

Doctoral Dissertation Proposal Defense

Data-adaptive SNP-set-based Association Tests of Longitudinal Traits

Yang Yang, M.S

UTSPH

Dec 15, 2014

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

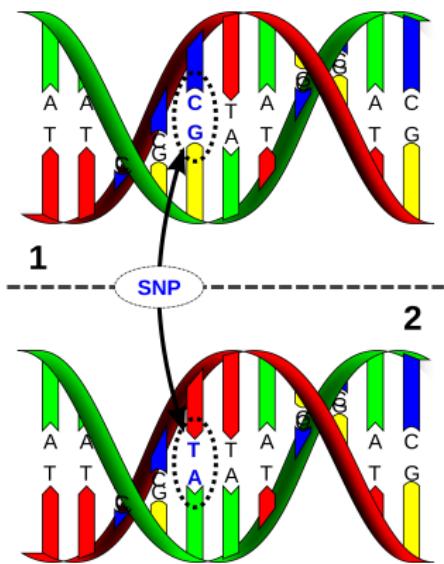
- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

Introduction to GWAS

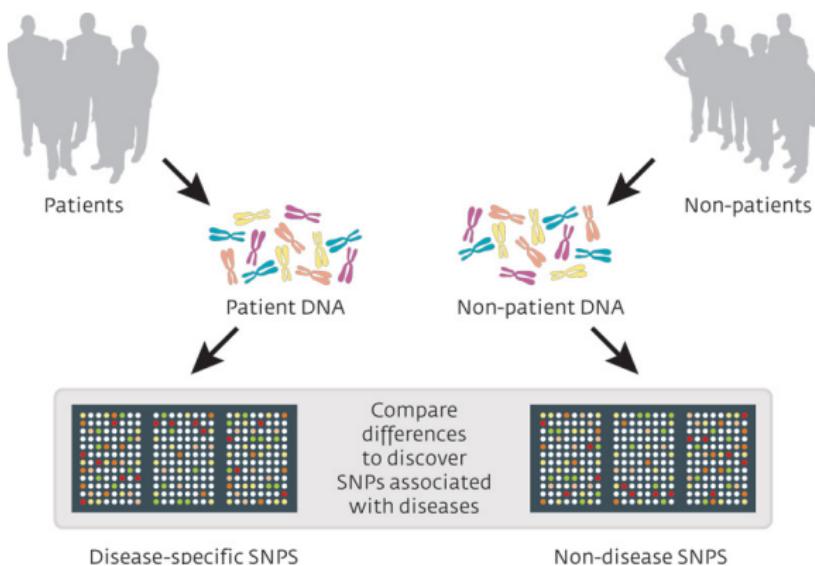
What is SNP?



A Single Nucleotide Polymorphism (SNP) is a DNA sequence variation occurring commonly within a population (e.g. 1%) in which a single nucleotide A, T, C or G in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes.

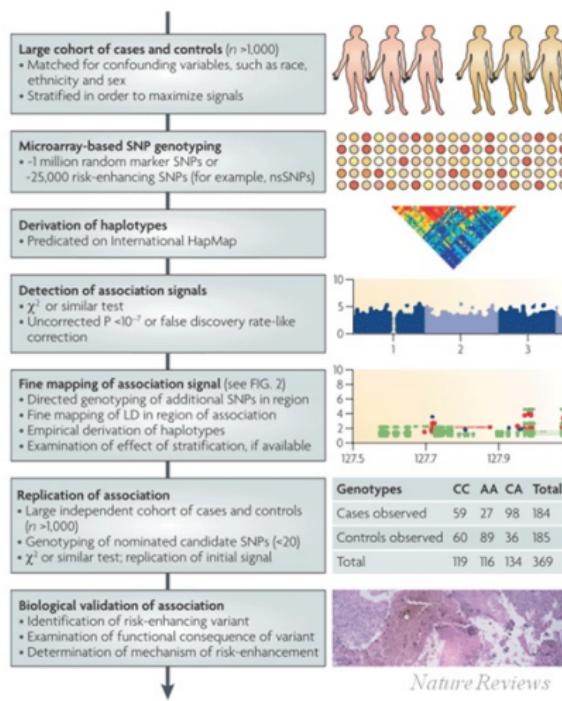
Introduction to GWAS

A simple flowchart



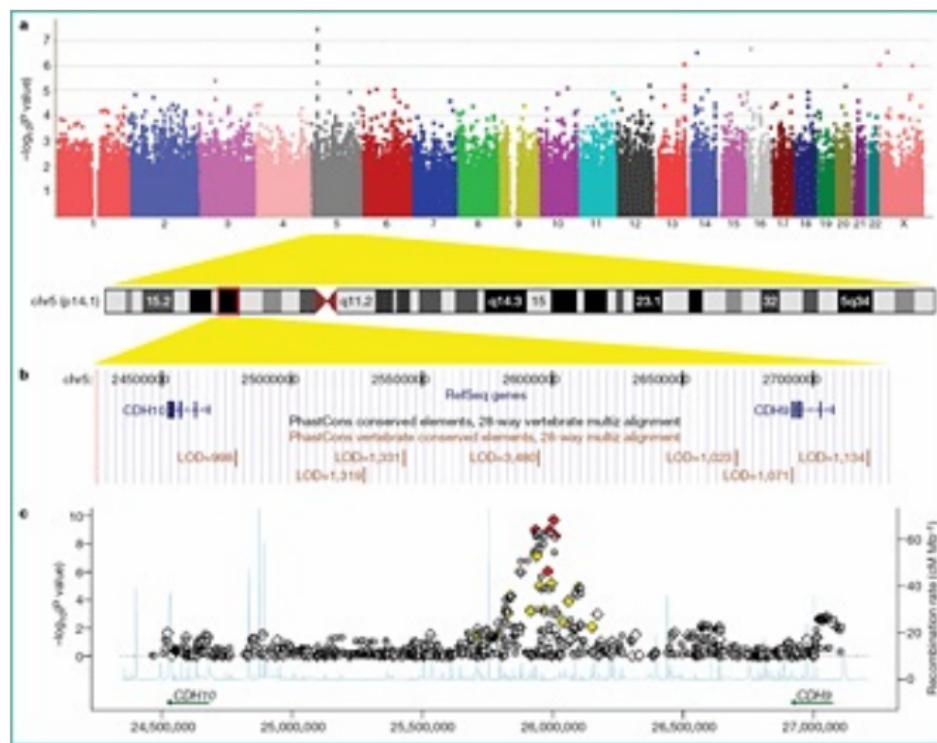
Introduction to GWAS

A more detailed flowchart



Introduction to GWAS

How does GWAS result look like?



Introduction to GWAS

Common variants and rare variants

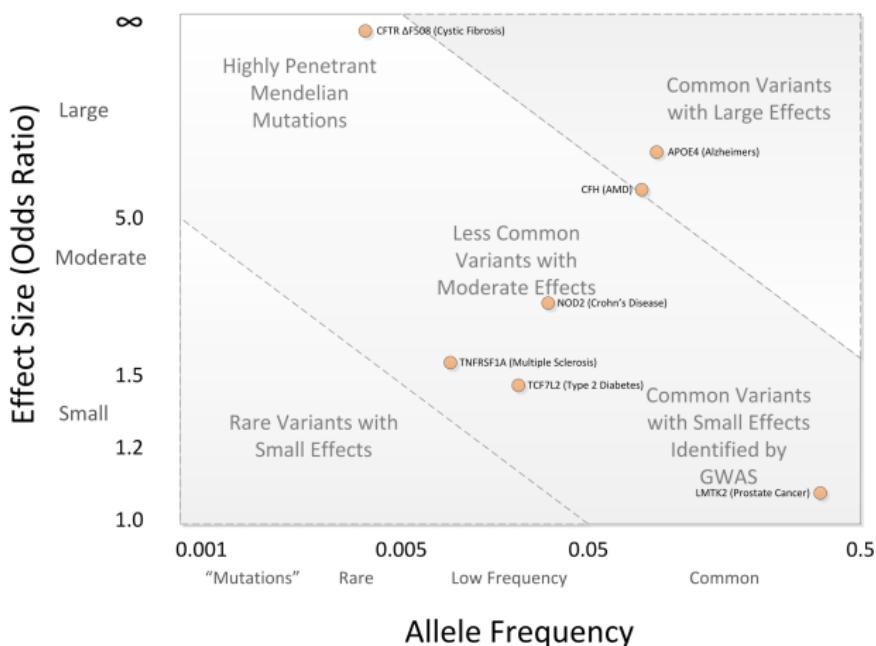


Figure: effect size of Single Nucleotide Variant [BM12]

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

single-SNP based association tests

the classical method

For individual i with SNP j coded as x_{ij} ($x_{ij} = 0, 1, 2$ representing copies of minor alleles) and a vector of covariates φ_i ,

$$g(\mu_i) = \beta_0 + x_{ij}\beta_j + z_i\varphi_i,$$

where $g(\mu_i)$ is a link function in Generalized Linear Model (GLM) to link types of outcome to the linear combination of predictors.

However, this method suffers from at least two disadvantages:

- 1), it will generate millions of tests thus increase the multiple test error correction burden;
- 2), the coefficient estimate of SNP j will become unstable or even the estimation algorithm cannot converge when SNP minor allele frequency (MAF) becomes smaller, e.g. $MAF < 0.01$.

SNP-set based association tests I

A brief review

By pooling multiple low MAF SNVs together, the SNP-set based association test can detect the signal(s) from a region (such as a gene) instead of from a single SNV.

