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

Yang Yang
M.S, Informatics
Ph.D Candidate, Biostatistics

Division of Biostatistics, School of Public Health, The University of Texas

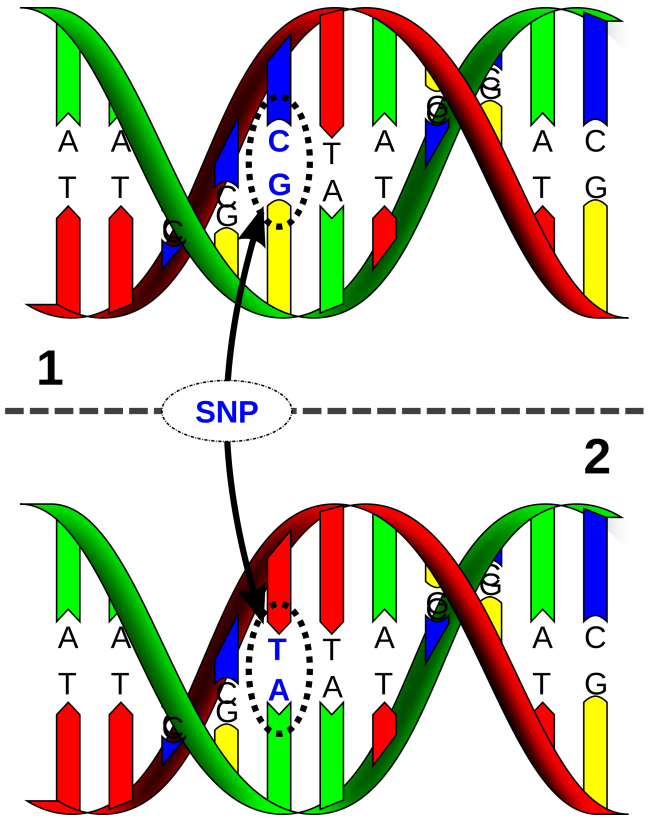
Jan.05 2015

Table of Contents

Table of Contents

Introduction to Genome-wide association study (GWAS)

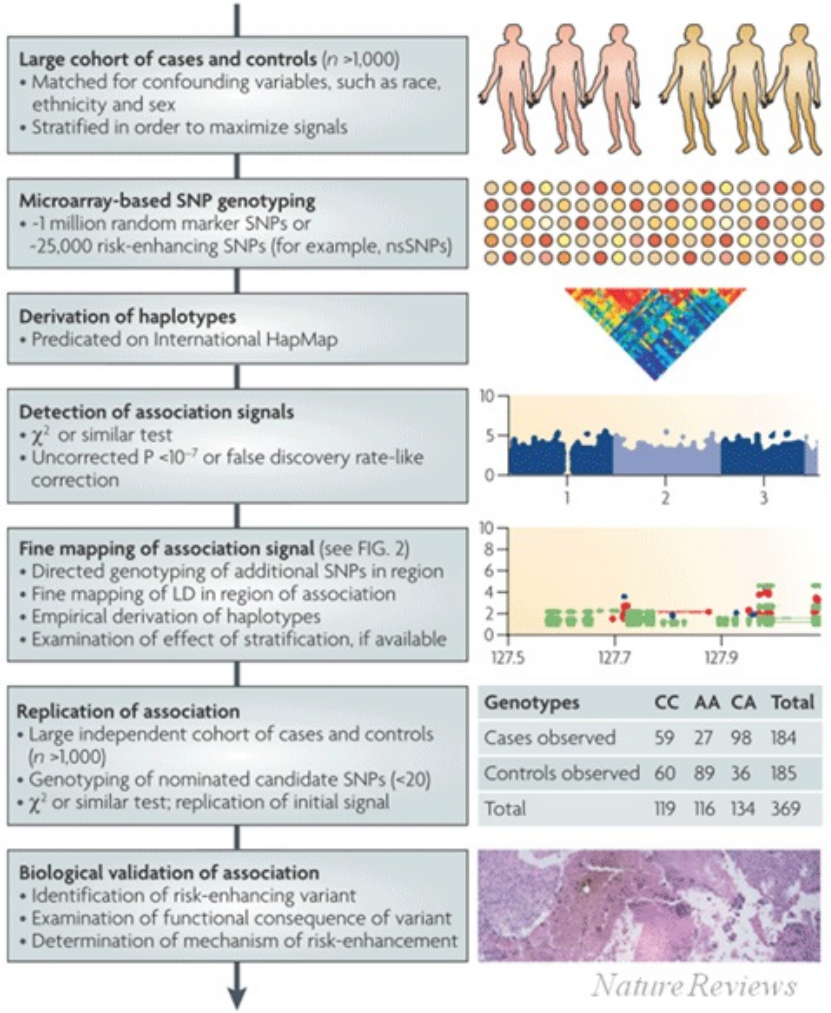
What is SNP?



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 flowchart of GWAS



Introduction to GWAS

How does GWAS result look like?

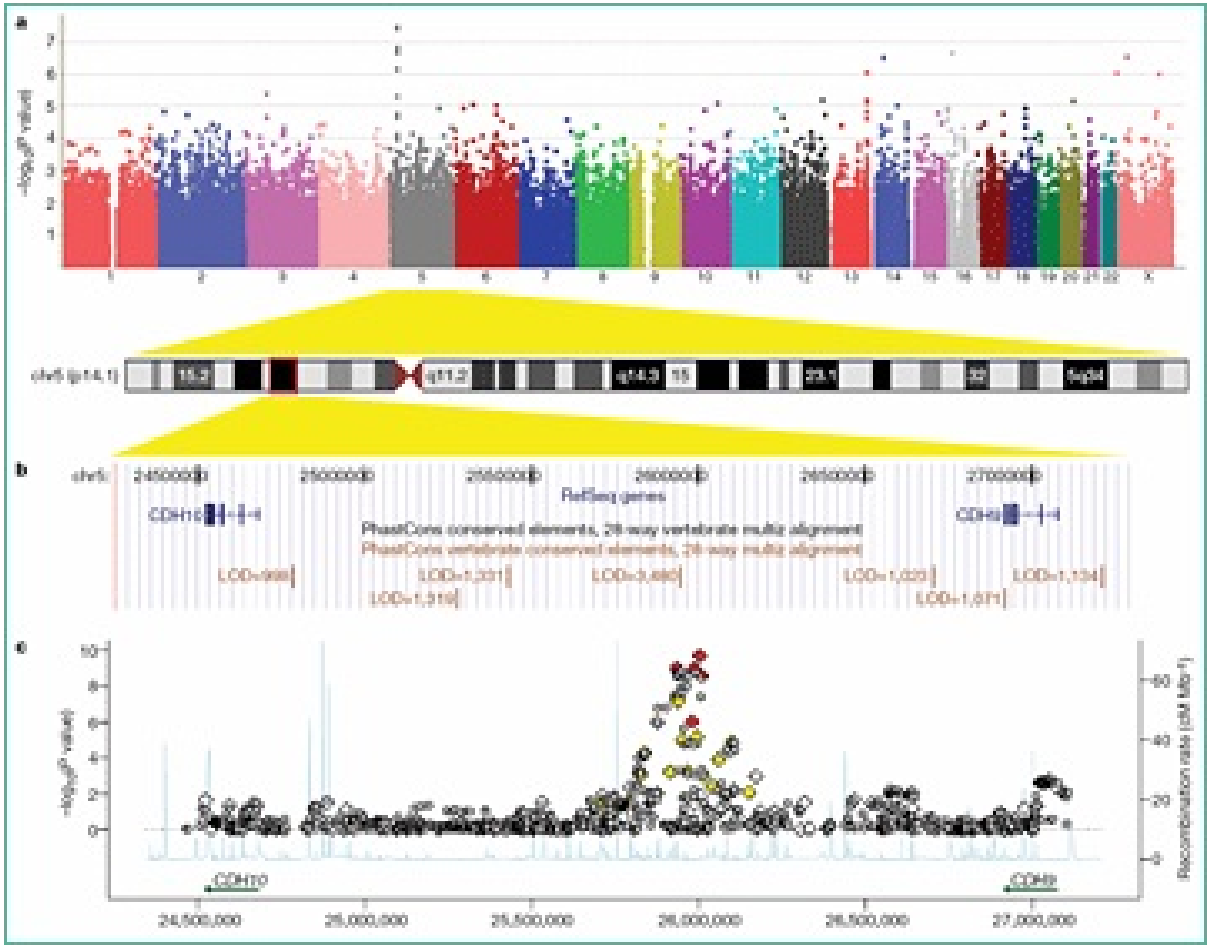


Figure: Common genetic variants on 5p14.1 associate with autism spectrum disorders [WZM+09]

