tau\mathbf{W})$, where $\mathbf{R}_{\rho} = (1-\rho)I + \rho 11'$. In SKAT-O, $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\beta}}$ are calculated under the null hypothesis using the same approach as in SKAT. The SKAT test statistic is a function of ρ,

$$\mathrm{Q}_{\rho} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)' \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{G} \mathbf{W}^{\frac{1}{2}} \mathbf{R}_{\rho} \mathbf{W}^{\frac{1}{2}} \mathbf{G}' \widehat{\boldsymbol{\Sigma}}^{-1} \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

It is a SKAT test when $\rho = 0$, and it is a Burden test when $\rho = 1$. The statistic Q_{ρ} is a quadratic form of $\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ and follows a mixture of chi-square distributions under *H₀*. Thus,

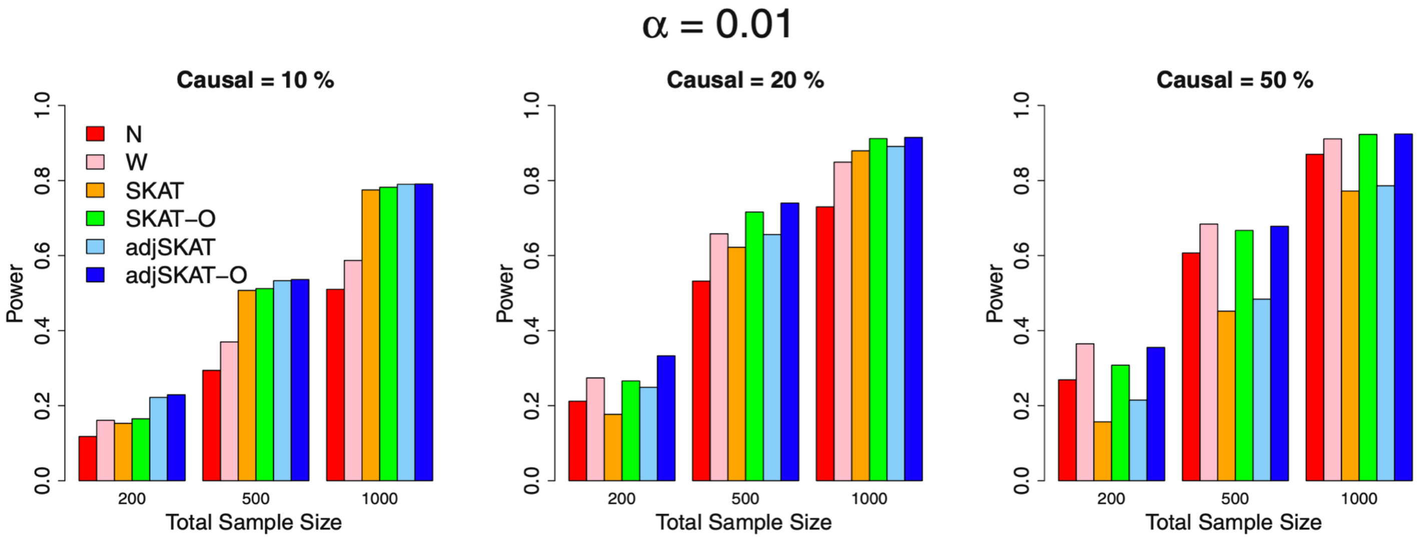
$$\mathrm{Q}_{\rho} \sim \textstyle\sum_{i=1}^{q} \lambda_i \chi_{1,i}^2,$$

where λ_i are the eigenvalues of the matrix $\mathbf{W}_{\rho}^{\frac{1}{2}} \mathbf{G}' \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{P}_{\mathbf{0}} \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{G} \mathbf{W}_{\rho}^{\frac{1}{2}}$ where $\mathbf{W}_{\rho} = \mathbf{W}^{\frac{1}{2}} \mathbf{R}_{\rho} \mathbf{W}^{\frac{1}{2}}$.

