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

Doctoral Dissertation Defense

Yang Yang
B.S, Biological Science
M.S, Bioinformatics
Ph.D. Candidate, Biostatistics

Department of Biostatistics, School of Public Health, The University of Texas Health Science Center at Houston

November 30th 2015

Dissertation Committee:

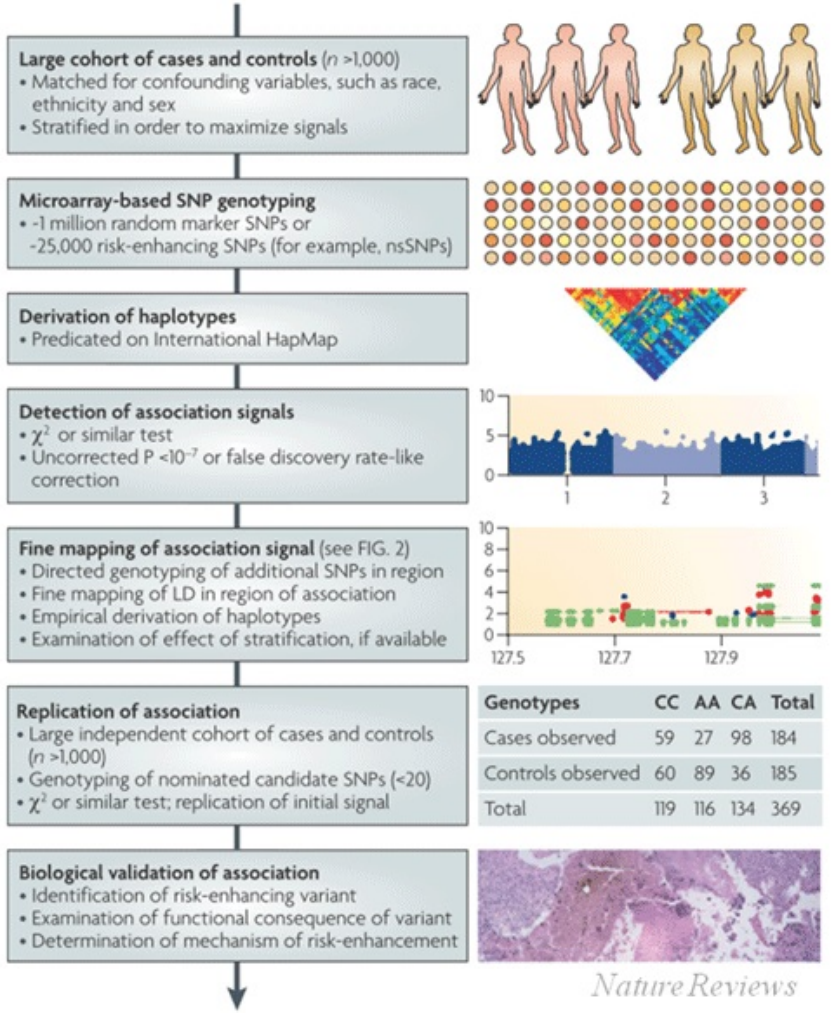
- Dissertation Chair and Academic Advisor, Peng Wei, PhD
- Minor Advisor, Alanna C. Morrison, PhD
- Breadth Advisor, Yun-Xin Fu, PhD
- External Advisor, Han Liang, PhD
- External Reviewer, Xiaoming Liu, PhD