Introduction to GWAS

GWAS Catalog

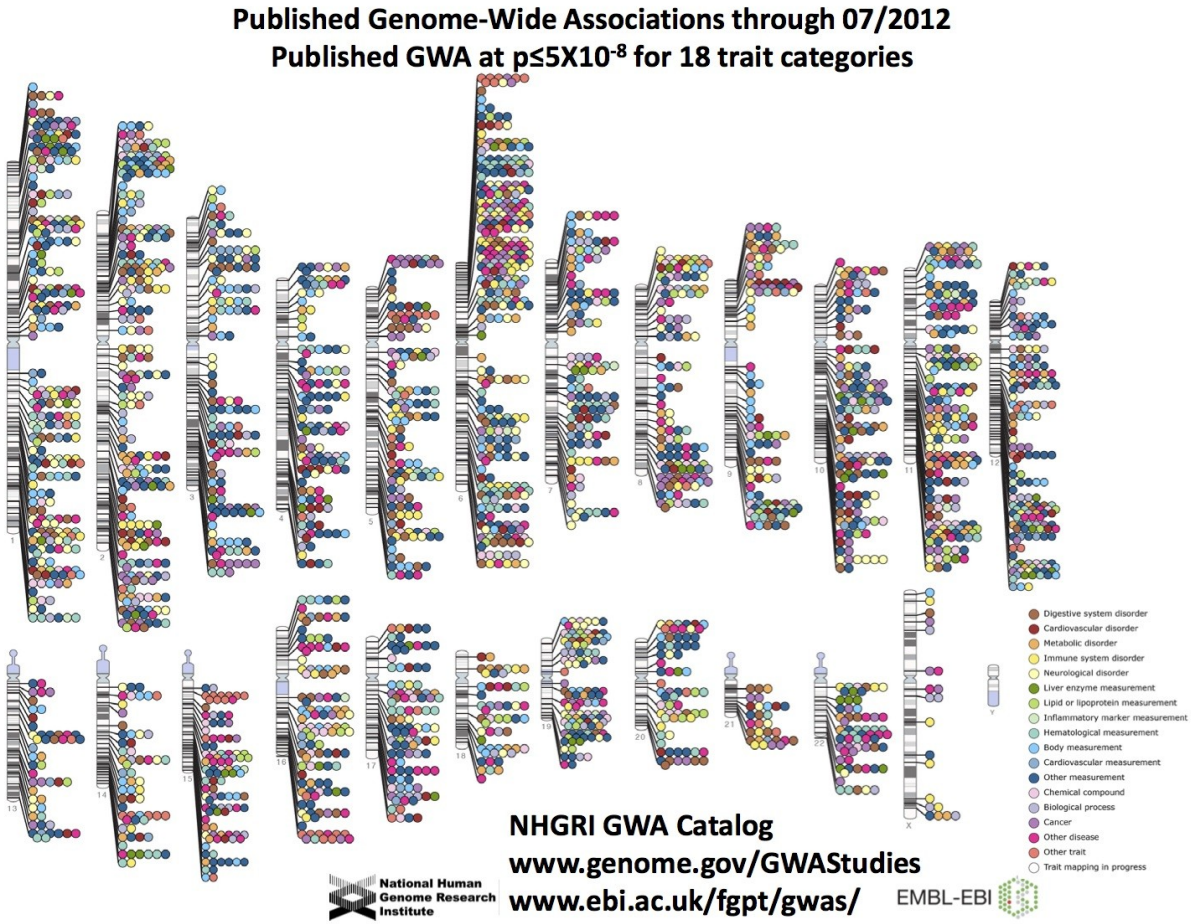


Figure: Published GWAS results for 18 trait categories

Introduction to GWAS

GWAS contribute to personalized medicine

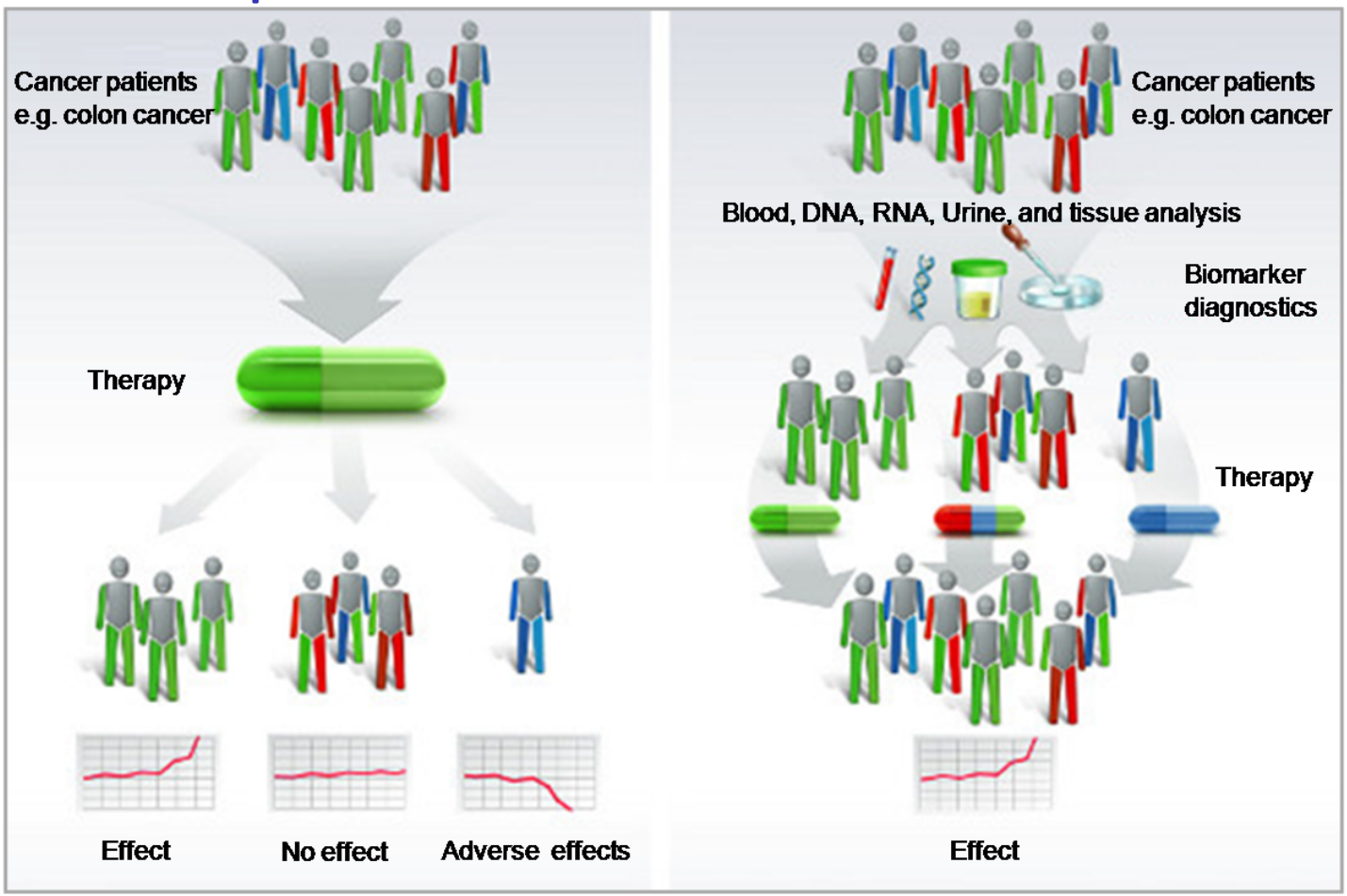


Figure: GWAS contribute to personalized medicine

Introduction to GWAS

Common variants and rare variants

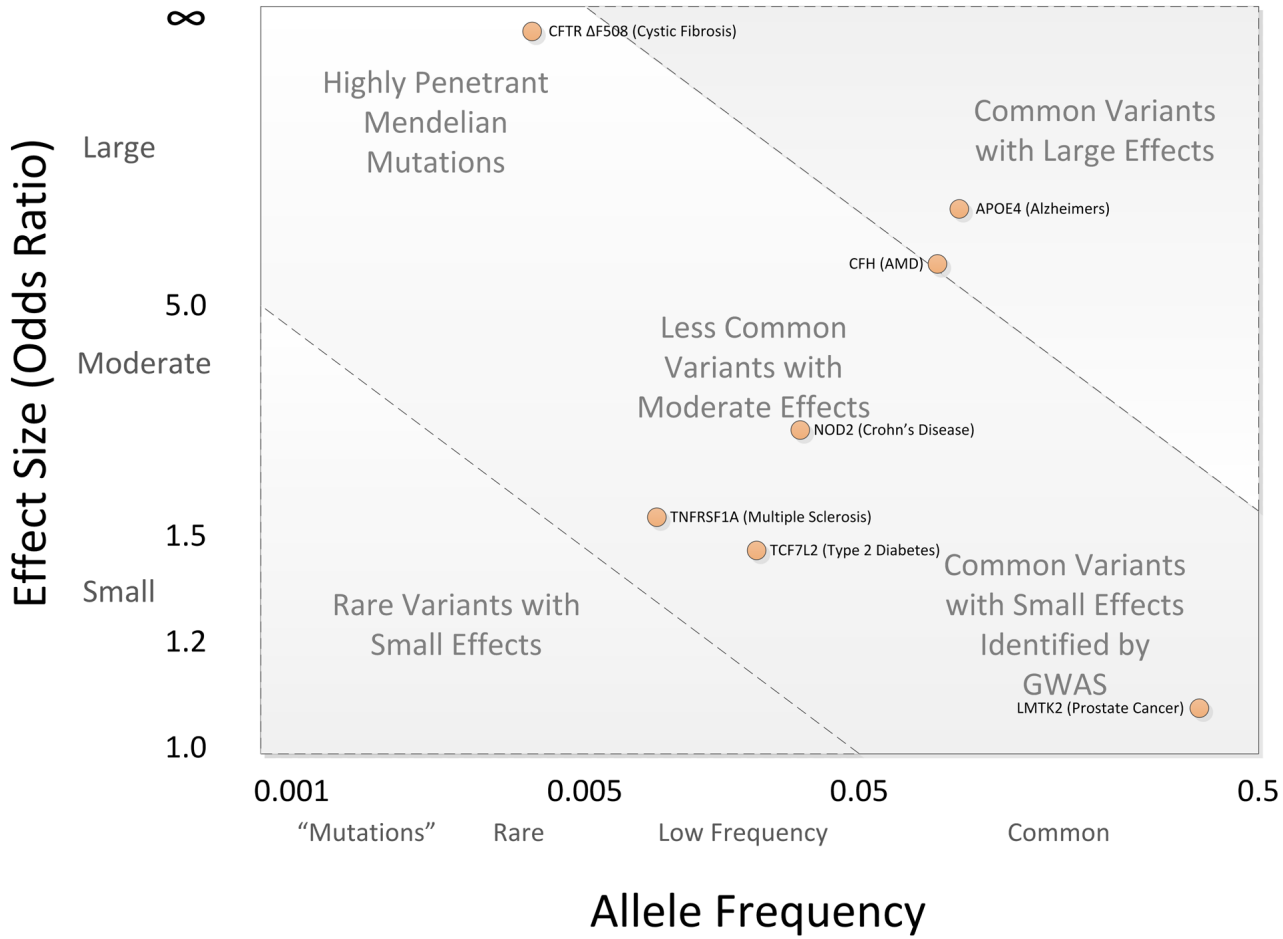


Figure: effect size of Single Nucleotide Variant [BM12]

Table of Contents

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,