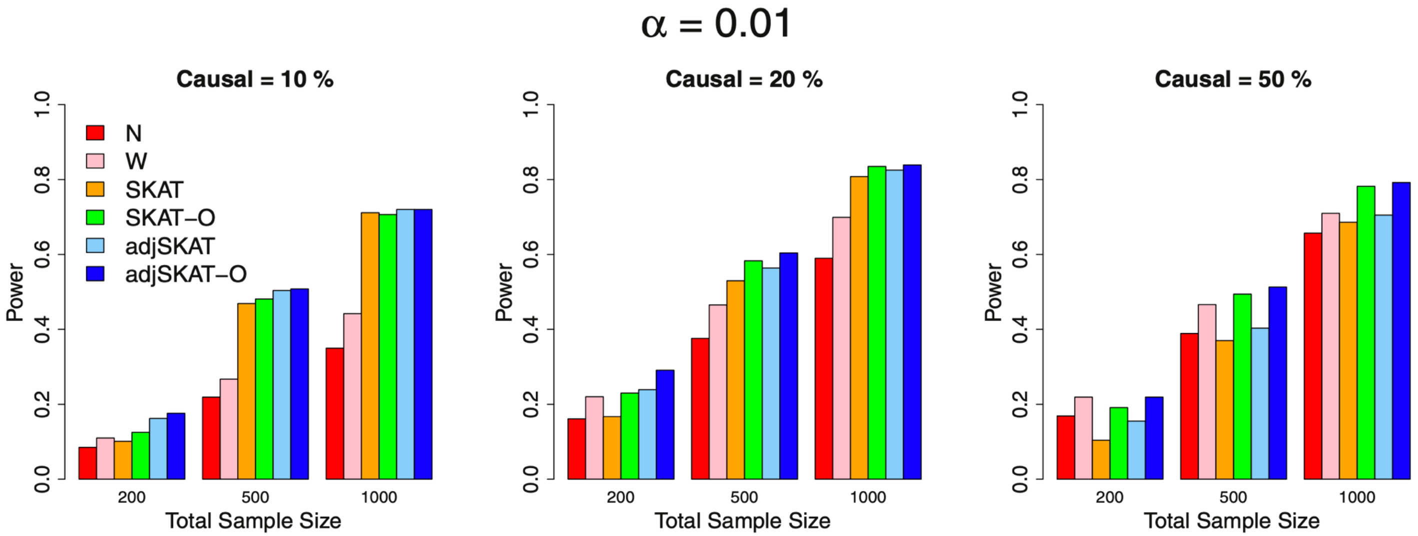
Key: auto search for $\boldsymbol{\rho}$.

Sequence Kernel Association Test (SKAT-O):

All Causal Variants Were Deleterious

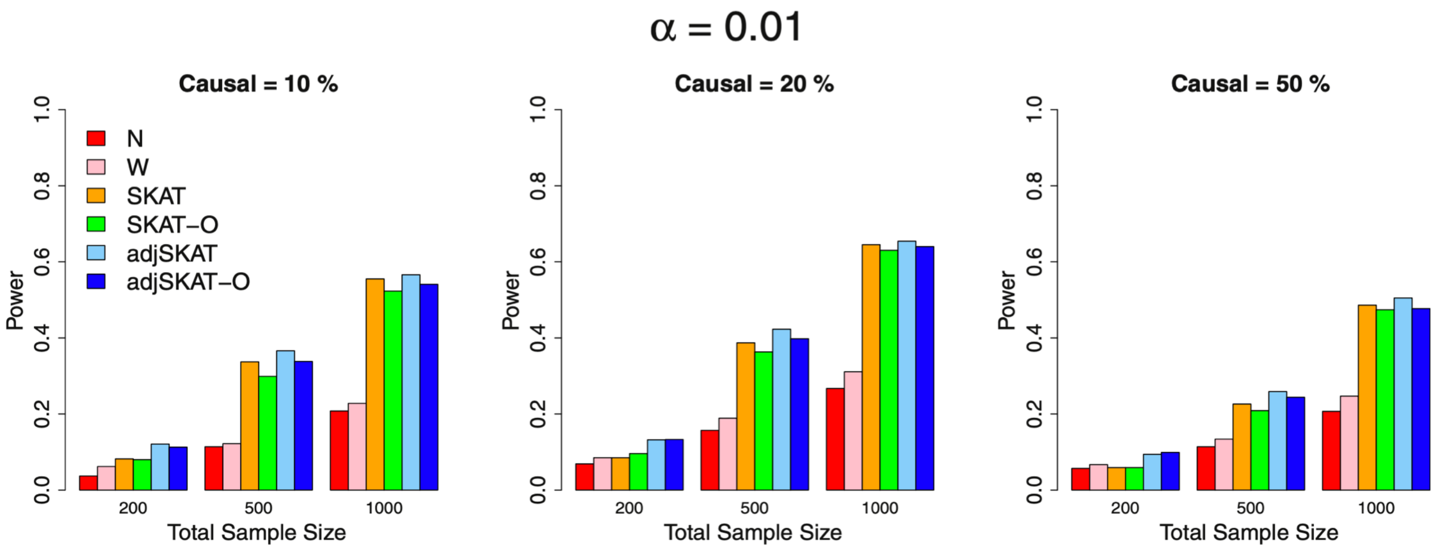


20%/80% of Causal Variants Were Protective/Deleterious



Sequence Kernel Association Test (SKAT-O):