Major categories of SNP-set based association tests:

- the so-called "burden test", which used MAF based weighting scheme to combine the sum statistics from multiple SNVs in a region [LL08, MB09];
- the variance-component test, which includes SKAT, C-alpha, SSU, etc [Pan09, NRV⁺11, WLC⁺11].
- the Lasso and group-penalized regression based methods [ZSSL10, KPS14].
- the functional linear model and functional principal component analysis based methods [LZX12b, LZX12a, LBX11, FWM⁺13].

SNP-set based association tests II

A brief review

- the adaptive test combines statistics of burden test and variance-component test, such as SKAT-O, aSum+aSSU+aScore, an exponential combination (EC) framework for set-based association tests, a robust and powerful test using Fisher's method to combine linear and quadratic statistics, a unified mixed-effect model, etc
[HP10, PS11, LEB⁺12, LWL12, CHG⁺12, DLS13, SZH13].

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS**
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

How do longitudinal data look like?

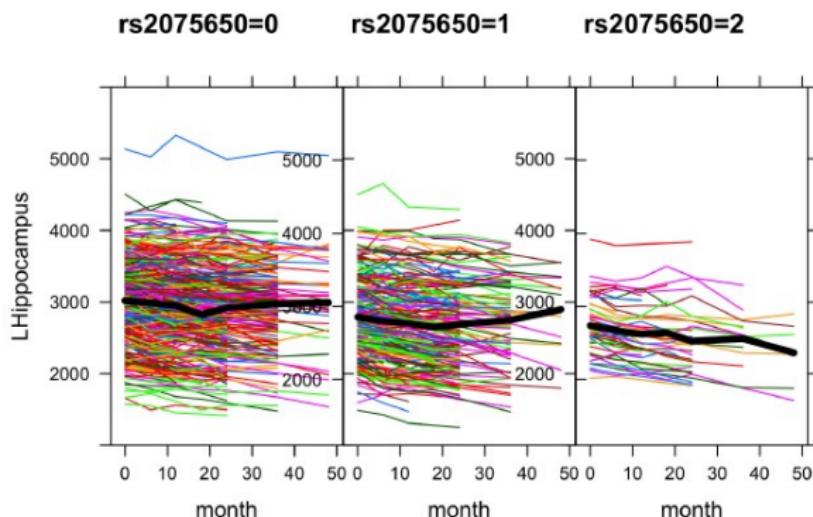


Figure: Trajectories of phenotype left hippocampus volume over time (in months) in three allele groups of SNP rs2075650 [XSP⁺14]

Why longitudinal? I

A recent study by Xu et al [XSP⁺14] demonstrates the power gain from longitudinal data analysis over traditional cross-sectional data analysis used in GWAS.

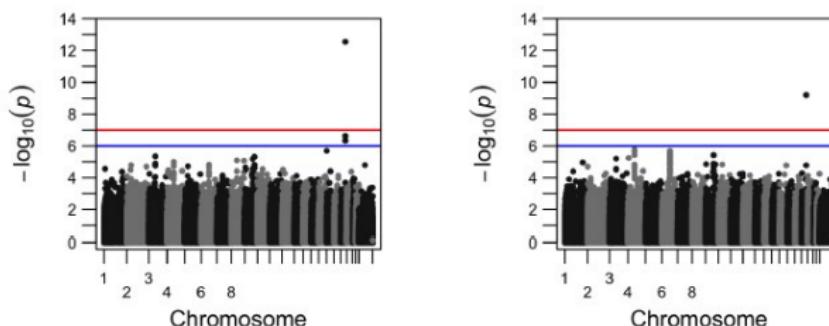


Figure: Comparison of the Manhattan plots for genome-wide p-values for phenotype left hippocampus volume from longitudinal analysis (left) and from cross-sectional analysis (right) [XSP⁺14]

Why longitudinal? II

Power

Model	rs2075650			rs439401		
	$P = 10^{-8}$	$P = 10^{-10}$	$P = 10^{-15}$	$P = 10^{-8}$	$P = 10^{-10}$	$P = 10^{-15}$
LME-RSI	1	1	0.859	0.872	0.677	0.148
GEE-Robust	1	1	0.857	0.871	0.676	0.170
Baseline	0.736	0.448	0.038	0.077	0.015	0

LME-RSI: a linear mixed-effects model with random slope and intercept; LME-RI: a linear mixed-effects model with only a random intercept term; GEE-Robust: GEE with the sandwich covariance estimator; GEE-Naive: GEE with the model-based covariance estimator; Baseline: a linear model at the baseline testing for the main effects of an SNP.

doi:10.1371/journal.pone.0102312.t008

Figure: Simulation results at significance level P with different methods [XSP⁺14]

A brief review of major longitudinal data analysis methods I

Major categories of longitudinal data analysis methods:

- random effect models

Random effect model is a two-stage models, which treat probability distributions for the response vectors of different individuals as a single family and the random-effects parameters which hold the same for the same individual as another distribution.

A brief review of major longitudinal data analysis methods II

- marginal effect models

Marginal effect model is an extension to quasi-likelihood method. Rather than giving subject-specific(SS) estimates as in random effect models, GEE gives population-averaged (PA) estimates by only describing the marginal expectation of the outcome variable as a function of the covariates and the variance is a known function of the mean, while accounting for the correlation among the repeated observations for a given subject by specifying a "working" correlation matrix.

A brief review of major longitudinal data analysis methods III

- transitional (Markov) models

The transitional (Markov) model, describes the conditional distribution of each response y_{ij} as an explicit function of first q prior observations

$y_{ij-1}, \dots, y_{ij-q}$ from history response vector: $H_{ij} = \{y_{ik}, k = 1, \dots, j-1\}$ and covariates x_{ij} . The integer q is referred as the order of the Markov models.

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

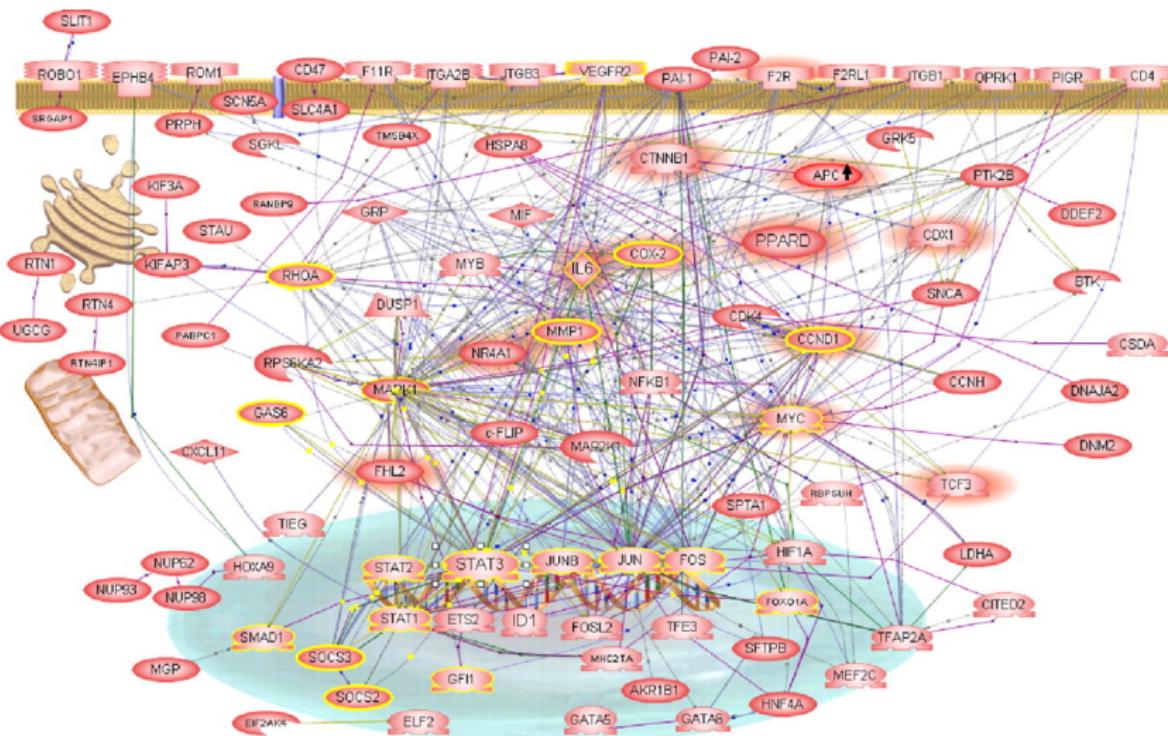
4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

A big picture



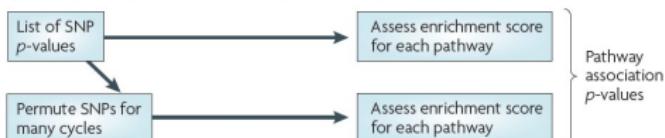
the advantage of using Gene-Set/Pathway based association test in GWAS

- it utilizes the information of biological pathway to help localize the association signal from close related genes
- it aggregates multiple Genes/RVs against testing each Gene/RV separately, which will boost the statistical power

Categories for Gene-Set/Pathway based association test in GWAS

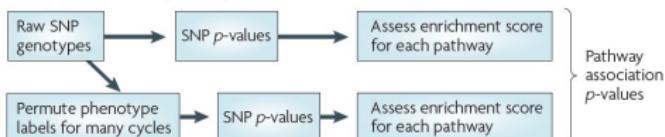
a SNP p-value enrichment approach:

Quick way to use precomputed whole-genome SNP p-values



Raw genotype approach:

In-depth analysis with phenotype permutation when raw genotype data are available



b 'Self-contained' tests



'Competitive' tests



Figure: Types of pathway association method [WLH10]