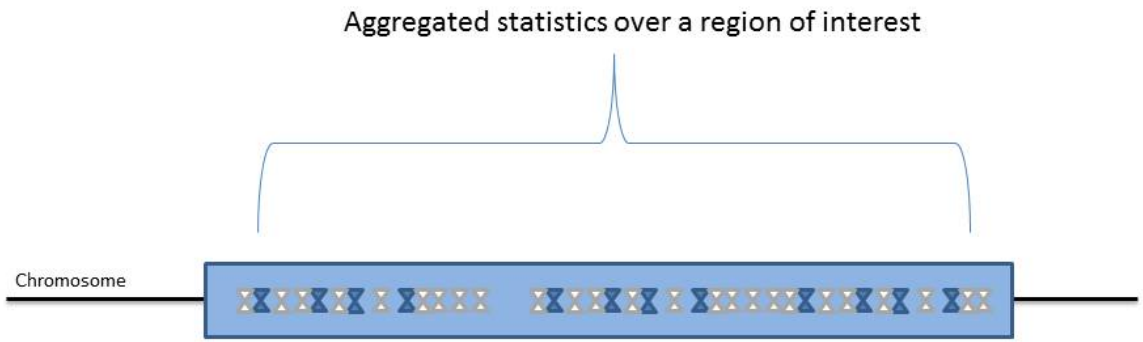
$$g(\mu_i) = \beta_0 + x_{ij}\beta_j + z_i\varphi_i,$$

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.



SNP-set based association tests II

A brief review

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].
- 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

How do longitudinal data look like?

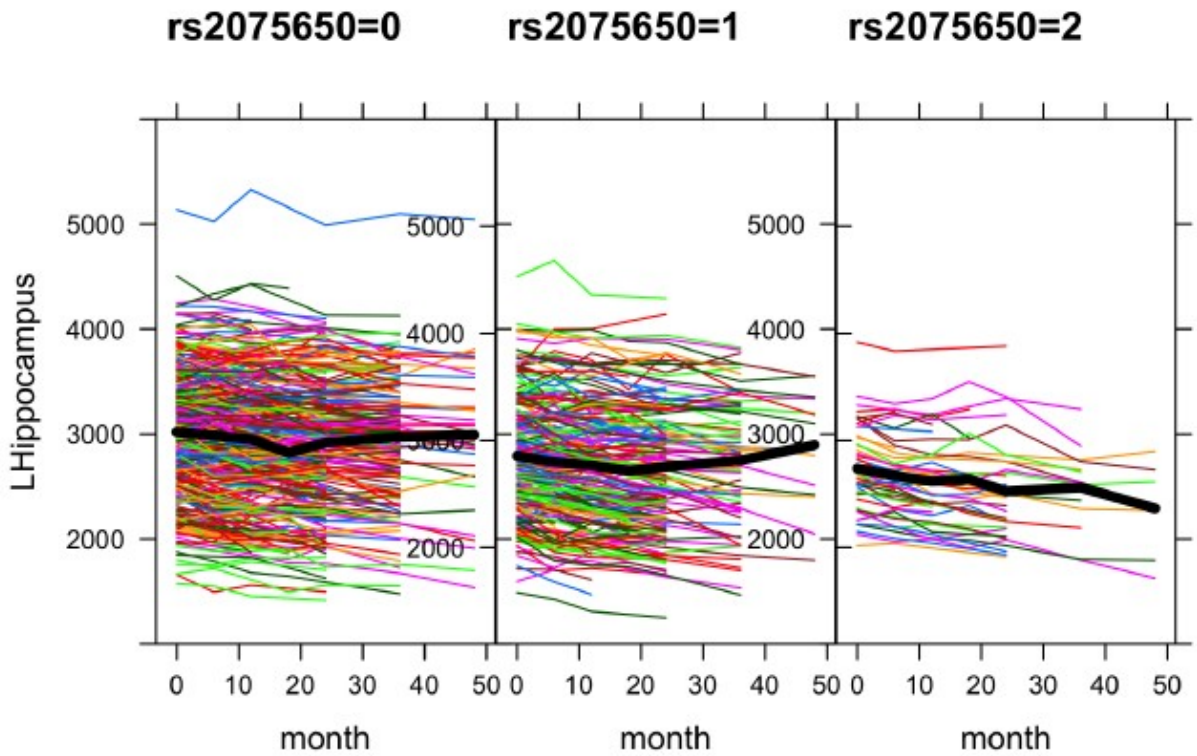


Figure: Trajectories of phenotype left hippocampus volume over time (in months) in three allele groups of SNP rs2075650 [XSP$^+$14]