50%/50% of Causal Variants Were Protective/Deleterious



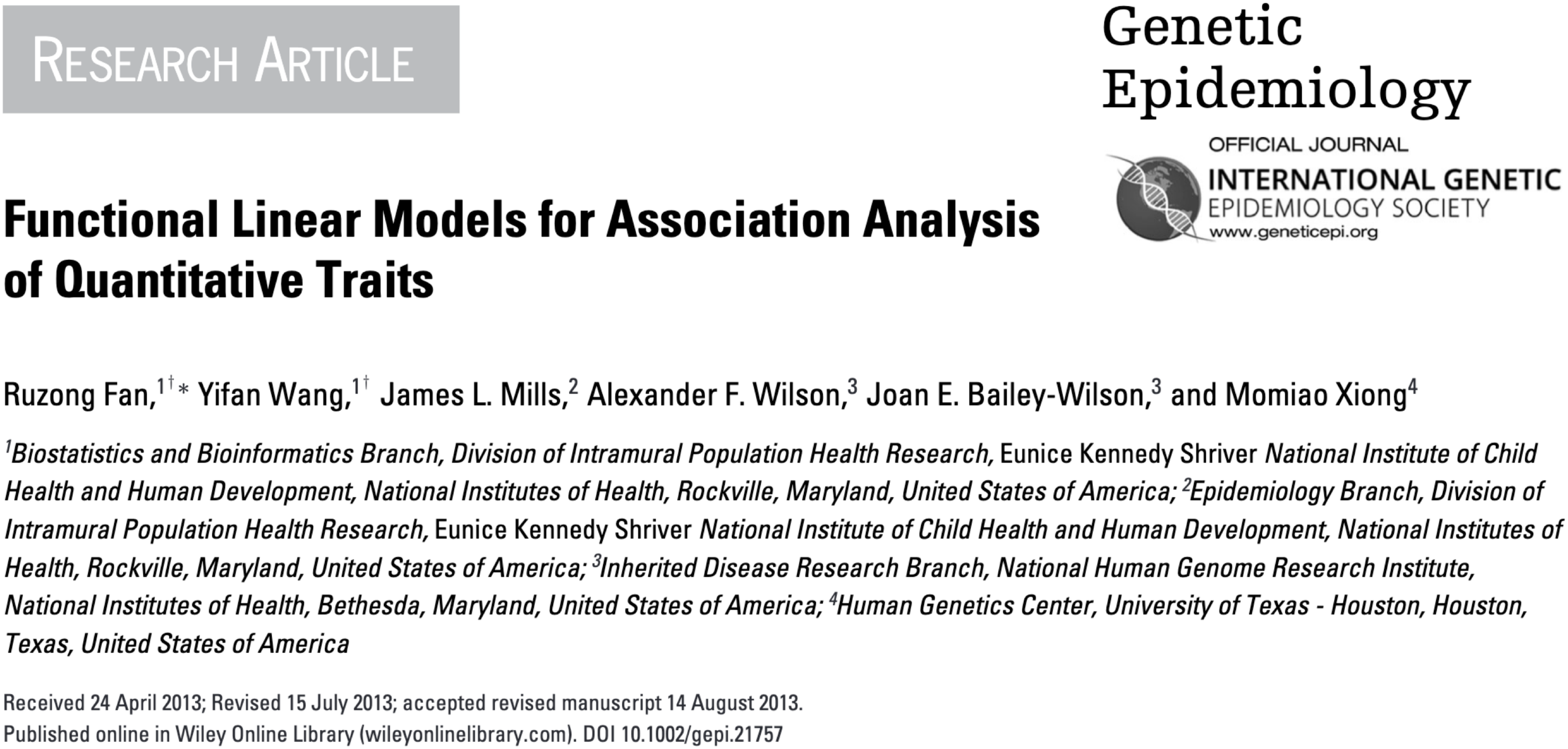
Alternate Tests for Rare Variant

- **Burden Test**

- **Sequence Kernel Association Test (SKAT)**

- **Function Linear Model (FLM)**

Functional Linear Model (FLM):



RESEARCH ARTICLE

Genetic
Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC
EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Functional Linear Models for Association Analysis of Quantitative Traits