A brief review of current Gene-Set/Pathway based association tests in GWAS I

- GSEA modification in GWAS; GSEA-SNP;i-GSEA4GWAS
- modification of Fishers method for combining SNP P-values for gene-level or gene-set-level association
- gene set ridge regression in association studies (GRASS)
- association list go annotator (ALIGATOR), which is a 'p-value enrichment approach' requiring only pre-computed SNP p-values, uses Fisher's exact test on SNP with minimum p-value for the gene-level association
- the SNP ratio test (SRT), tests the ratio of significant SNPs in a pathway and compute the empirical p-value based on permutation
- supervised principal component analysis with a Gumbel extreme value mixture distribution as test statistic distribution and simulation-based standardization procedure for pathway size

A brief review of current Gene-Set/Pathway based association tests in GWAS II

- the Gene-loci Set Analysis (GLOSSI), at first uses the Cochran-Armitage trend test at single-marker level assuming an additive SNP effect, then uses Fisher's combination test to combine individual p-values of markers and corrected by Brown's approximation to better control type I error
- an adaptive rank truncated product (ARTP) statistic and permutation-based p-value adjustment to combine marker-level p-values to derive gene-level significance level and/or combine gene-level p-values to derive pathway-level significance level

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

Dissertation Aims

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework;
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

Public Health Significance I

- ① Due to the complexity in genetics association with phenotype, e.g. specific association effect directions and sizes, a given test favoring one scenario may or may not perform well in other scenarios [Pan09, DLS13, PKZ⁺14, SZH13]. In other words, there is no single test the most powerful among all testing scenarios. Therefore, a few data-adaptive tests were developed as an ad hoc strategy, e.g. some tests tried to combine the advantage of burden test and variance-component test; some other tests tried to use a set of pre-determined weights for individual RVs. Compared to the previous limited data-adaptive tests, our proposed method's **data adaptability** will be more extensive and generalized. The new tests will provide a relative high power in almost all data scenarios;

Public Health Significance II

- ② There is not yet a **SNP-set** based **data-adaptive** association test method for **longitudinal** data analysis in GWAS: we will propose such a new method to fill in this gap;
- ③ CVs and RVs are both important in finding the missing heritability of human complex disease. Our proposed new method will have the ability to handle both of them (either CVs or RVs);
- ④ We will extend the SNP-set based method to **Gene-set** based method to allow incorporating the biological pathway information and further avoid too few minor allele counts scenario in the association test;
- ⑤ We will produce an R package or independent Linux command-line based software implementing proposed methods to facilitate the community usage.

In conclusion, my dissertation work will provide useful methods/tools for identifying the underlying genetic factors explaining the heritability of human complex disease, and in the long run this will contribute to the prevention, diagnosis and cure of complex diseases.

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

Aim 1

Develop the robust data-adaptive association test for longitudinal data analysis within the Generalized Estimating Equation framework, which has relatively high power in most data scenarios and avoid drastic power loss in any single data scenario, as compared to current available methods. This is the first data-adaptive association test method for longitudinal data as to my knowledge.

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

Aim 1(a)

To develop a data-adaptive longitudinal association test within GEE framework for common variants, which will be done in either sliding-window based or gene-based manner for real GWAS data.

Aim 1(a) I

Method: introduction to notation and formula

Suppose for each subject $i = 1, \dots, n$, we have k total longitudinal measurements

$y_i = (y_{i1}, y_{i2}, \dots, y_{ik})'$ with y_{im} as a element, p SNPs of interest as a row vector

$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ with x_{ij} coded as 0, 1 or 2 for the count of the minor allele for SNP

$j = 1, \dots, p$, and $z_i = (z_{i1}, z_{i2}, \dots, z_{iq})$ is a row vector for q variates. We assume common effect sizes of the SNPs and covariates on the longitudinal phenotype/trait measurements. Thus, we construct the design matrix for the SNPs and covariates as:

$$X_i = \begin{pmatrix} x_i \\ x_i \\ \vdots \\ x_i \end{pmatrix}, Z_i = \begin{pmatrix} 1 & z_i \\ 1 & z_i \\ \vdots & \vdots \\ 1 & z_i \end{pmatrix}$$

where x_i and z_i are row vectors of length p and q respectively. X_i is a $k \times p$ matrix, and Z_i is a $k \times (q + 1)$ matrix.

Denote the regression coefficient $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ and $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_{q+1})'$ for X_i and Z_i respectively. The marginal mean of each measurement, $E(y_{im}|x_i, z_i) = \mu_{im}$ where $m = 1, 2, \dots, k$ for k total measurements, or the vectorized format of all measurements, $E(y_i|x_i, z_i) = \mu_i$, relates to the SNPs and covariates through a generalized linear model (GLM):

$$g(\mu_i) = \eta_i = Z_i \varphi + X_i \beta = H_i \theta$$

Aim 1(a) II

Method: introduction to notation and formula

with $H_i = (Z_i, X_i)$, $\theta = (\varphi', \beta')'$ and $g(\cdot)$ as a suitable link function. For continuous outcome, an identity link is usually used.

The consistent and asymptotically Normal estimates of β and φ can be obtained by solving the GEE [LZ86]:

$$U(\varphi, \beta) = \sum_{i=1}^n U_i(\varphi, \beta) = \sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \theta'} \right)' V_i^{-1} (Y_i - \mu_i) = 0,$$

with

$$\frac{\partial \mu_i}{\partial \theta'} = \frac{\partial g^{-1}(H_i \theta)}{\partial \theta'}, V_i = \phi A_i^{\frac{1}{2}} R_w A_i^{\frac{1}{2}},$$

and

$$A_i = \begin{bmatrix} v(\mu_{i1}) & 0 & \cdots & 0 \\ 0 & v(\mu_{i2}) & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & v(\mu_{ik}) \end{bmatrix}$$

Aim 1(a) III

Method: introduction to notation and formula

ϕ in V_i is the dispersion parameter in GEE and is usually treated as nuisance parameter.

$v(\mu_{im}) = \phi \text{Var}(y_{im}|x_i, z_i)$. $R_w(\alpha)$ is a working correlation matrix depending on some unknown parameter α .

With a canonical link function and a working independence model, we have a closed form of the U vector with two parts corresponding to SNPs and covariates, and its covariance estimator:

$$U = (U'_{.1}, U'_{.2})' = \sum_i (Z_i, X_i)' (Y_i - \mu_i)$$

$$\tilde{\Sigma} = \widehat{\text{Cov}}(U) = \sum_i (Z_i, X_i)' \widehat{\text{var}}(Y_i) (Z_i, X_i) = \sum_i (Z_i, X_i)' (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' (Z_i, X_i) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \quad (1)$$

Aim 1(a) IV

Method: introduction to notation and formula

Quantitative traits

We use the identity link, i.e. $g(\mu_{im}) = \mu_{im}$ and $v(\mu_{im}) = \phi \times 1 = \phi$. Then we have:

$$\begin{aligned} U &= \sum_i (Z_i, X_i)' R_w^{-1} (Y_i - \mu_i) \\ \tilde{\Sigma} &= \sum_i (Z_i, X_i)' R_w^{-1} (Y_i - \hat{\mu}_i) (Y_i - \hat{\mu}_i)' R_w^{-1} (Z_i, X_i) \end{aligned} \quad (2)$$

if the assumption of a common covariance matrices across Y_i for i is valid, e.g. for quantitative continuous traits study [Pan01], we can adopt a more efficient covariance estimator:

$$\tilde{\Sigma} = \sum_{i=1}^n (Z_i, X_i)' \widehat{\text{var}(Y_i)} (Z_i, X_i) = \sum_{i=1}^n (Z_i, X_i)' \left(\sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'}{n} \right) (Z_i, X_i) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

which is used by default for its better finite-sample performance [Pan01].

Aim 1(a) V

Method: introduction to notation and formula

Binary traits

For binary traits (trait value coded as 0 and 1), we use the logit link function so that

$g(\mu_{im}) = \log \frac{\mu_{im}}{1-\mu_{im}}$ and $v(\mu_{im}) = \mu_{im}(1 - \mu_{im})$. Additionally the (m, l) th element of $\frac{\partial \mu_i}{\partial \theta'}$ is $H_{i,ml}\mu_{im}(1 - \mu_{im})$ with $H_{i,ml}$ as the (m, l) th element of H_i , which is the short notation for (Z_i, X_i) .

Then we have:

$$\begin{aligned} U &= \sum_{i=1} (\frac{\partial \mu_i}{\partial \theta'})' V_i^{-1} (Y_i - \mu_i) \\ &= \sum_{i=1} (\frac{\partial \mu_i}{\partial \theta'})' \phi A_i^{-\frac{1}{2}} R_w^{-1} A_i^{-\frac{1}{2}} (Y_i - \mu_i) \end{aligned}$$

and

$$\begin{aligned} \tilde{\Sigma} &= \sum_i \left(\frac{\partial \mu_i}{\partial \theta'} \right)' \phi A_i^{-\frac{1}{2}} R_w^{-1} A_i^{-\frac{1}{2}} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' \phi A_i^{-\frac{1}{2}} R_w^{-1} A_i^{-\frac{1}{2}} \left(\frac{\partial \mu_i}{\partial \theta'} \right) \\ &= \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \end{aligned}$$

Aim 1(a) VI

Method: introduction to notation and formula

In my dissertation, I will **focus on** the case with quantitative traits, since they are most typical traits used as the response variable in longitudinal data analysis. In general, the only difference lies in which canonical link we will use, with all other equations/formulas remaining the same.