Why longitudinal? I

In a cross-sectional study ($n_i = 1$) we are restricted to the model

$$Y_{i1} = \beta_C x_{i1} + \epsilon_{i1}, \quad i = 1, \ldots, m,$$

where β_C represents the difference in average Y across two sub-populations (samples) which differ by one unit in x. With repeated measurements, the above linear model can be extended to

$$Y_{ij} = \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1}) + \epsilon_{ij}, \quad i = 1, \ldots, m;\ j = 1, \ldots, n_i$$

[WDL+90].
Based on above formula, we can more obviously explain the merits of longitudinal studies over cross-sectional studies.

1. Longitudinal studies allow us to estimate both the cross-sectional difference (β_C) and the rate change over time (β_L).
2. Even when $\beta_C = \beta_L$, longitudinal studies tend to be more powerful than cross-sectional studies. This is due to the fact that in longitudinal studies, each person can be thought of serving as his/her own control.

Why longitudinal? II

3. Another merit of the longitudinal study is its ability to distinguish the between-subject variation and within-subject variation.
4. With longitudinal studies, we can estimate a person's current and future outcome (behavior trend).

Why longitudinal? III

Longitudinal study in GWAS

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.

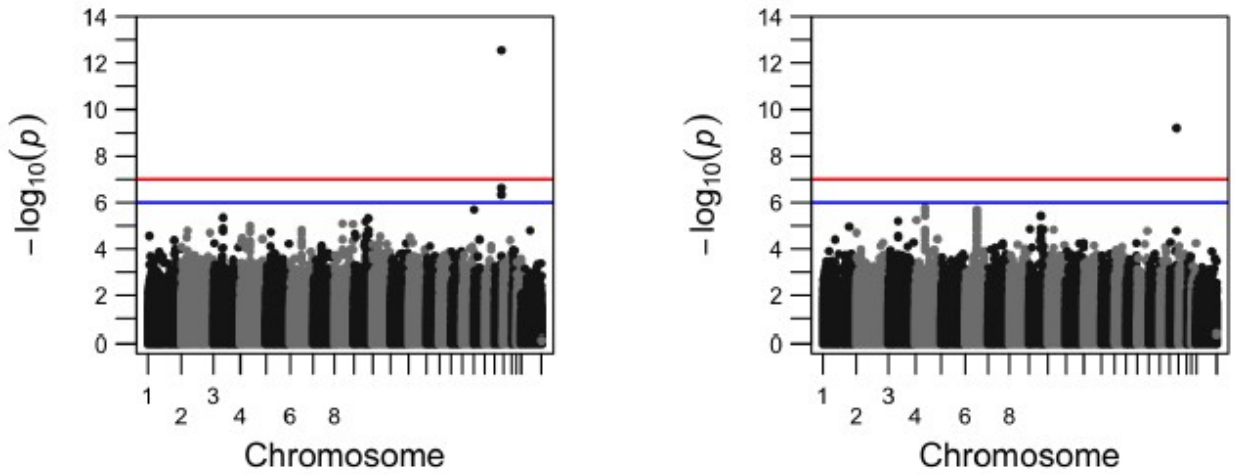


Figure: Comparison of the Manhattan plots for genome-wide p-values for phenotype left hippocampus volume from longitudinal analysis (left) and from cross-sectional analysis (right) [XSP+14]

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 [LW82].
- 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, marginal effect models by Generalized Estimating Equation (GEE) give population-averaged (PA) estimates.
- 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 $y_{ij-1}, \ldots, y_{ij-q}$ from history response vector: $H_{ij} = \{y_{ik}, k = 1, \ldots, j-1\}$ and covariates x_{ij}. The integer q is referred as the order of the Markov models.

Table of Contents

Research Aims

- Aim 1: Data-adaptive SNP-set-based association tests (aSPU) for longitudinal data analysis within GEE framework for **Common Variants**;
- Aim 2: Longitudinal aSPU family tests on **Rare Variants**

Table of Contents

Aim 1

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.

Table of Contents

Aim 1 I

Methods: introduction to notation and formula

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

$$y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})'$$

with y_{im} as a element, p SNPs of interest as a row vector

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

with x_{ij} coded as 0,1 or 2 for the count of the minor allele, and

$$z_i = (z_{i1}, z_{i2}, \ldots, z_{iq})$$

as a row vector for q variates.

Thus, we have:

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

X_i is a $k \times p$ matrix, and Z_i is a $k \times (q+1)$ matrix.

Aim 1 II

Methods: introduction to notation and formula

We then have the GLM equation as,

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

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

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

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

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

Aim 1 III

Methods: introduction to notation and formula

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:

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

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

Aim 1 IV

Methods: 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:

$$U = \sum_i (Z_i, X_i)' \, R_w^{-1} (Y_i - \mu_i)$$

$$\widetilde{\Sigma} = \sum_i (Z_i, X_i)' \, R_w^{-1} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' R_w^{-1} \, (Z_i, X_i) \tag{2}$$

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:

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

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

Aim 1 V

Methods: 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:

$$U = \sum_{i=1}(\frac{\partial \mu_i}{\partial \theta'})' V_i^{-1}(Y_i - \mu_i)$$

$$= \sum_{i=1}(\frac{\partial \mu_i}{\partial \theta'})' \phi A_i^{-\frac{1}{2}} R_w^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i)$$

and

$$\widetilde{\Sigma} = \sum_i \left(\frac{\partial \mu_i}{\partial \theta'}\right)' \phi A_i^{-\frac{1}{2}} R_w^{-1} A_i^{-\frac{1}{2}}(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' \phi A_i^{-\frac{1}{2}} R_w^{-1} A_i^{-\frac{1}{2}} \left(\frac{\partial \mu_i}{\partial \theta'}\right)$$

$$= \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

Aim 1 VI

Methods: 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

$$H_o : \beta = (\beta_1, \beta_2, \ldots, \beta_p)' = 0$$

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

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

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

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

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

Aim 1 VII

Methods: 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 $T \sim \chi^2_p$. 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:** $T = U'_{.2}\Sigma^{-1}_{.2}U^{-1}_{.2}$, where $U_{.2}$ and $\Sigma_{.2}$ are discussed above; the statistic is asymptotically equivalent to the Wald test with the same null distribution $T \sim \chi^2_p$. 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].
- **The UminP Test:** $T = \max_j \frac{U^2_{.2,j}}{\Sigma_{.2,jj}}$ for $j \in 1, 2, \ldots, p$, of jth SNP effect. The $\Sigma_{.2,jj}$ is the jth entry on the diagonal of $\Sigma_{.2}$. With max 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]. An asymptotic multivariate normal distribution numerical integration based method provided an alternative to calculate its p-value [PHS09, Pan09].