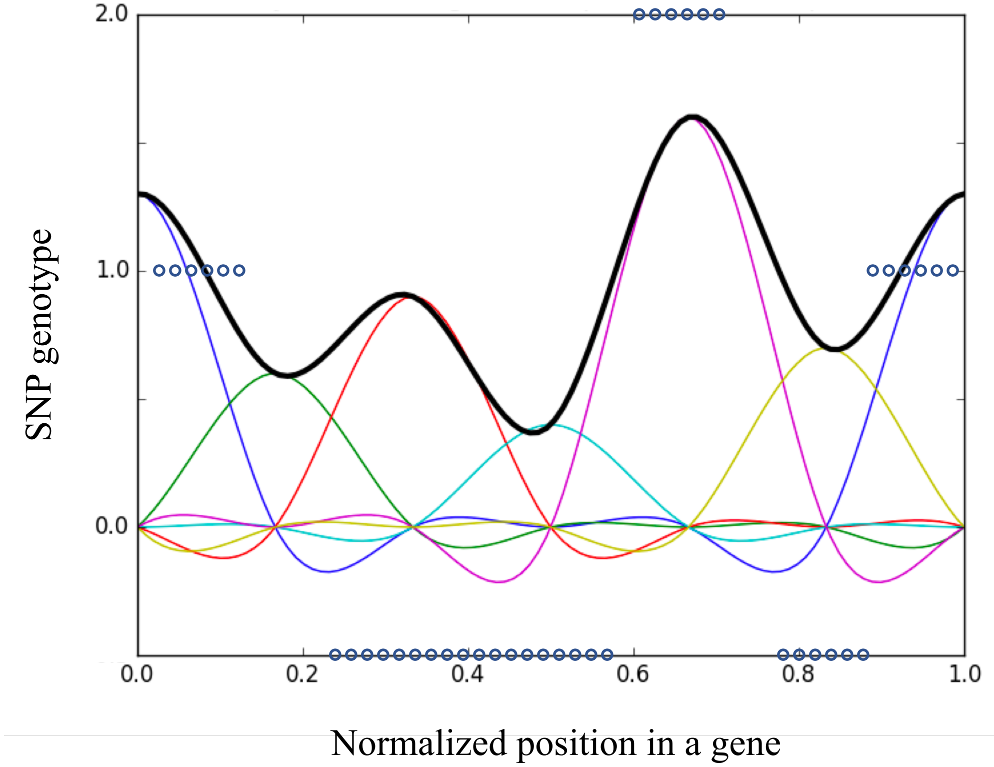
Ruzong Fan,[1†*] Yifan Wang,[1†] James L. Mills,[2] Alexander F. Wilson,[3] Joan E. Bailey-Wilson,[3] and Momiao Xiong[4]

[1]*Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research,* Eunice Kennedy Shriver *National Institute of Child Health and Human Development, National Institutes of Health, Rockville, Maryland, United States of America;* [2]*Epidemiology Branch, Division of Intramural Population Health Research,* Eunice Kennedy Shriver *National Institute of Child Health and Human Development, National Institutes of Health, Rockville, Maryland, United States of America;* [3]*Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America;* [4]*Human Genetics Center, University of Texas - Houston, Houston, Texas, United States of America*

Received 24 April 2013; Revised 15 July 2013; accepted revised manuscript 14 August 2013.
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21757

Functional Linear Model (FLM):

Fan et al., 2013, for a quantitative trait, we still consider a linear model,



Functional Linear Model (FLM):

Fan et al., 2013, for a quantitative trait, we still consider a linear model,

$$y_i = X_i'\beta + \int_0^1 G_i(t)\,\gamma(t)dt + \varepsilon_i$$

$$G_i(t) = \boxed{\left(G_i(t_1), \dots, G_i(t_q)\right)}\,\boxed{\Phi}[\Phi'\Phi]^{-1}\boxed{\phi(t)}$$

A series of basis functions of SNP positions (e.g., B-spline, Fourier) $K_1 \times 1$

$1 \times q$

$q \times K_1$ contains values of $\phi(t)$

Functional Linear Model (FLM):

Fan et al., 2013, for a quantitative trait, we still consider a linear model,

$$y_i = X_i'\beta + \int_0^1 G_i(t)\,\gamma(t)dt + \varepsilon_i$$

$$G_i(t) = \underbrace{\left(G_i(t_1),\dots,G_i(t_q)\right)}_{1\times q}\underbrace{\Phi[\Phi'\Phi]^{-1}}_{q\times K_1}\underbrace{\phi(t)}_{K_1\times 1}$$

$$\gamma(t) = \boxed{\theta'(t)}\boxed{(\gamma_1,\dots,\gamma_K)'} \qquad 1\times K$$

A series of basis functions
of SNP positions
(e.g., B-spline, Fourier)
1×K

Functional Linear Model (FLM):