Aim 1(a) VII

Method: introduction to notation and formula

Our goal is to detect whether there is any association between the longitudinal trait and the SNPs via testing on hypothesis $H_0 : \beta = (\beta_1, \beta_2, \dots, \beta_p)' = 0$. We have under the null hypothesis with $g(Y_i) = Z_i\varphi$ to obtain φ and predict $\hat{\mu} = g^{-1}(Z\hat{\varphi})$. We hereby have score vector under the null hypothesis, with a working independence model, is:

$$U(\hat{\varphi}, 0) = (U_{.1}', U_{.2}')' = \sum_{i=1}^n (U_{i1}', U_{i2}')'$$

where

$$U_{.1} = \sum_i Z_i'(Y_i - \hat{\mu}_i), \quad U_{.2} = \sum_i X_i'(Y_i - \hat{\mu}_i)$$

As U asymptotically follows a multivariate normal distribution under H_0 , then the score vector for β also has an asymptotic normal distribution:

$$U_{.2} \sim N(0, \Sigma_{.2}), \quad \widehat{\text{Cov}}(U_{.2}) = V_{22} - V_{21}V_{11}^{-1}V_{12}$$

, where V_{xx} are defined in Equation 1.

Aim 1(a) VIII

Method: introduction to notation and formula

Several classical tests:

- **The Wald Test:** The Wald Test known as $T = \hat{\beta}'\text{cov}(\hat{\beta})\hat{\beta}$ is most commonly used, where $\hat{\beta}$ is the estimate of β after fitting the full GEE model with $g(\mu_i) = Z_i\varphi + X_i\beta$. Under H_0 , we have $T \sim \chi_p^2$. The Wald test is more time consuming by fitting full model, may fail to converge with many SNPs put on RHS of the regression-like equation to test, and more importantly, the type I error tends to inflate in such case [PKZ⁺¹⁴, ZXSP14].
- **The Score Test:** $T = U_{.2}'\Sigma_{.2}^{-1}U_{.2}^{-1}$, where $U_{.2}$ and $\Sigma_{.2}$ are discussed above; the statistic is asymptotically equivalent to the Wald test with the same null distribution $T \sim \chi_p^2$. Since we only need to fit the null model with covariates, it is computationally easier and less likely to have numerical convergence problems. More importantly, the score test controls the type I error well [PKZ⁺¹⁴, ZXSP14].

Aim 1(a) IX

Method: introduction to notation and formula

- The UminP Test:** $T = \max_j \frac{U_{.2,j}^2}{\Sigma_{.2,jj}}$ for $j \in 1, 2, \dots, p$, of j th SNP effect. The $\Sigma_{.2,jj}$ is the j th entry on the diagonal of $\Sigma_{.2}$. With $\max_j T$, we can get minimal p-value accordingly. A simulation method based on the asymptotic normal distribution of the score vector can be used to calculate its p-value [PKZ⁺14, ZXSP14]. Specifically, we first simulate the score vector $U_{(b)} = (U_{(b).1}, U_{(b).2}, \dots, U_{(b).p})'$ from its null distribution $U_{(b)} \sim N(0, \Sigma_{.2})$ for $b = 1, 2, \dots, B$, then calculate a total number of B null statistics:

$$T^{(b)} = \max_{j=1,\dots,p} \frac{U_{(b),j}^2}{\Sigma_{.2,jj}}, \text{ and the p-value is calculated as } \sum_{b=1}^B \frac{I(T^{(b)} \geq T) + 1}{B+1}.$$

With a working independence correlation matrix $R_w = I$, every element $\frac{U_{.2,j}^2}{\Sigma_{.2,jj}}$ is equivalent to running the model on each single SNP (e.g. j th) one by one and get the Score test statistics. Hence, in this condition, the GEE-UminP test is equivalent to the usual UminP test that combines multiple single-SNP based longitudinal association test statistics.

Aim 1(a) I

Method: A new class of tests and a data-adaptive test in longitudinal data settings

Before I introduce the proposed new test method, let me explain the logic in current GEE and Score test based methods.

$$T_{Sum} = \mathbf{1}' \mathbf{U} = \sum_{j=1}^p U_j, \quad T_{SSU} = \mathbf{U}' \mathbf{U} = \sum_{j=1}^p U_j^2,$$

These two tests are called Sum test and SSU test [Pan09].

Sum test and SSU test are all based on score vector. A more general form of score-based statistic can be generalized as:

$$T_w = \mathbf{W}' \mathbf{U} = \sum_{j=1}^p W_j U_j$$

where $\mathbf{W} = (W_1, \dots, W_p)'$ is a vector of weights for the p SNVs [LT11].

Aim 1(a) II

Method: A new class of tests and a data-adaptive test in longitudinal data settings

If we choose weight to be

$$W_j = U_{.2,j}^{\gamma-1}$$

for a series of integer value $\gamma = 1, 2, \dots, \infty$, leading to the sum of powered score (U) tests called SPU tests:

$$T_{SPU(\gamma)} = \sum_{j=1}^p U_{.2,j}^{\gamma-1} U_{.2,j}$$

Aim 1(a) III

Method: A new class of tests and a data-adaptive test in longitudinal data settings

When $\gamma \rightarrow \infty$ as an extreme situation, where ∞ is assumed to be an even number, we have

$$T_{SPU(\gamma)} \propto \|U\|_\gamma = \left(\sum_{j=1}^p |U_{.2,j}|^\gamma \right)^{\frac{1}{\gamma}} \rightarrow \|U\|_\infty = \max_{j=1}^p |U_{.2,j}| \equiv T_{SPU(\infty)}.$$

which takes only the largest element (in absolute value) of score vector. Apparently, $SPU(\infty)$ is equivalent to UminP test except the variance of each score component is replaced by 1 as in the denominator part.

In our experience, $SPU(\gamma)$ test with a large $\gamma > 8$ usually gave similar results as that of $SPU(\infty)$ test [PKZ⁺14], thus we will only use $\gamma \in \Gamma = \{1, 2, \dots, 8, \infty\}$ for the whole dissertation work.

Aim 1(a) IV

Method: A new class of tests and a data-adaptive test in longitudinal data settings

Simulation-based P-value estimation of $T_{SPU(\gamma)}$

Suppose the sample size is large enough or MAF of SNV is large enough for the asymptotic normal distribution of score vector to hold under null hypothesis, we will use a simulation method to calculate the p-value from each $T_{SPU(\gamma)}$ [Lin05, SMM05].

Specifically, suppose T is short notation of $T_{SPU(\gamma)}$ for a specific γ and $\hat{\Sigma}_{.2}$ is the covariance matrix of the score vector $U_{.2}$ based on original data (see 1). We draw B samples of the score vector from its null distribution: $U_{.2}^{(b)} \sim MVN\left(0, \hat{\Sigma}_{.2}\right)$, with $b = 1, 2, \dots, B$, and thus obtain a statistics under null hypothesis: $T^{(b)} = \sum_{j=1}^p U_{.2,j}^{(b)\gamma}$. We then can calculate the p-value of $T_{SPU(\gamma)}$ as $P_{SPU(\gamma)} = \sum_{b=1}^B \frac{I(T^{(b)} \geq T^{obs}) + 1}{B+1}$.

Aim 1(a) V

Method: A new class of tests and a data-adaptive test in longitudinal data settings

The aSPU test

Although we have a list of $SPU(\gamma)$ statistics and p-values, we are not sure which one is the most powerful in a specific data situation. Thus, it will be convenient to have a test which data-adaptively and automatically select/combine the best $SPU(\gamma)$ test(s). We hereby propose an adaptive SPU (aSPU) test to achieve such purpose. As for different γ , it is difficult to characterize the power curve of an SPU test in real data situation, we will use the p-value of a SPU test to approximate its power; this idea has been prevalent in practice. Accordingly, we will have the aSPU test statistic:

$$T_{aSPU} = \min_{\gamma \in \Gamma} P_{SPU(\gamma)},$$

where $P_{SPU(\gamma)}$ is the p-value of a specific $SPU(\gamma)$ test.

Aim 1(a) VI

Method: A new class of tests and a data-adaptive test in longitudinal data settings

Simulation-based P-value estimation of T_{aSPU}

Similarly as the above simulation method to get p-value of $T_{SPU(\gamma)}$, the *same strategy* can be applied to get the p-value of T_{aSPU} and actually it fully utilizes the previous simulated intermediate result, hereby saves another *unnecessary* simulation work. Specifically, at the SPU test stage we already have the $U_{.2}^{(b)}$ for $b = 1, 2, \dots, B$. We then calculate the corresponding SPU test statistics $T_{SPU(\gamma)}^{(b)}$ and p-value

$$P_{SPU(\gamma)}^{(b)} = \sum_{b_1 \neq b}^B \frac{I(T_{SPU(\gamma)}^{(b_1)} \geq T_{SPU(\gamma)}^{(b)}) + 1}{(B - 1) + 1}$$

for every γ and every b . Then, we will have $T_{aSPU}^{(b)} = \min_{\gamma \in \Gamma} P_{SPU(\gamma)}^{(b)}$, and the final p-value of aSPU test is:

$$P_{aSPU} = \sum_{b=1}^B \frac{I(T_{aSPU}^{(b)} \leq T_{aSPU}^{obs}) + 1}{B + 1}.$$

It is worth noting again that the same B simulated score vectors have been used in calculating the P_{aSPU} .

Aim 1(a) VII

Method: A new class of tests and a data-adaptive test in longitudinal data settings

"data-adaptive" genome wide scan strategy

In practice for genome wide scan purpose, we can use a "data-adaptive" aSPU test strategy that is: we first start with a smaller B , say $B = 1000$, to scan the genomes, then gradually increase B to say 10^6 for a few groups of SNVs, e.g. specific genes or windows, which pass an pre-determined significance cutoff (e.g. p-value $\leq 5/B$) in the previous step; repeat this process according to user's specific need until satisfying the significance level accuracy, e.g. a p-value of $\leq 10^{-7}$ requires $B \geq 10^7$. In this "data-adaptive" way of implementing the simulation based p-value calculating method for aSPU test, we will be able to apply the aSPU test to GWA data.