Table of Contents

Table of Contents

Introduction to GWAS

A work flow of GWAS



Introduction to GWAS

GWAS contributes to precision medicine

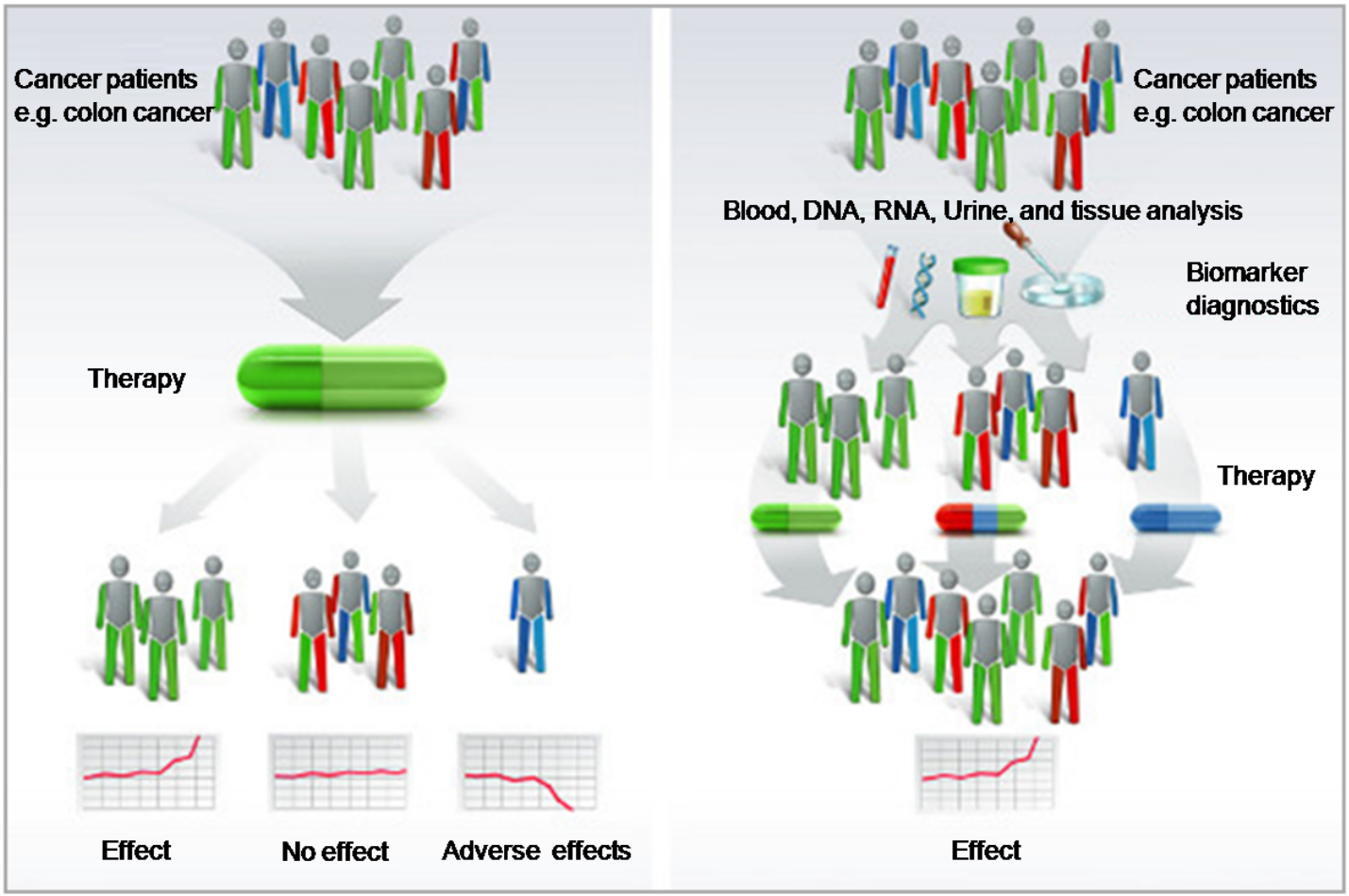
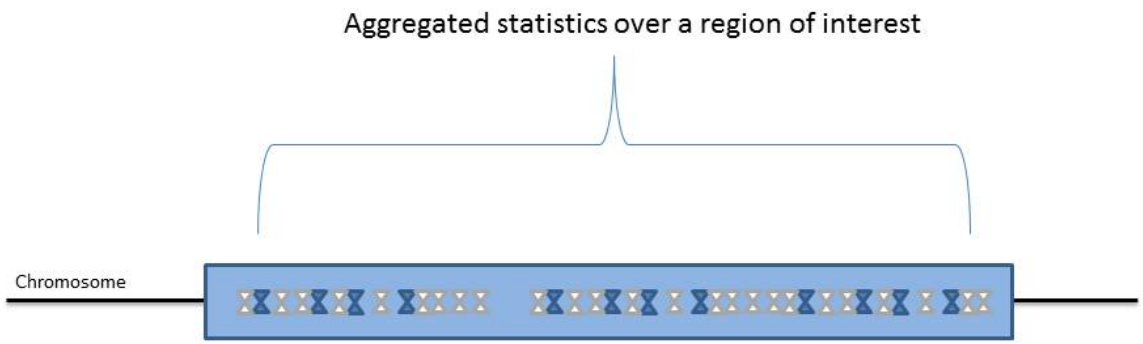