Fan et al., 2013, for a quantitative trait, we still consider a linear model,

$$y_i = X_i'\beta + \int_0^1 G_i(t)\,\gamma(t)dt + \varepsilon_i$$

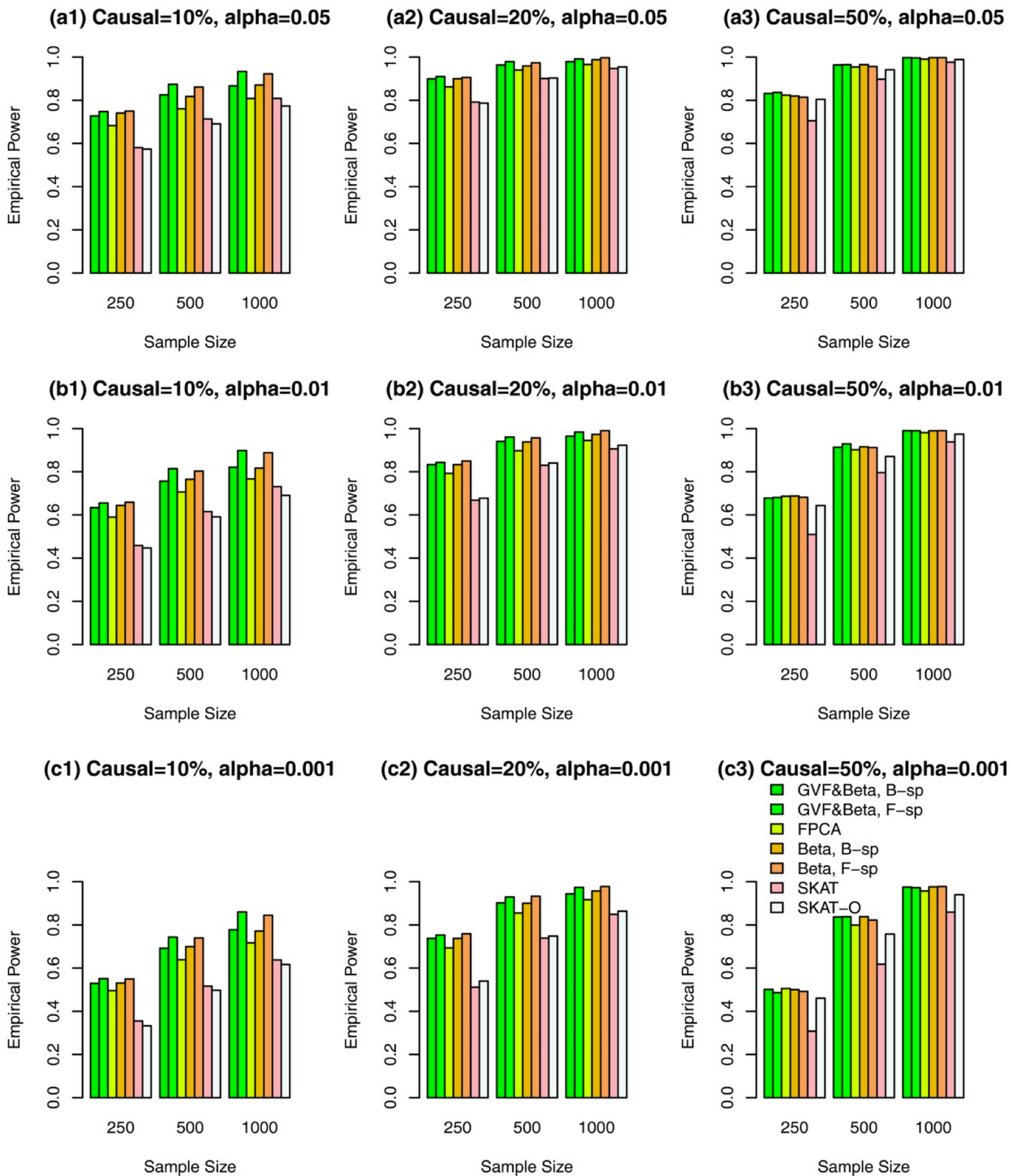
$$G_i(t) = \underbrace{\boxed{\left(G_i(t_1), \dots, G_i(t_q)\right)}}_{1\times q}\underbrace{\boxed{\Phi[\Phi'\Phi]^{-1}}}_{q\times K_1}\underbrace{\boxed{\phi(t)}}_{K_1\times 1} \qquad \gamma(t) = \underbrace{\boxed{\theta'(t)}}_{1\times K}\underbrace{\boxed{(\gamma_1, \dots, \gamma_K)'}}_{K\times 1}$$