Aim 1(a) VIII

Method: A new class of tests and a data-adaptive test in longitudinal data settings

Other versions of aSPU test

- **aSPUw test**

The SPUw test is a *diagonal-variance-weighted* version of the SPU test, defined as:

$$T_{SPUw(\gamma)} = \sum_{j=1}^p \left(\frac{U_{.2,j}}{\sqrt{\hat{\Sigma}_{.2,ij}}} \right)^\gamma$$

Accordingly, **the aSPUw test** statistic is defined as

$$T_{aSPUw} = \min_{\gamma \in \Gamma} P_{SPUw(\gamma)}$$

where $P_{SPUw(\gamma)}$ is the p-value from $T_{SPUw(\gamma)}$. The procedures of getting these values are exactly the same as in above **aSPU** test based on simulation. Finally, aSPUw p-value can be get by:

$$P_{aSPUw} = \sum_{b=1}^B \frac{I(T_{aSPUw}^{(b)} \leq T_{aSPUw}^{obs}) + 1}{B + 1},$$

Aim 1(a) IX

Method: A new class of tests and a data-adaptive test in longitudinal data settings

- **aSPU(w).Score test**

Although the **GEE Score test** will lose power in some scenario of gene-based GWA analysis as mentioned before, it still has the unique advantage in some scenarios when the correlation structure among SNVs really matters. GEE Score test in the form of $T = U_{.2}' \Sigma_{.2}^{-1} U_{.2}$ will keep the covariance matrix in the denominator, which preserves the information of possible linkage disequilibrium among SNVs.

$$T_{aSPU.Score} = \min \left\{ \min_{\gamma \in \Gamma} P_{SPU(\gamma)}, P_{Score} \right\},$$

where P_{Score} is the p-value of the Score test. To calculate the p-value of the aSPU(w).Score test, it is just as simple as to include the Score test p-value along with the other $SPU(\gamma)$ p-values, select the minimum p-value among them to form the new statistic $T_{aSPU.Score}$, then use the same simulation algorithm as discussed earlier to get the the $P_{aSPU.Score}$.

Aim 1(a) I

Method in data simulation

Simulation of genotype data

- ① a latent vector $G_i = (G_{i1}, \dots, G_{ip})'$ was first drawn from a **multivariate Normal distribution** $N(0, R)$, where R had a AR(1) correlation structure with its (i, j) th element in terms of purely correlation $r_{ij} = \text{Corr}(G_{if}, G_{ig}) = \rho^{|f-g|}$ between any two latent components, G_{if} and G_{ig} for $f \neq g$. In our simulations we set $\rho = 0.8$;
- ② the latent vector G_i was dichotomized to yield a haplotype with each latent element G_{ij} dichotomized to 0 or 1 with probability $\text{Prob}(G_{ij} = 1) = \text{MAF}$ of j th SNP; the MAFs were randomly drawn from a uniform distribution: for causal SNPs the MAFs were set between 0.3 and 0.4; for null SNPs the MAFs were set between 0.1 and 0.5;
- ③ we combined two independent haplotypes to form the genotype $X_i = (X_{i1}, \dots, X_{ip})'$ for subject i . The haplotypes for different subject were generated independently.

Aim 1(a) II

Method in data simulation

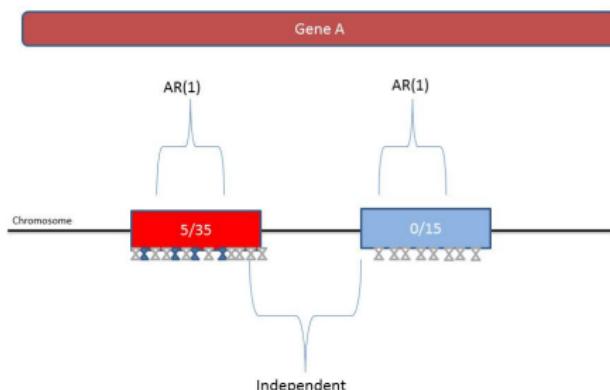


Figure: Demo graph of genotype simulation

in the test.

By this strategy we placed 35 SNPs in first block with AR(1) correlation structure to imitate the real LD structure among these SNPs; out of 35 SNPs we randomly set 5 SNPs to be causal (i.e. has a non-zero coefficient in later introduced regression model); to mimic the real data situation in SNP genotyping platforms, e.g. tag SNPs are usually in LD with casual SNPs but not the casual SNPs themselves, we excluded the 5 causal SNPs in the test (thus in first block, only null SNPs in LD with these 5 causal SNPs will enter the test). We further placed 15 null SNPs in the second block with AR(1) correlation structure as the same as we did in the first block. Note the first block and second block are independent though. All the SNPs from second block will participate

Aim 1(a) III

Method in data simulation

Simulation of phenotype data

We setup the mixed effect model to achieve the AR(1) correlation structure as:

$$y_{im} = \mu_i + b_i + \underbrace{\rho e_{i,m-1} + s_{i,m}}_{e_{i,m}}, \quad (3)$$

with $m = 1, \dots, k$ indexes the longitudinal measurements within subject i as already stated in ??; $\mu_i = Z_i\varphi + X_i\beta = H_i\theta$ as in quantitative trait case (see ??); b_i is the random intercept representing the subject-level random effect, and

$$\rho e_{i,m-1} + s_{i,m} = e_{i,m},$$

where ρ is lag-one autocorrelation coefficient, so we can plugin our estimate from real data here by setting up $\rho = 0.7$. We assume the following distribution:

$$b_i \sim N(0, \sigma_b^2)$$

$$e_{i,m} \sim N(0, \sigma_e^2)$$

$$s_{i,m} \sim N(0, (1 - \rho^2)\sigma_e^2)$$

Aim 1(a) IV

Method in data simulation

Under this assumption, the variance-covariance matrix across longitudinal measurements becomes (assuming $k = 4$ for the number of longitudinal measurements):

$$\Sigma_{4 \times 4} = \text{Var} \begin{pmatrix} b_i + e_{i1} \\ b_i + \rho e_{i1} + s_{i2} \\ b_i + \rho e_{i2} + s_{i3} \\ b_i + \rho e_{i3} + s_{i4} \end{pmatrix} = \sigma_b^2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \quad (4)$$

Aim 1(a) V

Method in data simulation

Connect phenotype data with genotype data

Let we first introduce the below splitting of the phenotype variance:

$$\text{Var}(y_{im}) = \text{Var}(X_{ij})\beta_j^2 + \sigma_{oth}^2 = 2f(1-f)\beta_j^2 + \sigma_{oth}^2 \quad (5)$$

where the Hard-Weinberg equilibrium is assumed to hold. f is the MAF of the SNP; σ_{oth}^2 is the residual variance after removing the effect of j th SNP. Obviously we can see σ_b^2 and σ_e^2 are contained in σ_{oth}^2 (see equation (3)), and if other SNPs' effect are negligible, we expect $\sigma_b^2 + \sigma_e^2 \approx \sigma_{oth}^2$. Now let we look at the relationship between genetic heritability (narrow-sense heritability) and equation (5):

$$h^2 = \frac{\text{Var}(A)}{\text{Var}(P)} \quad (6)$$

this is the classical formula of narrow-sense heritability, with $\text{Var}(A)$ represents the variance due to the additive effects of the alleles, and $\text{Var}(P)$ represents the total variance in the phenotype. In our situation for j th SNP, this can be extended to:

$$h_j^2 = \frac{\text{Var}_j(A)}{\text{Var}(P)} = \frac{\text{Var}(X_{ij})\beta_j^2}{\text{Var}(y_{im})} = \frac{\text{Var}(y_{im}) - \sigma_{oth}^2}{\text{Var}(y_{im})} \approx \frac{\text{Var}(y_{im}) - \sigma_b^2 - \sigma_e^2}{\text{Var}(y_{im})} \quad (7)$$

Aim 1(a) VI

Method in data simulation

Summary of parameter setup in simulation studies

After this point, by systematically solving the equations (5) and (7), we can easily calculate the β_j for j th SNP once we have determined the value of h_j^2 , σ_b^2 , σ_e^2 and f . Usually a h_j^2 for a single SNP j will not be high for complex disease and we used $h_j^2 = 0.001$ in our simulation study to control β_j , with other parameters set as: $\sigma_b^2 = 1$, $\sigma_e^2 = 1$ and $k = 4$ representing the number of longitudinal measurements for a single subject. Without special indication, we will use the simulated data set with 1000 replicates; significance level is set at 0.05.