Figure: GWAS contributes to precision medicine

SNP-set based association tests

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



Longitudinal data analysis in GWAS

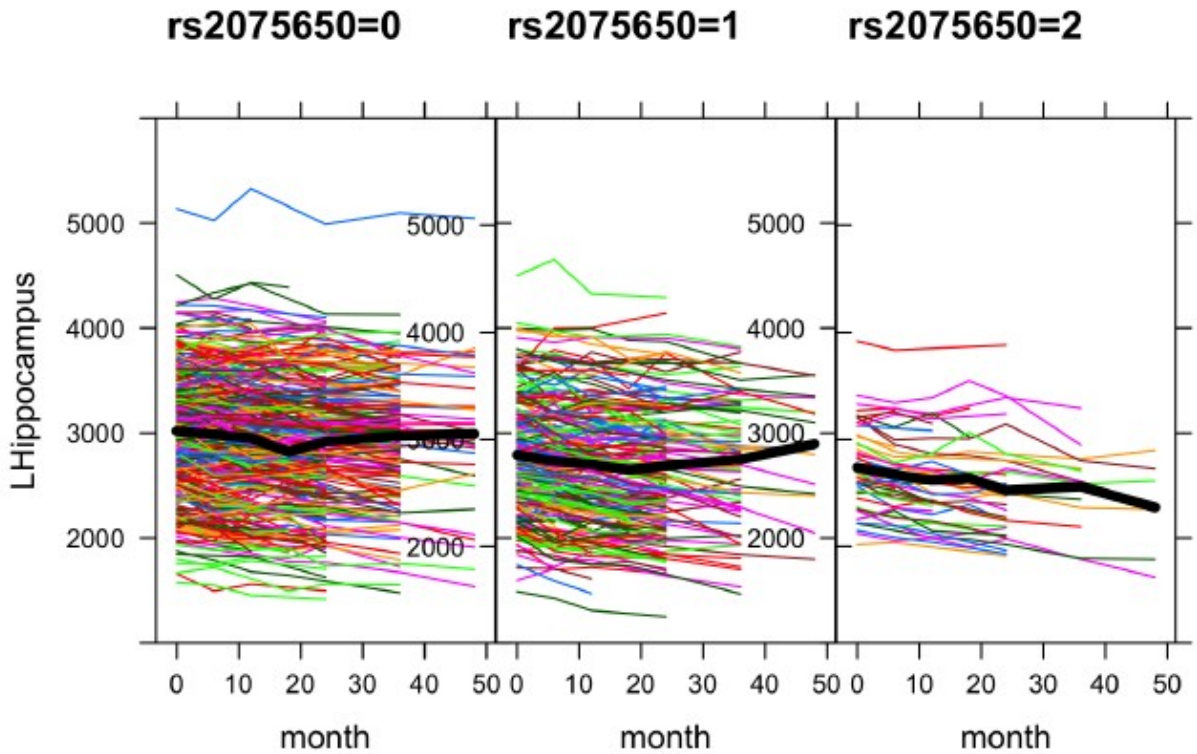


Figure: Trajectories of phenotype left hippocampus volume over time (in months) in three allele groups of SNP rs2075650 [XSP+14]