Aim 1 I

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

A general form of score-vector-based statistic can be generalized as:

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

where $W = (W_1, \ldots, W_p)'$ is a vector of weights for the p SNVs [LT11].
with special cases:

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

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

Aim 1 II

Methods: 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 $\gamma = 1, 2, \ldots, \infty$, leading to the sum of powered score (U) tests called **SPU** tests:

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

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

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

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

Aim 1 III

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

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

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 Equation 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 $b = 1, 2, \ldots, B$, 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 IV

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

The aSPU test

Although we have a list of SPU(γ) 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(γ) test(s).

We hereby propose an adaptive SPU (aSPU) test to achieve such purpose. Accordingly, we will have the aSPU test statistic:

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

Aim 1 V

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

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

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

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

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

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

Aim 1 VI

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

The "data-adaptive" genome wide scan strategy

In practice for genome wide scan purpose, we can use a "data-adaptive" aSPU test strategy that is:

1. we first start with a smaller B, say $B = 1000$
2. we increase B to say 10^6 for just a few groups of SNVs, which passed an pre-determined significance cutoff (e.g. p-value $\leq 5/B$) in 1
3. repeat 2 until a pre-determined B number reached

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 VII

Methods: 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:

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

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

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

Table of Contents

Aim 1 I

Methods in data simulation

Simulation of genotype data

1. a latent vector $G_i = (G_{i1}, \ldots, G_{ip})'$ 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 $r_{ij} = \mathrm{Corr}(G_{if}, G_{ig}) = \rho^{|f-g|}$ 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 $\mathrm{Prob}(G_{ij} = 1) =$ MAF of jth 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 $X_i = (X_{i1}, \ldots, X_{ip})'$ for subject i. The haplotypes for different subject were generated independently.

Aim 1 II

Methods in data simulation

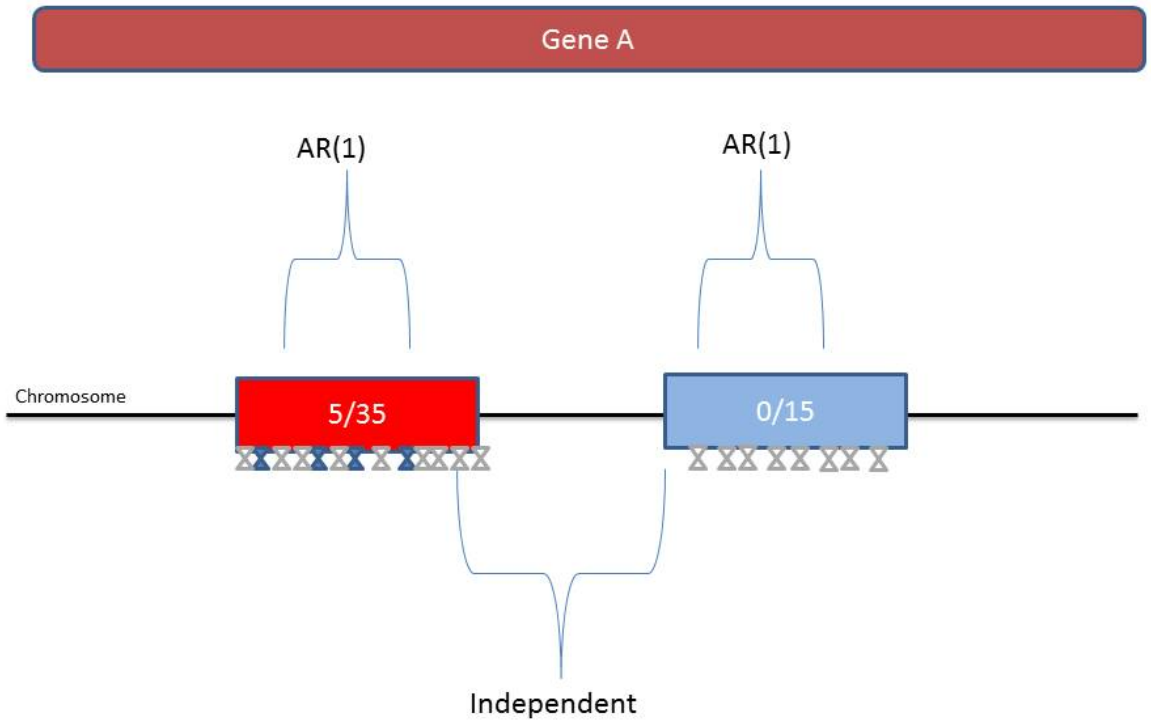


Figure: Demo graph of genotype simulation

Aim 1 III

Methods in data simulation

Simulation of phenotype data

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

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

with $m = 1, \ldots, k$ indexes the longitudinal measurements within subject i;

$$\mu_i = Z_i\varphi + X_i\beta = H_i\theta$$

as in quantitative trait case; 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:

Aim 1 IV

Methods in data simulation

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

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

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

Aim 1 V

Methods in data simulation

Connect phenotype data with genotype data
Let we first introduce the below splitting of the phenotype variance:

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

Now let we look at the relationship between genetic heritability (narrow-sense heritability) and equation (5):

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

In our situation for jth SNP, this can be extended to:

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

Aim 1 VI

Methods 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 jth 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. We summarize the parameters used in simulation studies here:

- $h_j^2 = 0.001$
- $\sigma_b^2 = 1$
- $\sigma_e^2 = 1$
- n varies between 500 and 3000
- $k = 4$
- 1000 replicates of simulated dataset
- $\alpha = 0.05$
- $\rho_y = 0.7$
- $\rho_x = 0.8$
- $R = AR(1)$
- $Rw = I$

Table of Contents

Aim 1 I

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 II

Simulation results

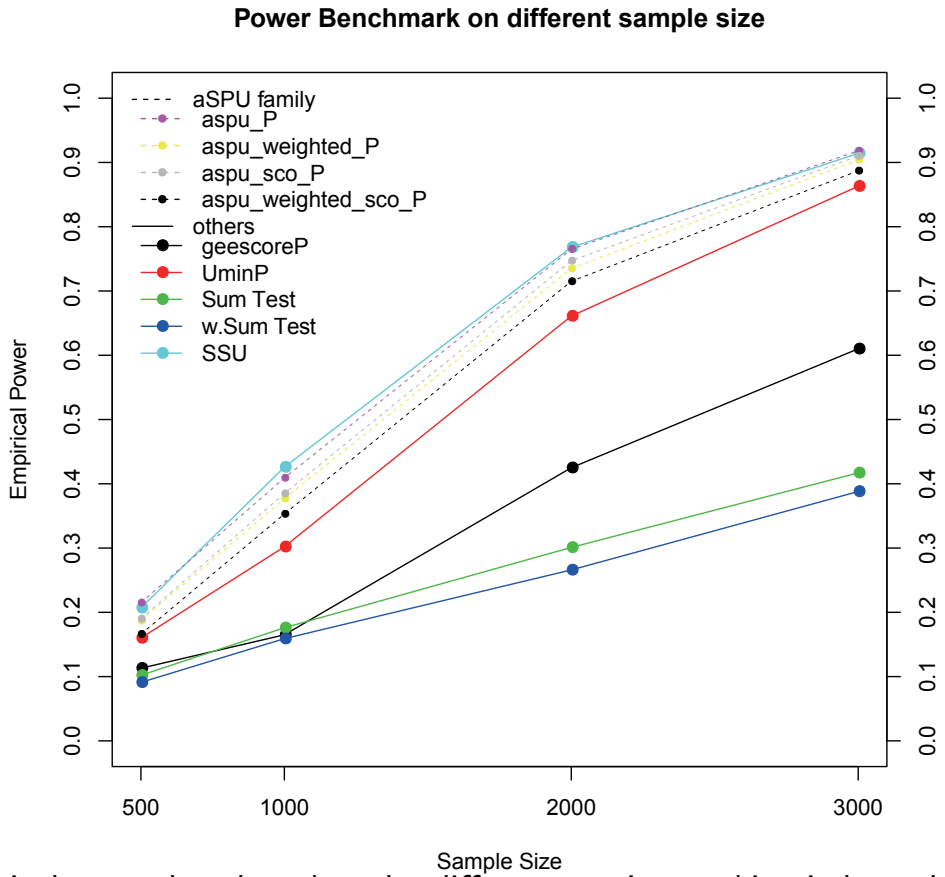


Figure: Empirical power benchmark under different n using working independence R_w