Aim 1(a) I

Preliminary simulation results

- Tests under default simulation settings with varying sample size

n	Score	UminP	SumP	SumP.w	SSU	aSPU	aSPUw	aSPU.sco	aSPUw.sco
500	0.038	0.056	0.058	0.053	0.044	0.052	0.051	0.050	0.048
1000	0.047	0.054	0.048	0.049	0.065	0.065	0.064	0.059	0.057
2000	0.055	0.041	0.053	0.053	0.059	0.052	0.055	0.058	0.058
3000	0.055	0.054	0.057	0.060	0.065	0.063	0.054	0.056	0.059

Table: Type I error under using working independence R_w



Aim 1(a) II

Preliminary simulation results

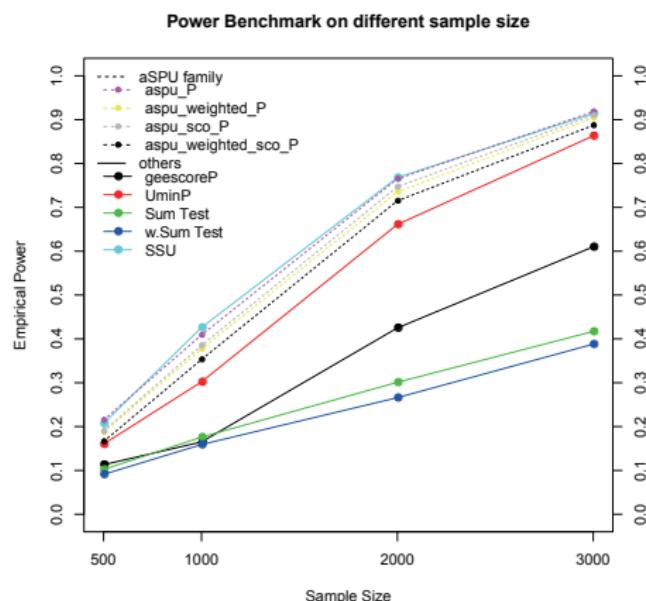


Figure: Empirical power benchmark under different n using working independence R_w

Aim 1(a) III

Preliminary simulation results

- **Tests with half number of SNPs in opposite effect direction**

In 5 causal SNPs, we set 2 of them to have opposite effect direction to the left 3 SNPs. The other settings kept the same as the above. We have the empirical power benchmark result as below:

Aim 1(a) IV

Preliminary simulation results

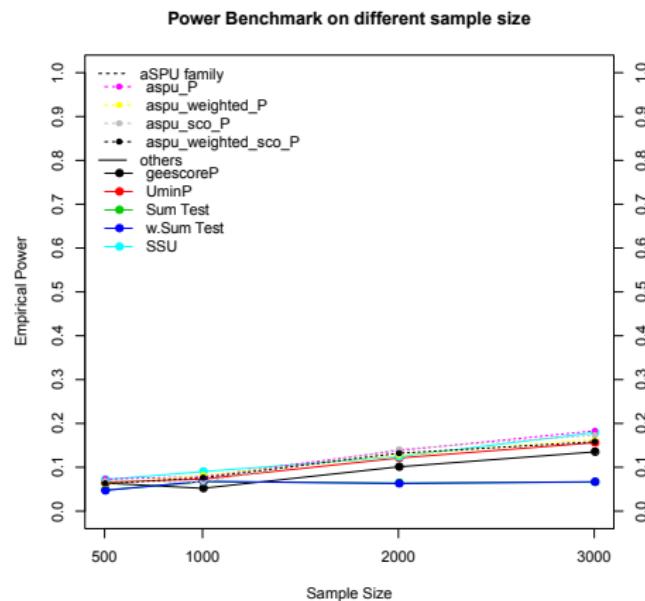


Figure: Empirical power benchmark under a mixed SNP effects

Aim 1(a) V

Preliminary simulation results

- **Tests with growing number of Null SNPs**

In previous two tests scenarios, we confirmed the ability of aSPU family members and concluded the SPU(2) is most powerful and contribute to the good performance of aSPU family. Now we are curious how higher γ will bring aSPU family to the edge. We gradually increased the number of null SNPs number from 50 to 75, 100, 200, then finally a seemingly extreme number 400. We used $n = 3000$ as the sample size. We kept all other settings the same with previous scenarios. The empirical power benchmark result is shown below:

Aim 1(a) VI

Preliminary simulation results

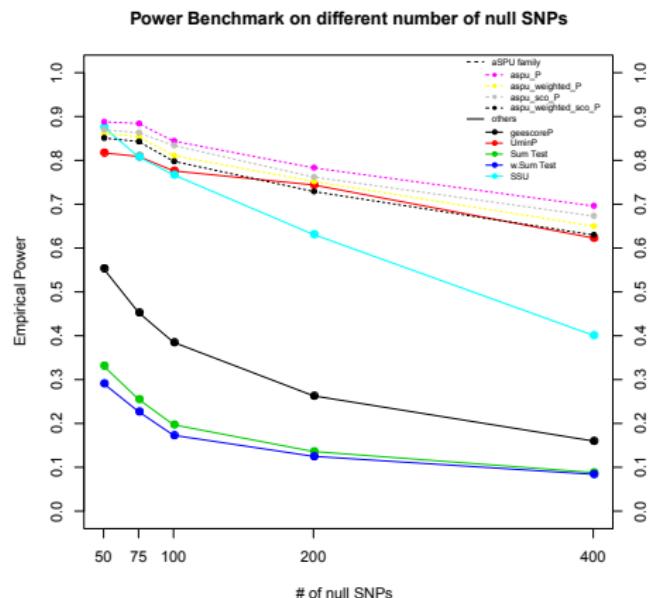


Figure: Empirical power benchmark under an increasing number of Null SNPs

Aim 1(a) VII

Preliminary simulation results

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

Aim 1(b)

Extend the data-adaptive longitudinal association test within GEE framework to work for rare variants in a gene-based manner. Since RVs has much lower MAF than CVs, some assumptions like coefficient estimator follows an asymptotic normal distribution may hold or not. Special procedure like permutation or parametric bootstrap specially designed for the longitudinal data settings should be adopted to provide an accurate association significance level.

Aim 1(b) I

Method

Previously we have:

$$U_{.2}^{(b)} \sim MVN \left(0, \hat{\Sigma}_{.2} \right)$$

with $b = 1, 2, \dots, B$, and thus obtain a statistics under null hypothesis: $T^{(b)} = \sum_{j=1}^p U_{.2,j}^{(b)\gamma}$. We then calculate the p-value of $T_{SPU(\gamma)}$ as $P_{SPU(\gamma)} = \sum_{b=1}^B \frac{I(T^{(b)} \geq T^{obs}) + 1}{B+1}$.

While MAF of RVs are usually low, e.g. between 0.001 to 0.01, the asymptotically Normal distribution of either beta coefficient or score vector may or may not hold. The above algorithms will hold in RV case by large, except that the $U_{.2}^{(b)}$ may not follow the multivariate Normal distribution any longer. As a remedy, we propose a permutation algorithm that generates the empirical null distribution of $U_{.2}^{(b)}$ and in the same time maintain the relationship between longitudinal traits and possible covariates such as age, gender, etc, for subject i . The algorithm is required to be also robust to missing data as this is a usual case in longitudinal data settings.



Aim 1(b) II

Method

The permutation algorithm can be implemented as follows:

- ① identify the max k across all n subjects, which is the number of longitudinal measurements, e.g. $k = 4$.
- ② detect if the data has missing values, if yes, fill the missing value with NA to complement the data dimension (for example, subject i with $Y_i = (y_{i,1}, \dots, y_{i,4})'$ has two missing measurements at time 2 and time 3. After missing value complementing, it becomes $Y_i = (y_{i,1}, NA, NA, y_{i,4})'$). Now we should have all the subjects with each Y_i of dimension equal to $k \times 1$.
- ③ complement H_i to be of full dimension, i.e. $k \times (p + q + 1)$, for covariates and SNVs. Now we should have $(Y_i \quad H_i)$ as an augmented matrix of dimension $k \times (p + q + 2)$ for each subject i , where $H_i = (Z_i, X_i)$. For total n subjects, we have row-wise binded matrix

$$M = \begin{pmatrix} Y_1 & H_1 \\ Y_2 & H_2 \\ \vdots & \vdots \\ Y_n & H_n \end{pmatrix}$$

of dimension $nk \times (p + q + 2)$.

Aim 1(b) III

Method

- ④ permute the SNV chunk among different individuals, i.e. the X_i in $(Y_i \quad Z_i, X_i)$ with the X_j in $(Y_j \quad Z_j, X_j)$, where $i \neq j$.
- ⑤ with permuted

$$M^{*(b)} = \begin{pmatrix} Y_1 & Z_1, X_1^{*(b)} \\ Y_2 & Z_1, X_2^{*(b)} \\ \vdots & \vdots \\ Y_n & Z_1, X_n^{*(b)} \end{pmatrix}$$

we refit the GEE model and get the $U_{.2}^{*(b)}$

- ⑥ repeat step 4 - 5 B times to produce $U_{.2}^{*(b)}$ with $b = 1, 2, \dots, B$.

After we get enough $U_{.2}^{*(b)}$ to form an empirical null distribution, the left work of aSPU test for RVs will be exactly the same as we did on CVs. The only difference is, previously we get simulation based null distribution of score vector under CVs situation, but now we rely on special permutation algorithm in the longitudinal data settings to generate the null distribution of score vector.

Aim 1(b) I

Methods in data simulation

The simulation strategy of RV data is almost the same with previous strategy for generating CV data , except that:

- ① the MAF of RVs, regardless of casual one or null one, are set between 0.001 and 0.01.
- ② the casual RVs are not excluded from later test as we expect the whole-genome sequencing or exome sequencing/Chip platform will identify high density SNVs including the real casual ones.

We will use the same simulated longitudinal phenotype data as for CVs.

Aim 1(b) I

Preliminary simulation results

- Simulation-based Test under default settings with varying sample size

n	pSSU	pSSUw	pScore	pSum	mvn.UminP	UminP	SPU(1)	SPUw(1)	SPU(2)	SPUw(2)	aSPU	aSPUw	aSPU.sco	aSPUw.sco
500	0.053	0.054	0.052	0.049	0.047	0.022	0.052	0.026	0.063	0.025	0.056	0.021	0.059	0.035
1000	0.055	0.040	0.042	0.048	0.054	0.049	0.048	0.046	0.061	0.044	0.045	0.045	0.053	0.047
2000	0.054	0.050	0.048	0.049	0.046	0.045	0.053	0.044	0.063	0.061	0.066	0.062	0.062	0.062
3000	0.045	0.044	0.039	0.060	0.053	0.055	0.057	0.058	0.058	0.052	0.049	0.055	0.055	0.057

Table: Empirical type I error using simulation-based method in RV analysis. mvn.UminP: UminP method based MVN distribution; UminP: UminP method based on simulation.