Longitudinal data analysis in GWAS

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

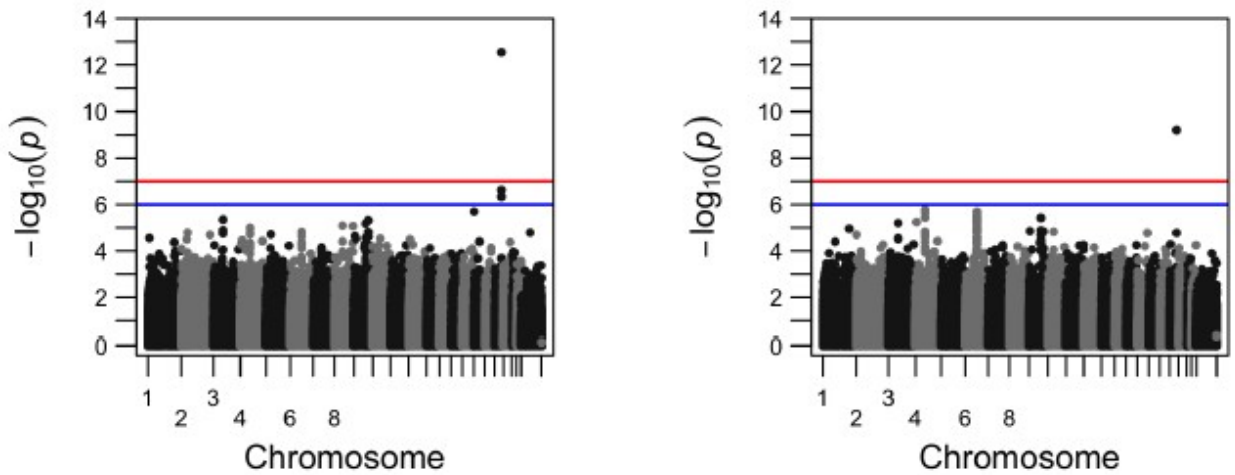


Figure: Comparison of the Manhattan plots for genome-wide p-values for phenotype left hippocampus volume from longitudinal analysis (left) and from cross-sectional analysis (right) [XSP+14]

Table of Contents

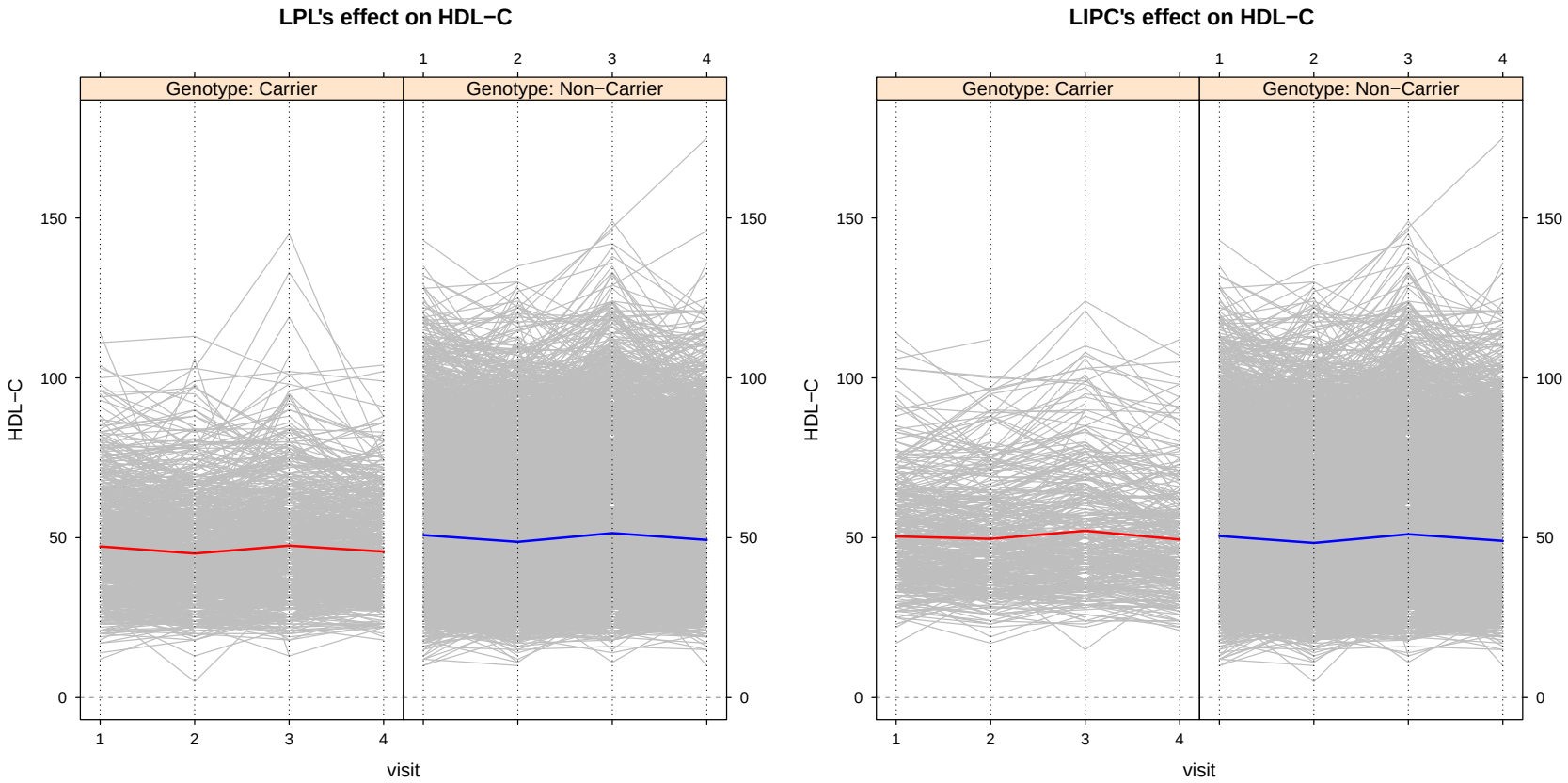
Simulation Studies

General Purpose

Evaluate the proposed methods' performance in

- maintaining the nominal Type I error rate
- maintaining a higher power under different simulation scenarios