Therefore, after some algebra,

$$y_i = X_i'\beta + R_i'\gamma + \varepsilon_i$$

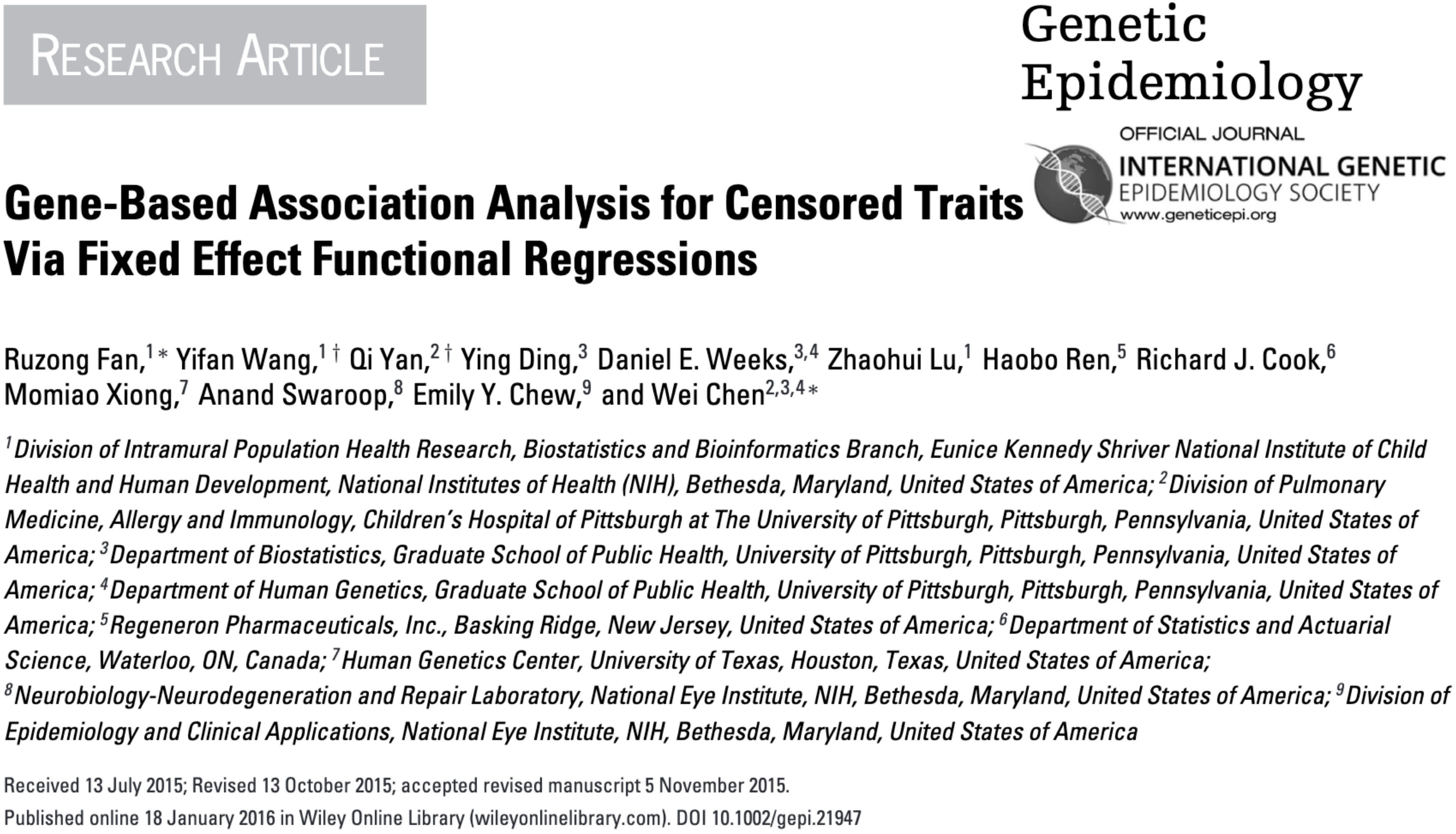
$$R_i = \underbrace{\left(G_i(t_1), \dots, G_i(t_q)\right)\Phi[\Phi'\Phi]^{-1}\int_0^1 \phi(t)\theta'(t)\,dt}_{1\times K}$$

Functional Linear Model (FLM):



Extensions: continuous, binary, family, multivariate, survival, meta …

Functional Linear Model (FLM):



RESEARCH ARTICLE

Genetic
Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC
EPIDEMIOLOGY SOCIETY
www.geneticepi.org