Aim 1(b) II

Preliminary simulation results

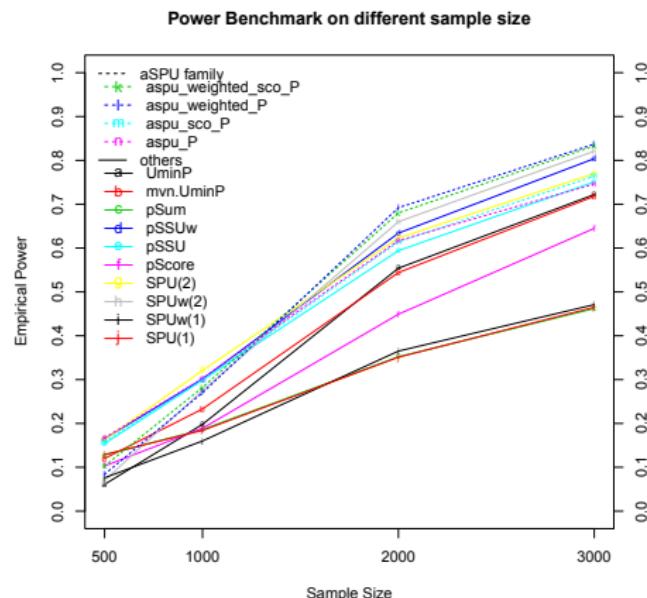


Figure: Empirical power benchmark using simulation-based method in RV analysis

Aim 1(b) III

Preliminary simulation results

- **Permutation-based Test under default settings with varying sample size**

As noted before, there are some minor issues in using simulated-based aSPU method to test RVs, we thus tested the aSPU performance based on permutation algorithm. The type I error is shown below.

n	pSSU	pSSUw	pScore	pSum	mvn.UminP	UminP	SPU(1)	SPUw(1)	SPU(2)	SPUw(2)	aSPU	aSPUw	aSPU.sco	aSPUw.sco
500	0.053	0.054	0.052	0.049	0.047	0.046	0.050	0.049	0.056	0.061	0.054	0.053	0.060	0.056
1000	0.055	0.040	0.042	0.048	0.054	0.056	0.048	0.049	0.056	0.043	0.047	0.045	0.052	0.051
2000	0.054	0.050	0.048	0.049	0.046	0.046	0.049	0.043	0.053	0.052	0.063	0.057	0.058	0.056
3000	0.045	0.044	0.039	0.060	0.053	0.050	0.058	0.058	0.047	0.048	0.049	0.053	0.049	0.053

Table: Empirical type I error using permutation-based method in RV analysis.

mvn.UminP: UminP method based MVN distribution; UminP: UminP method based on permutation.

Aim 1(b) IV

Preliminary simulation results

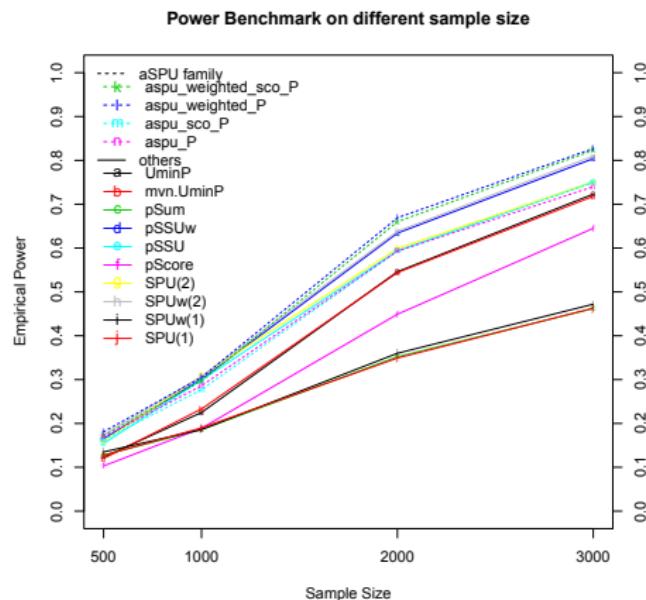


Figure: Empirical power benchmark using simulation-based method in RV analysis

Aim 1(b) V

Preliminary simulation results

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

Aim 2

To extend the data-adaptive longitudinal association test within the GEE framework to work for common variants or rare variants in a gene-set/pathway-based manner, i.e. pathway-based association test. Currently, there are no statistical models designed for pathway-based association test in longitudinal data settings, not to mention the data-adaptive property.

Aim 2 I

Method

A pathway analysis involves multiple genes (e.g. 20 as a typical number). As the genes within a pathway may contain different numbers of RVs, we need to modify the aSPU test to adjust for various gene length to avoid dominant influence from a large (or small) gene.

Suppose we let the short notation $U_{g\cdot}$ to represent $U_{g,2}$ for the RVs X_i ' part in the whole score vector, and $U_{g\cdot} = (U_{g,1}, U_{g,2}, \dots, U_{g,p_g})'$ is the score vector for gene g with p_g RVs of itself. Given a pathway (or gene set) S , the gene-specific SPU statistic is as follows:

$$T_{SPU(\gamma;g)} \propto \|U_{g\cdot}\|_\gamma = \left(\frac{\sum_{j=1}^{p_g} |U_{g,j}|^\gamma}{p_g} \right)^{\frac{1}{\gamma}} \quad (8)$$

Then accordingly, the pathway-based SPU statistic is

$$T_{Path-SPU(\gamma,\gamma_2;S)} = \sum_{g \in S} (T_{SPU(\gamma;g)})^{\gamma^2} \quad (9)$$

Note the $T_{SPU(\gamma;g)}$ is now standardized by gene-specific number of RVs, p_g ; for a given gene g , $T_{SPU(\gamma;g)}$ is equivalent to previous notation $T_{SPU(\gamma)}$ by large.

Aim 2 II

Method

For any given (γ, γ_2) , we recourse to same simulation or permutation strategy to calculate the p-value $P_{Path-SPU}(\gamma, \gamma_2; S)$ from $T_{Path-SPU}(\gamma, \gamma_2; S)$. Then we will have the **pathway-based aSPU** test statistic:

$$T_{Path-aSPU}(S) = \min_{\gamma, \gamma_2} P_{Path-SPU}(\gamma, \gamma_2; S) \quad (10)$$

we again adopt the same strategy as previous (which utilized the same simulated U in last step for calculating $P_{Path-SPU}(\gamma, \gamma_2; S)$) to calculate the final **pathway-based aSPU** p-value $P_{Path-aSPU}(S)$.

We propose to use $\gamma_2 \in \Gamma_2 = \{1, 2, 4, 8\}$. The 1, 2, 4, 8 will cover Sum-like test, SSU-like test, and two more tests preferring the sparse-casual-gene situation (e.g. only 2 or 3 genes are associated with traits in a pathway, say with 20 genes).

Aim 2 I

Methods in data simulation

- We will simulate a pathway with 20 genes; each gene g will contain p_g RVs with p_g randomly drawn from a uniform distribution $U(5, 30)$; 10 of the 20 genes will be randomly selected to be causal, with each causal gene containing 1 causal RV. The RVs within each gene will be simulated as before. The phenotype data in the simulation study will be the same as before.
- We will test Path-aSPU family on the simulated data to evaluate the type I error and power, with comparison to other existing tests like GRASS [CHP⁺10], which executes lasso regression (L1-norm) of eigenSNPs within each gene to achieve variable selection, while performing ridge regression (L2-norm) of eigenSNPs at the gene-set-level to achieve gene effect estimates shrinkage simultaneously; ALIGATOR [HGP⁺09], the association list go annotator, which is a 'p-value enrichment approach' requiring only pre-computed SNP p-values, uses Fisher's exact test on SNP with minimum p-value for the gene-level association; Plink [PNTB⁺07], which is a very popular GWAS analysis tool and plinkSet module within it implements the set-based associate test; the famous GSEA test in association study settings by [WLB07].
- We can further consider more extensive simulation tests, such as changing the causal RV number within a causal gene or changing the number of causal genes; use independent RVs within a gene instead of correlated RVs in AR(1); test Path-aSPU family with different working correlation matrix for modeling the longitudinal response.

Table of Contents

1 Background

- Introduction to GWAS
- SNP-set based association tests
- Longitudinal data analysis strategy in GWAS
- Gene-Set/Pathway based association tests in GWAS

2 Dissertation Aims

3 Public Health Significance

4 Specific Aims, Methods, and Preliminary Simulation Results

- Aim 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework
- Aim 1(b): Longitudinal aSPU family tests on Rare Variants
- Aim 2: Pathway-based longitudinal aSPU family tests: Path-aSPU
- Aim 3: Package/software development

5 Real Data Introduction

6 References

Aim 3

To provide an R package or Linux command-line based software program to enable convenient implementation of above methods. The package/software will be released to public (e.g. CRAN) eventually.