Application to ARIC ExomeChip Data



- Phenotype: HDL-C levels measured at four visits in each of n=11,478 ARIC EA subjects
- Genotype: rare/low frequency variants (MAF $< 5\%$) on the ExomeChip
- Statistical methods: LaSPU(path) proposed here and its extensions
- Baseline (aSPU) vs all four measures (LaSPU)
- Gene-based and biological pathway-based association analysis

Table of Contents

Journal Article 1

Title of Journal Article

Data-adaptive SNP-set based association tests for longitudinal phenotypes within the Generalized Estimating Equations framework.

Table of Contents

Methods I

Journal Article 1

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 covariates, e.g., time, gender and leading principal components for population substructure.

Thus, we have:

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

Methods II

Journal Article 1

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

Denote the regression coefficient $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ and $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_{q+1})'$ for X_i and Z_i respectively. We then have the GLM equation as,

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

In the analysis of the ARIC data,

- $Z_i = time + gender + BMI(baseline) + age(baseline) + age(baseline)^2 + PC1 + PC2$
- $X_i = SNP_1 + SNP_2 + \ldots + SNP_p$

Methods III

Journal Article 1

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

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

Methods IV

Journal Article 1

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:

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

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

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

$$\widetilde{\Sigma} = \sum_{i=1}^{n} (Z_i, X_i)' \, \widehat{\mathrm{var}(Y_i)} \, (Z_i, X_i) = \sum_{i=1}^{n} (Z_i, X_i)' \left(\sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'}{n} \right) (Z_i, X_i) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \tag{2}$$

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

Methods V

Journal Article 1

Testing null hypothesis

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

where we have $g(Y_i) = Z_i\varphi$ to obtain φ and predict $\hat{\mu} = g^{-1}(Z\hat{\varphi})$. Then, we have score vector under 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 2.

Methods VI

Journal Article 1

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].

Methods VII

Journal Article 1

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 used $\gamma \in \Gamma = \{1, 2, \ldots, 8, \infty\}$ for the whole dissertation work.

Methods VIII

Journal Article 1

P-value calculation of the SPU test

Let T denotes $T_{SPU(\gamma)}$ for a specific γ.

Because the null distribution for SPU with γ larger than 2 is difficult to derive, so is aSPU to be introduced later, we recourse to simulation or permutation based method to calculate the null distribution.

With empirical distribution of $U_{.2}$ constructed, we can obtain the statistics under the null hypothesis: $T^{(b)} = \sum_{j=1}^{p} U_{.2,j}^{(b)\gamma}$ and calculate the p-value of $T_{SPU(\gamma)}$ as

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

Methods IX

Journal Article 1

Empirical Null Distribution of the Score Vector

We can obtain the empirical distribution of the score vector $U_{.2}$ under the null hypothesis by either the simulation or permutation procedure.

By simulation procedure, we draw B samples of $U_{.2}$ from its asymptotic distribution:

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

with $b = 1, 2, \ldots, B$.

Methods X

Journal Article 1

The permutation procedure is more robust to the failure of asymptotic property, e.g., the analysis of rare variants. The permutation procedure 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 SNPs. 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 $nk \times (p + q + 2)$.

Methods XI

Journal Article 1

4. permute the genotype codes across 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 recalculate $U_{.2}^{*(b)}$ by $U_{.2}^{*(b)} = \sum_i (X_i^{*(b)})'(Y_i - \hat{\mu}_i)$
6. repeat step 4 - 5 B times to produce $U_{.2}^{*(b)}$ with $b = 1, 2, \ldots, B$.

Methods XII

Journal Article 1

The longitudinal adaptive SPU test (LaSPU)

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 unknown association pattern. Thus, it will be convenient to have a test adaptively and automatically **selects the best** SPU(γ) test.

We therefore propose a longitudinal adaptive SPU (LaSPU) test:

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

Methods XIII

Journal Article 1

P-value calculation of LaSPU test

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 LaSPU test is:

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

It is worth noting again that the same B simulated score $(U_{.2})$ vectors have been used in calculating the P_{LaSPU}.