Aim 1 III

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 IV

Simulation results

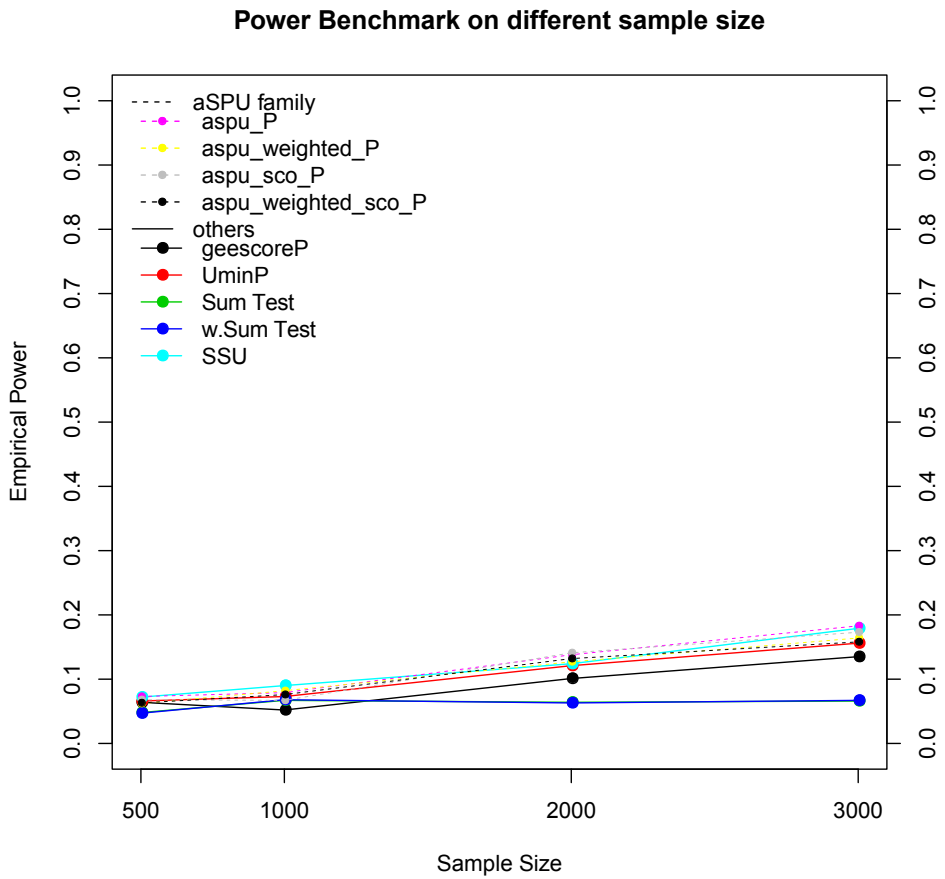


Figure: Empirical power benchmark under a mixed SNP effects

Aim 1 V

Simulation results

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

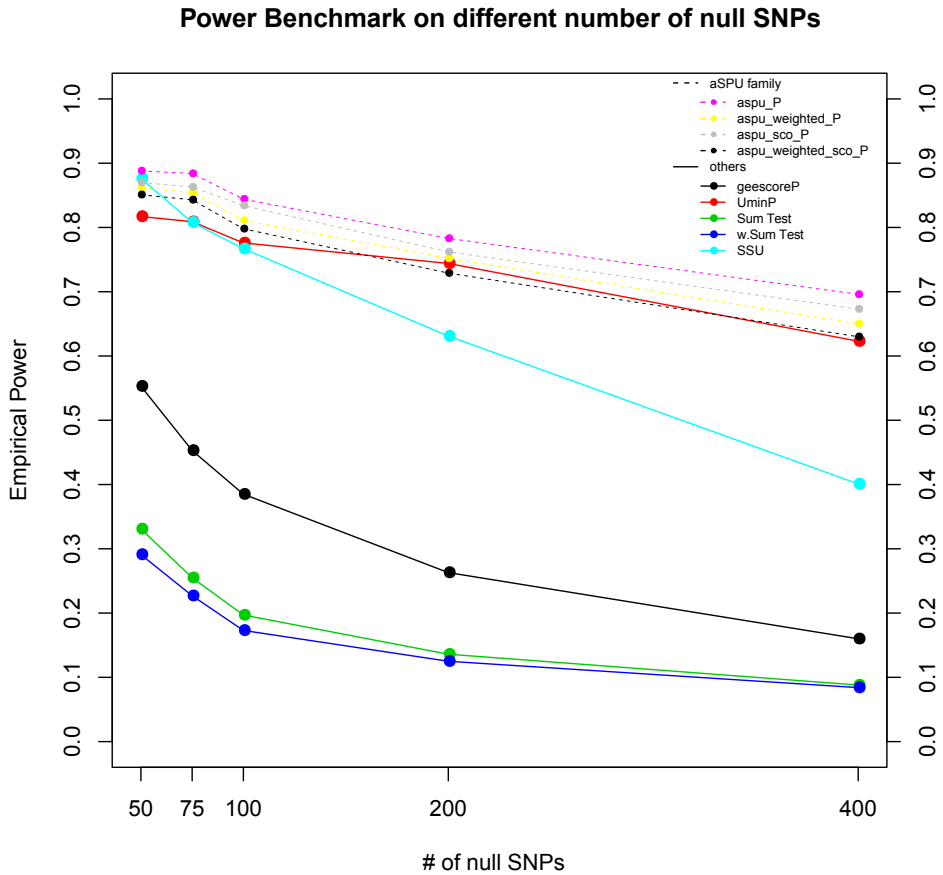


Figure: Empirical power benchmark under an increasing number of Null SNPs

Table of Contents

Aim 2

Extend the data-adaptive longitudinal association test within GEE framework to work for **rare variants** in a gene-based manner.

Table of Contents

Aim 2 I

Methods

For CVs we have:

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

with $b = 1, 2, \ldots, B$, 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 $P_{SPU(\gamma)} = \sum_{b=1}^{B} \frac{I(T^{(b)} \geq T^{obs})+1}{B+1}$.

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 will also be robust to **missing data** as this is a usual case in longitudinal data settings.

Aim 2 II

Methods

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 $Y_i = (y_{i,1},,,y_{i,4})'$ has two missing measurements at time 2 and time 3. After missing value complementing, it becomes $Y_i = (y_{i,1}, \mathrm{NA}, \mathrm{NA}, y_{i,4})'$). 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 $\begin{pmatrix} Y_i & H_i \end{pmatrix}$ 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 2 III

Methods

4. permute the SNV chunk among different individuals, i.e. the X_i in $\begin{pmatrix} Y_i & Z_i, X_i \end{pmatrix}$ with the X_j in $\begin{pmatrix} Y_j & Z_j, X_j \end{pmatrix}$, 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 $b = 1, 2, \ldots, B$.

Table of Contents

Aim 2 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.