Aim 3

Method

- ① the package/software will be straightforward to install and use for 1st-time user
- ② the package/software will have the ability to run in a very flexible parallel computation framework, e.g. can use single node with multiple cores or use multiple nodes with multiple cores. The parallel protocol we will adopt is either SOCKET or MPI.
- ③ the package/software will have state-of-the-arts technique to enable efficient implementation of aSPU algorithms, such as hash table, radix sort, memory-efficient task send & collect among nodes, some intensive loops consider calling C++ code, etc.
- ④ the package/software will have a help document with demo examples

Real Data Introduction

The real data used in my dissertation will be obtained from the Atherosclerosis Risk in Communities (ARIC) Study (<https://www2.cscc.unc.edu/aric/>).

The Cohort Component of the ARIC study began in 1987. A total of 15,792 participants received an extensive examination, including medical, social, and demographic data. These participants were re-examined every three years with the first screen (baseline) occurring in 1987-89, the second in 1990-92, the third in 1993-95, and the fourth exam was in 1996-98. In 2009, the NHLBI funded a fifth exam, which is currently being conducted.

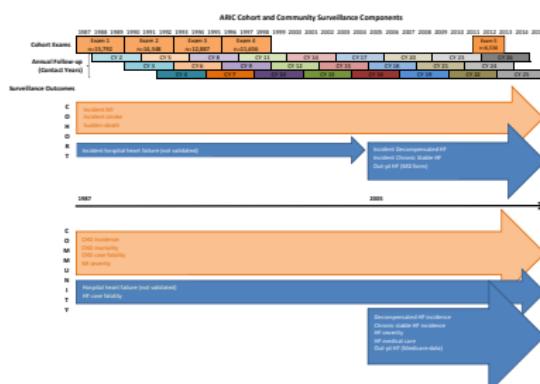


Figure: ARIC Cohort and Community Surveillance Components. Figure adopted from the ARIC website

We will apply our proposed method on ARIC data. Specifically, we will use the four closely cardiovascular-disease-related traits measured in ARIC cohort data, which are total cholesterol (tch), High-density lipoprotein (HDL), Low-density lipoprotein (LDL) and triglycerides (trgs). We will exclusively use Caucasian samples ($n = 11478$). For the covariates, we will include but not limited to subject's demographic information such as age, gender, BMI, etc.

References I

-  William S Bush and Jason H Moore, *Genome-wide association studies*, PLoS computational biology **8** (2012), no. 12, e1002822.
-  Global Lipids Genetics Consortium et al., *Discovery and refinement of loci associated with lipid levels*, Nature genetics (2013).
-  Lin S Chen, Li Hsu, Eric R Gamazon, Nancy J Cox, and Dan L Nicolae, *An exponential combination procedure for set-based association tests in sequencing studies*, The American Journal of Human Genetics **91** (2012), no. 6, 977–986.
-  Lin S Chen, Carolyn M Hutter, John D Potter, Yan Liu, Ross L Prentice, Ulrike Peters, and Li Hsu, *Insights into colon cancer etiology via a regularized approach to gene set analysis of gwas data*, The American Journal of Human Genetics **86** (2010), no. 6, 860–871.

References II

-  Andriy Derkach, Jerry F Lawless, and Lei Sun, *Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests*, *Genetic epidemiology* **37** (2013), no. 1, 110–121.
-  Ruzong Fan, Yifan Wang, James L Mills, Alexander F Wilson, Joan E Bailey-Wilson, and Momiao Xiong, *Functional linear models for association analysis of quantitative traits*, *Genetic epidemiology* **37** (2013), no. 7, 726–742.
-  Peter Holmans, Elaine K Green, Jaspreet Singh Pahwa, Manuel AR Ferreira, Shaun M Purcell, Pamela Sklar, Michael J Owen, Michael C O'Donovan, and Nick Craddock, *Gene ontology analysis of gwa study data sets provides insights into the biology of bipolar disorder*, *The American Journal of Human Genetics* **85** (2009), no. 1, 13–24.

References III

-  Fang Han and Wei Pan, *A data-adaptive sum test for disease association with multiple common or rare variants*, Human heredity **70** (2010), no. 1, 42–54.
-  Sunkyung Kim, Wei Pan, and Xiaotong Shen, *Penalized regression approaches to testing for quantitative trait-rare variant association*, Frontiers in genetics **5** (2014).
-  Li Luo, Eric Boerwinkle, and Momiao Xiong, *Association studies for next-generation sequencing*, Genome research **21** (2011), no. 7, 1099–1108.

References IV

-  Seunggeun Lee, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, N. H. L. B. I G. O Exome Sequencing Project-E. S. P Lung Project Team , David C. Christiani, Mark M. Wurfel, and Xihong Lin, *Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.*, Am J Hum Genet **91** (2012), no. 2, 224–237 (eng).
-  Leslie A Lange, Youna Hu, He Zhang, Chenyi Xue, Ellen M Schmidt, Zheng-Zheng Tang, Chris Bizon, Ethan M Lange, Joshua D Smith, Emily H Turner, et al., *Whole-exome sequencing identifies rare and low-frequency coding variants associated with ldl cholesterol*, The American Journal of Human Genetics **94** (2014), no. 2, 233–245.

References V

-  DY Lin, *An efficient monte carlo approach to assessing statistical significance in genomic studies*, Bioinformatics **21** (2005), no. 6, 781–787.
-  Bingshan Li and Suzanne M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.*, Am J Hum Genet **83** (2008), no. 3, 311–321 (eng).
-  Dan-Yu Lin and Zheng-Zheng Tang, *A general framework for detecting disease associations with rare variants in sequencing studies*, The American Journal of Human Genetics **89** (2011), no. 3, 354–367.
-  Seungeun Lee, Michael C. Wu, and Xihong Lin, *Optimal tests for rare variant effects in sequencing association studies.*, Biostatistics **13** (2012), no. 4, 762–775 (eng).

References VI

-  Kung-Yee Liang and Scott L Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika **73** (1986), no. 1, 13–22.
-  Li Luo, Yun Zhu, and Momiao Xiong, *Quantitative trait locus analysis for next-generation sequencing with the functional linear models*, Journal of medical genetics **49** (2012), no. 8, 513–524.
-  _____, *Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation*, European Journal of Human Genetics **21** (2012), no. 2, 217–224.
-  Bo Eskerod Madsen and Sharon R. Browning, *A groupwise association test for rare mutations using a weighted sum statistic.*, PLoS Genet **5** (2009), no. 2, e1000384 (eng).

References VII

-  Taylor J Maxwell, Christie M Ballantyne, James M Cheverud, Cameron S Guild, Chiadi E Ndumele, and Eric Boerwinkle, *Apoe modulates the correlation between triglycerides, cholesterol, and chd through pleiotropy, and gene-by-gene interactions*, *Genetics* **195** (2013), no. 4, 1397–1405.
-  Benjamin M Neale, Manuel A Rivas, Benjamin F Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M Purcell, Kathryn Roeder, and Mark J Daly, *Testing for an unusual distribution of rare variants*, *PLoS genetics* **7** (2011), no. 3, e1001322.

References VIII

-  Gina M Peloso, Paul L Auer, Joshua C Bis, Arend Voorman, Alanna C Morrison, Nathan O Stitziel, Jennifer A Brody, Sumeet A Khetarpal, Jacy R Crosby, Myriam Fornage, et al., *Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks*, The American Journal of Human Genetics **94** (2014), no. 2, 223–232.
-  Wei Pan, *On the robust variance estimator in generalised estimating equations*, Biometrika **88** (2001), no. 3, 901–906.
-  _____, *Asymptotic tests of association with multiple snps in linkage disequilibrium*, Genetic epidemiology **33** (2009), no. 6, 497–507.
-  Wei Pan, Junghi Kim, Yiwei Zhang, Xiaotong Shen, and Peng Wei, *A powerful and adaptive association test for rare variants*, Genetics (2014), genetics–114.

References IX

-  Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al., *Plink: a tool set for whole-genome association and population-based linkage analyses*, The American Journal of Human Genetics **81** (2007), no. 3, 559–575.
-  Wei Pan and Xiaotong Shen, *Adaptive tests for association analysis of rare variants.*, Genet Epidemiol **35** (2011), no. 5, 381–388 (eng).
-  SR Seaman and B Müller-Myhsok, *Rapid simulation of p values for product methods and multiple-testing adjustment in association studies*, The American Journal of Human Genetics **76** (2005), no. 3, 399–408.

References X

-  Jianping Sun, Yingye Zheng, and Li Hsu, *A unified mixed-effects model for rare-variant association in sequencing studies*, Genetic epidemiology **37** (2013), no. 4, 334–344.
-  Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al., *Biological, clinical and population relevance of 95 loci for blood lipids*, Nature **466** (2010), no. 7307, 707–713.
-  Kai Wang, Mingyao Li, and Maja Bucan, *Pathway-based approaches for analysis of genomewide association studies*, The American Journal of Human Genetics **81** (2007), no. 6, 1278–1283.

References XI

-  Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin, *Rare-variant association testing for sequencing data with the sequence kernel association test.*, Am J Hum Genet **89** (2011), no. 1, 82–93 (eng).
-  Kai Wang, Mingyao Li, and Hakon Hakonarson, *Analysing biological pathways in genome-wide association studies.*, Nat Rev Genet **11** (2010), no. 12, 843–854 (eng).
-  Zhiyuan Xu, Xiaotong Shen, Wei Pan, Alzheimer's Disease Neuroimaging Initiative, et al., *Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes*, PloS one **9** (2014), no. 8, e102312.

References XII

-  Hua Zhou, Mary E Sehl, Janet S Sinsheimer, and Kenneth Lange, *Association screening of common and rare genetic variants by penalized regression*, Bioinformatics **26** (2010), no. 19, 2375.
-  Yiwei Zhang, Zhiyuan Xu, Xiaotong Shen, and Wei Pan, *Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data*, NeuroImage **96** (2014), 309–325.