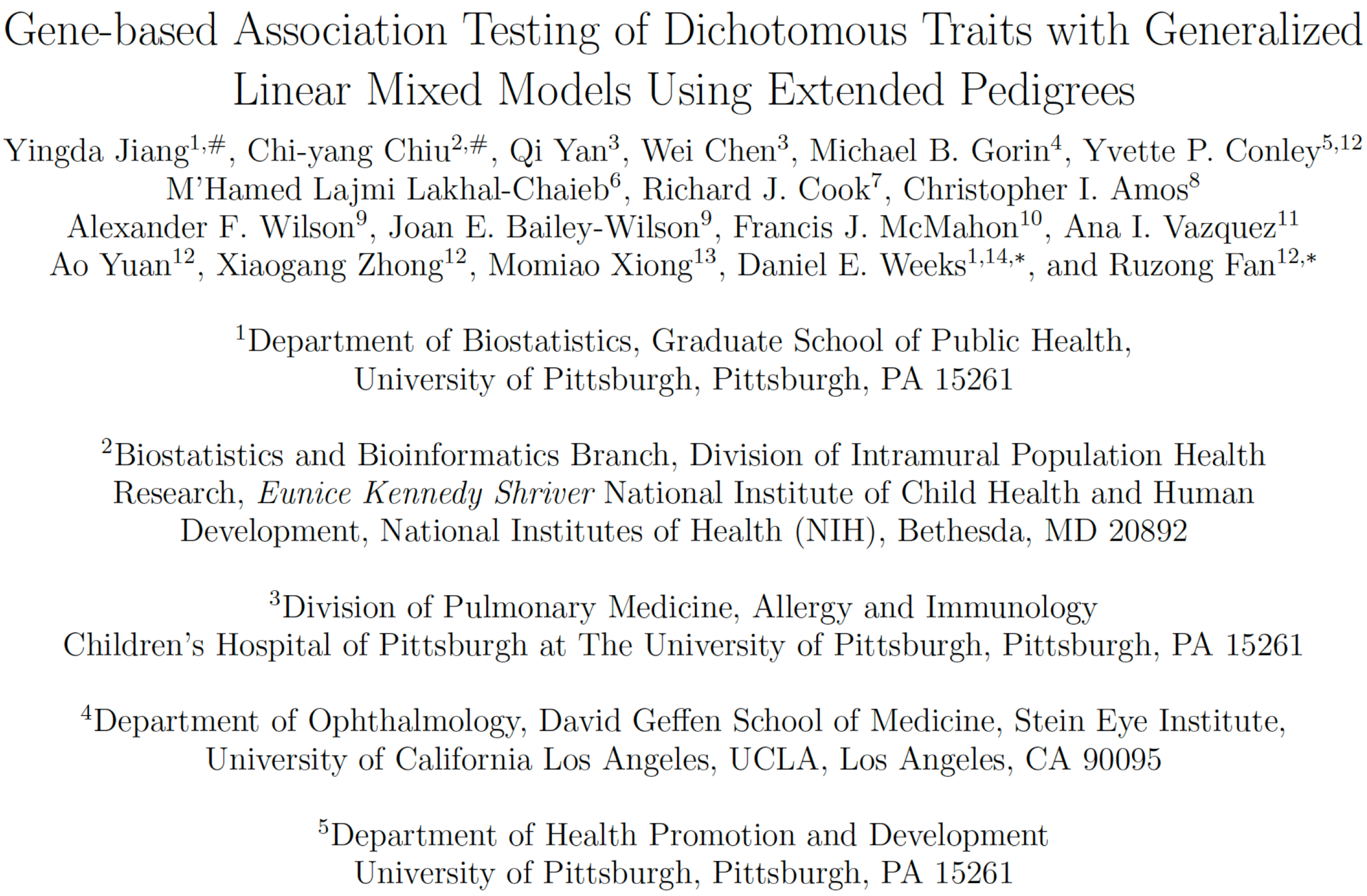
Gene-Based Association Analysis for Censored Traits Via Fixed Effect Functional Regressions

Ruzong Fan,[1]* Yifan Wang,[1]† Qi Yan,[2]† Ying Ding,[3] Daniel E. Weeks,[3,4] Zhaohui Lu,[1] Haobo Ren,[5] Richard J. Cook,[6] Momiao Xiong,[7] Anand Swaroop,[8] Emily Y. Chew,[9] and Wei Chen[2,3,4]*