Table of Contents

Aim 2 I

Simulation results

If we still use the CVs' strategy on RVs, we will have

- **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 2 II

Simulation results

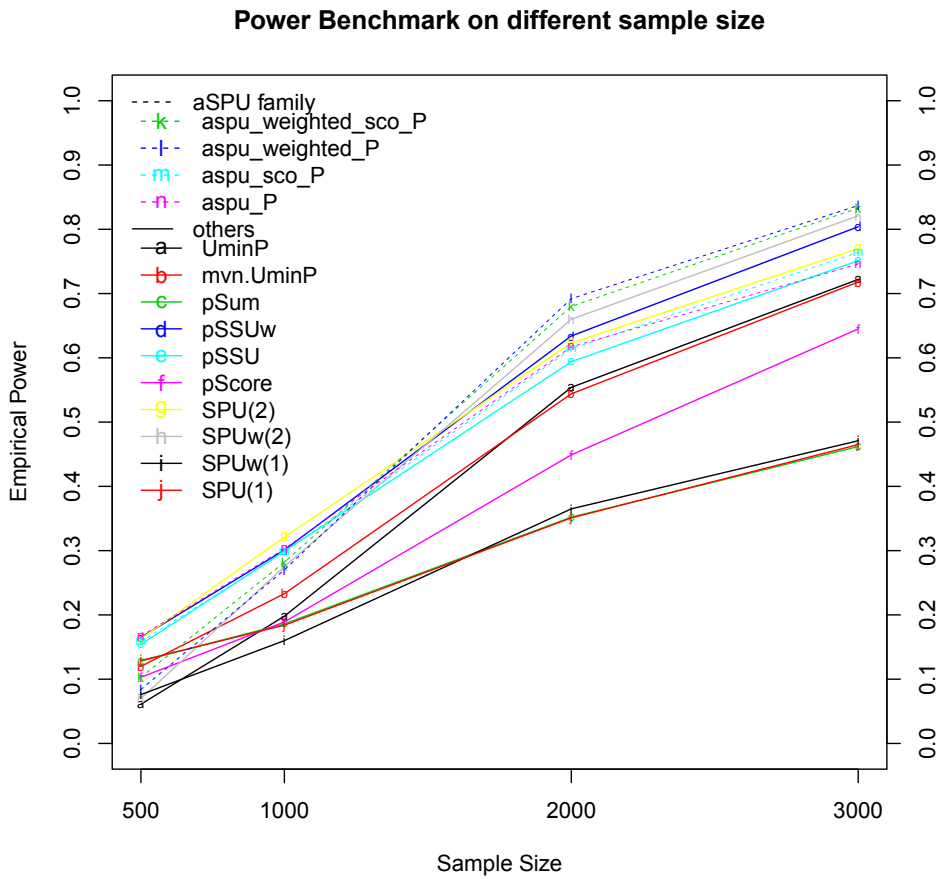


Figure: Empirical power benchmark using simulation-based method in RV analysis

Aim 2 III

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 2 IV

Simulation results

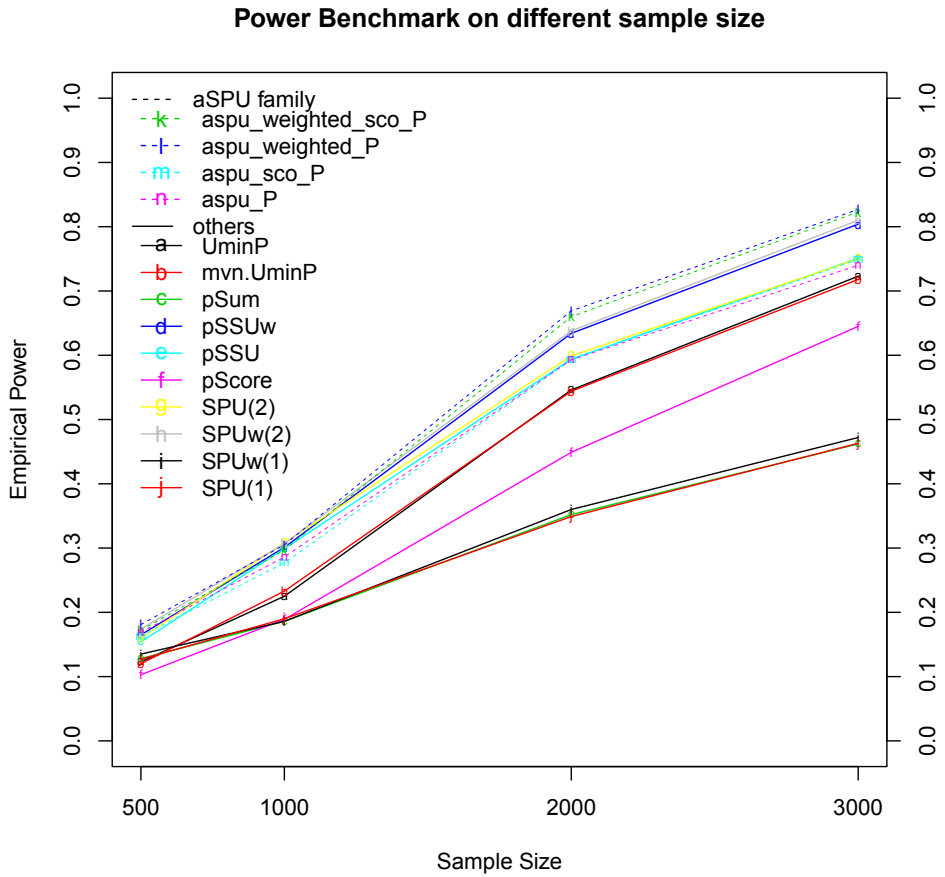


Figure: Empirical power benchmark using simulation-based method in RV analysis

Aim 2 V

Simulation results

- An effort to combine the advantages from aSPU, aSPUw and score test. The aSPU.aSPUw.Score test can save user's effort in deploying a best version of aSPU family test on a specific dataset with only a small amount of power loss in the process of compromising among different versions.

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

Aim 2 VI

Simulation results

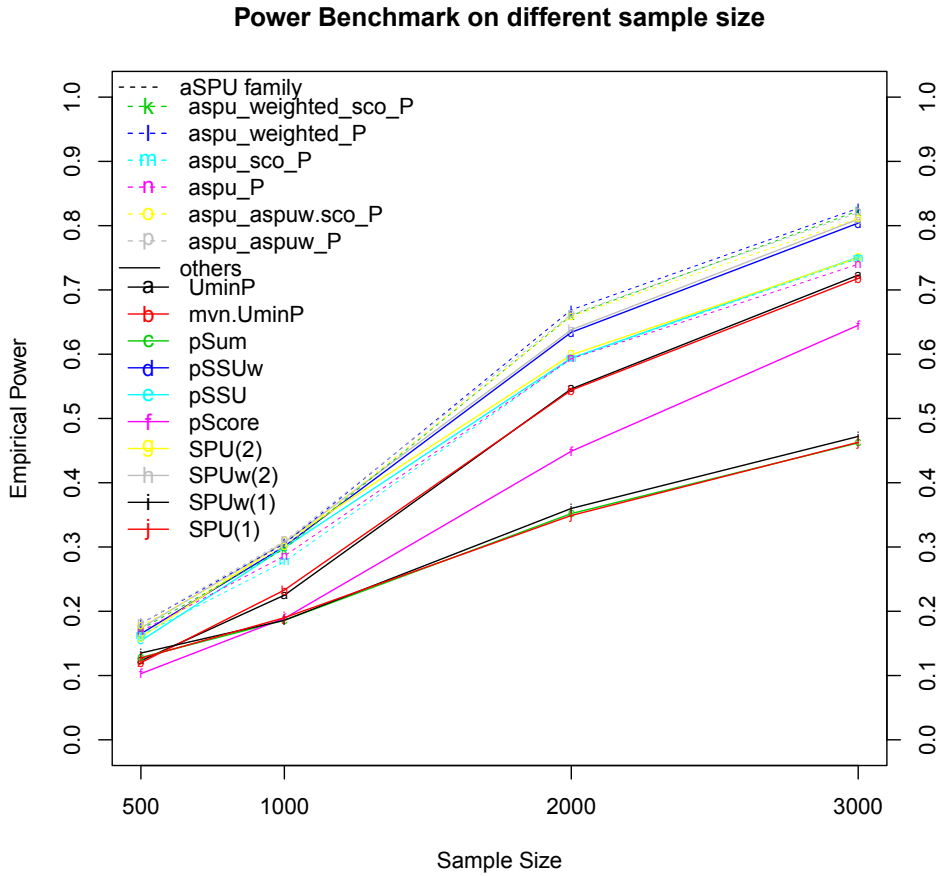


Figure: Empirical power benchmark with aSPU.aSPUw.Score test in RV analysis

Table of Contents

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.

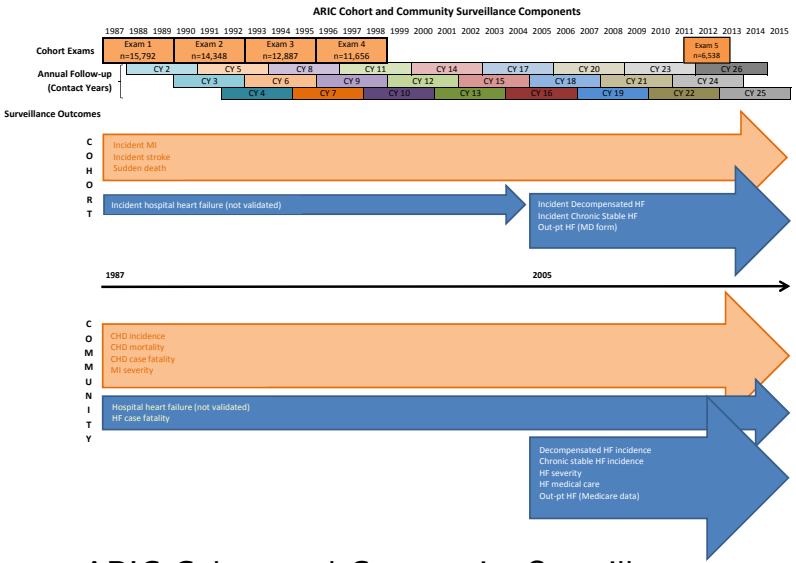


Figure: ARIC Cohort and Community Surveillance Components. Figure adopted from the ARIC website

We applied our novel 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.

Real Data Result Demo

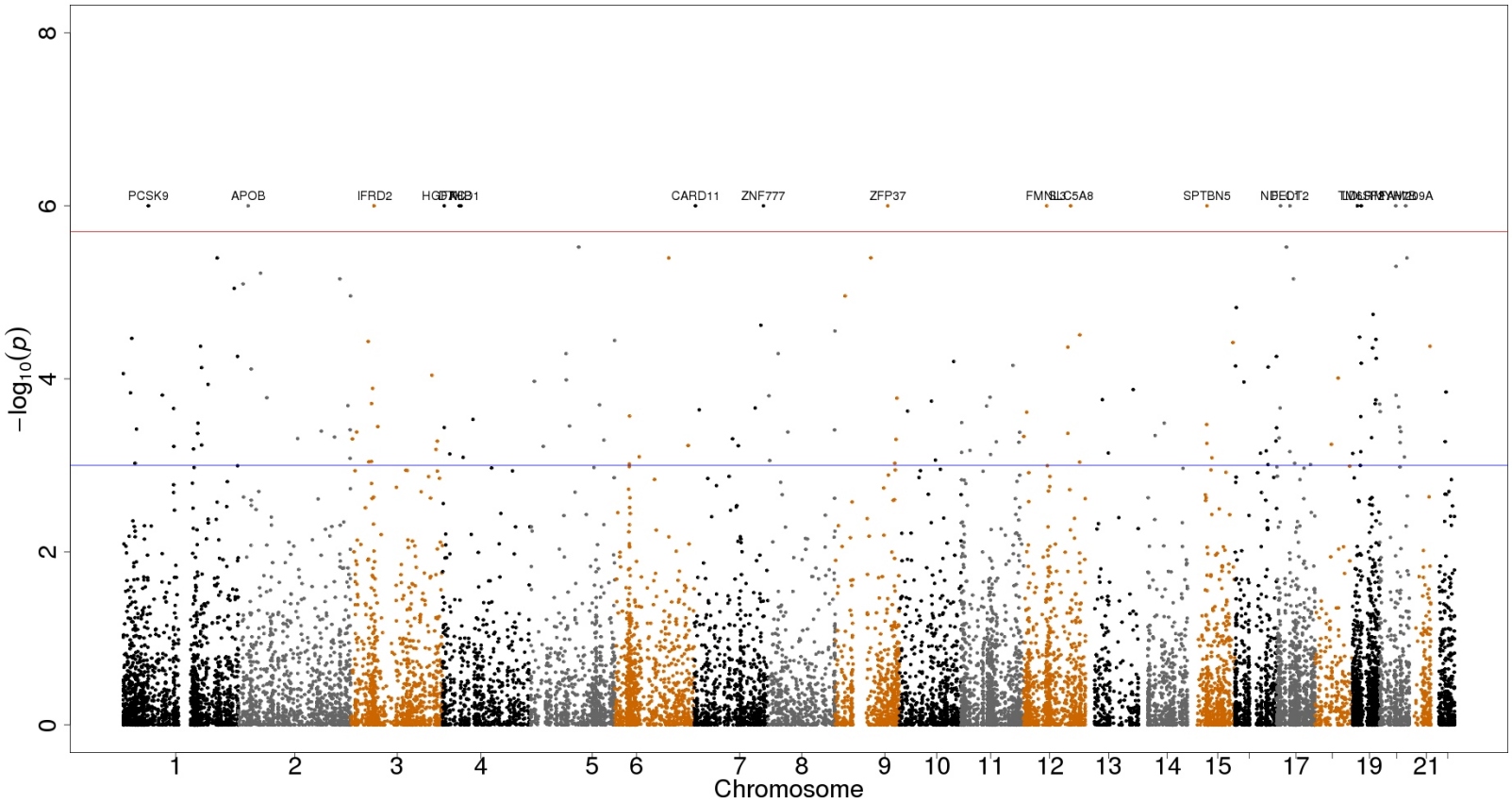


Figure: Manhattan Plot of aSPUw.score test on ARIC data Total Cholesterol trait

Table of Contents

Acknowledgement

Advisers:

- Peng Wei, Ph.D, Associate Professor, Division of Biostatistics, School of Public Health, University of Texas
- Wei Pan, Ph.D, Professor, Division of Biostatistics, School of Public Health, University of Minnesota

Supporting Grant:
Title: Association Analysis of Rare Variants with Sequencing Data
Funding Source: NIH/NHLBI (1R01HL116720)
Total cost: $\$1,043,901$

Thanks to Audience

Thank you for your participation!

References I


William S Bush and Jason H Moore, *Genome-wide association studies*, PLoS computational biology **8** (2012), no. 12, e1002822.


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.


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.


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.


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).


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).

References II


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.


Nan M Laird and James H Ware, *Random-effects models for longitudinal data*, Biometrics (1982), 963–974.


Seunggeun 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).


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).


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.


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.

References III


Wei Pan, Fang Han, and Xiaotong Shen, *Test selection with application to detecting disease association with multiple snps*, Human heredity **69** (2009), no. 2, 120–130.


Wei Pan, Junghi Kim, Yiwei Zhang, Xiaotong Shen, and Peng Wei, *A powerful and adaptive association test for rare variants*, Genetics (2014), genetics–114.


Wei Pan and Xiaotong Shen, *Adaptive tests for association analysis of rare variants.*, Genet Epidemiol **35** (2011), no. 5, 381–388 (eng).


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.


JAMES H WARE, DOUGLAS W DOCKERY, THOMAS A LOUIS, XIPING XU, BENJAMIN G FERRIS, and FRANK E SPEIZER, *Longitudinal and cross-sectional estimates of pulmonary function decline in never-smoking adults*, American journal of epidemiology **132** (1990), no. 4, 685–700.


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, Haitao Zhang, Deqiong Ma, Maja Bucan, Joseph T Glessner, Brett S Abrahams, Daria Salyakina, Marcin Imielinski, Jonathan P Bradfield, Patrick MA Sleiman, et al., *Common genetic variants on 5p14. 1 associate with autism spectrum disorders*, Nature **459** (2009), no. 7246, 528–533.


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 IV


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.