[1] Division of Intramural Population Health Research, Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health (NIH), Bethesda, Maryland, United States of America; [2] Division of Pulmonary Medicine, Allergy and Immunology, Children's Hospital of Pittsburgh at The University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [3] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [4] Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [5] Regeneron Pharmaceuticals, Inc., Basking Ridge, New Jersey, United States of America; [6] Department of Statistics and Actuarial Science, Waterloo, ON, Canada; [7] Human Genetics Center, University of Texas, Houston, Texas, United States of America; [8] Neurobiology-Neurodegeneration and Repair Laboratory, National Eye Institute, NIH, Bethesda, Maryland, United States of America; [9] Division of Epidemiology and Clinical Applications, National Eye Institute, NIH, Bethesda, Maryland, United States of America

Received 13 July 2015; Revised 13 October 2015; accepted revised manuscript 5 November 2015.
Published online 18 January 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21947

Functional Linear Model (FLM):



Gene-based Association Testing of Dichotomous Traits with Generalized Linear Mixed Models Using Extended Pedigrees

Yingda Jiang[1,#], Chi-yang Chiu[2,#], Qi Yan[3], Wei Chen[3], Michael B. Gorin[4], Yvette P. Conley[5,12]
M'Hamed Lajmi Lakhal-Chaieb[6], Richard J. Cook[7], Christopher I. Amos[8]
Alexander F. Wilson[9], Joan E. Bailey-Wilson[9], Francis J. McMahon[10], Ana I. Vazquez[11]
Ao Yuan[12], Xiaogang Zhong[12], Momiao Xiong[13], Daniel E. Weeks[1,14,*], and Ruzong Fan[12,*]

[1]Department of Biostatistics, Graduate School of Public Health,
University of Pittsburgh, Pittsburgh, PA 15261

[2]Biostatistics and Bioinformatics Branch, Division of Intramural Population Health
Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human
Development, National Institutes of Health (NIH), Bethesda, MD 20892

[3]Division of Pulmonary Medicine, Allergy and Immunology
Children's Hospital of Pittsburgh at The University of Pittsburgh, Pittsburgh, PA 15261

[4]Department of Ophthalmology, David Geffen School of Medicine, Stein Eye Institute,
University of California Los Angeles, UCLA, Los Angeles, CA 90095

[5]Department of Health Promotion and Development
University of Pittsburgh, Pittsburgh, PA 15261

Journal of the American Statistical Association

Have single variant association tests been performed?

- Start with single variant tests
    - even though under-powered
    - provides a quality check
- Examine genome-wide QQ plots

From Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet. 2012 Oct 15;21(R1):R1-9. Epub 2012 Sep 13. PubMed PMID: 22983955; PubMed Central PMCID: PMC3459641.

What type of rare variant test to perform?

- Group rare variants, and compare to trait distribution
- Two major types:
  - with effect of all alleles in the same direction
  - allowing for alleles with variable effect directions
- Use variable threshold implementations
- Examine QQ plots (all analyses, combined with single variant results)

From Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet. 2012 Oct 15;21(R1):R1-9. Epub 2012 Sep 13. PubMed PMID: 22983955; PubMed Central PMCID: PMC3459641.

What allele frequency threshold to use for gene based tests?

- If can’t use variable threshold methods, then use a variety of frequency cut-offs
- Additional analysis: Examine homozygotes or compound heterozygotes for deleterious mutations.

From Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet. 2012 Oct 15;21(R1):R1-9. Epub 2012 Sep 13. PubMed PMID: 22983955; PubMed Central PMCID: PMC3459641.

What variants to include in the rare variant test?

- Include all missense, splice or stop altering variants, excluding only synonymous and non-coding variants.
- Focus on subset of variants predicted to have deleterious consequences.
- Focus on only splice, frame, and stop-altering variants.

From Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet. 2012 Oct 15;21(R1):R1-9. Epub 2012 Sep 13. PubMed PMID: 22983955; PubMed Central PMCID: PMC3459641.

What approach to correct for multiple testing?

- Use permutation-based approaches to assess statistical significance.
- Or proposed rule of thumb: need a p-value less than 5×10^{-7}.

From Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet. 2012 Oct 15;21(R1):R1-9. Epub 2012 Sep 13. PubMed PMID: 22983955; PubMed Central PMCID: PMC3459641.

Conclusions

- Mixture of risk, neutral, and protective variants
  - Probably should not assume all have same direction of effect
- Avoid arbitrary thresholds
  - Variable threshold models
- Many different statistics, with differing power under different conditions
  - Sensitivity analyses with a few different methods
- Always good to incorporate measures of data quality
  - Model uncertainty

From Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet. 2012 Oct 15;21(R1):R1-9. Epub 2012 Sep 13. PubMed PMID: 22983955; PubMed Central PMCID: PMC3459641.