Methods XIV

Journal Article 1

The stage-wise genome wide scan strategy

In practice for genome wide computation purpose, we can use a stage-wise aSPU test strategy:

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

Methods XV

Journal Article 1

Other versions of LaSPU test

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

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

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

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

- **LaSPU omnibus test**

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

Table of Contents

Simulation Methods I

Journal Article 1

Simulation of genotype data following [WE07]

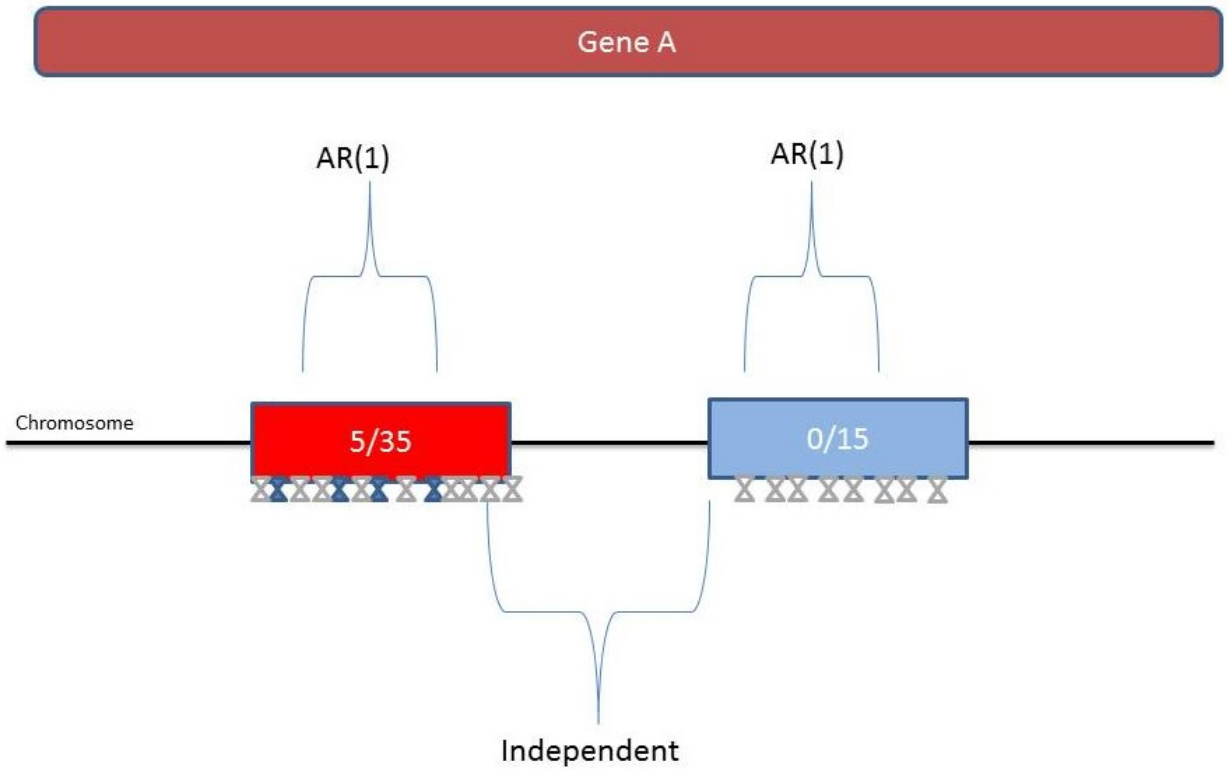


Figure: Demo graph of genotype simulation

Simulation Methods II

Journal Article 1

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:

Simulation Methods III

Journal Article 1

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

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

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

Simulation Methods IV

Journal Article 1

Summary of parameter setup in simulation studies

- $h_j^2 = 0.001$
- $\sigma_b = 7$
- $\sigma_e = 27$
- n varies between 500 and 3000
- $k = 4$
- 1000 replicates of simulated dataset (5000 replicates as optional validation, result not shown)
- $\alpha = 0.05$
- $\rho_y = 0.7$
- $\rho_x = 0.8$
- $R = AR(1)$
- $Rw = I$

Table of Contents

Simulation Results I

Journal Article 1

- **Simulation setting A**
  - $B = 1000$ for simulation of U
  - sample size of 500, 1000, 2000 and 3000
  - $5/15 + 0/35$ SNPs allocation
  - $0.05 < \text{MAF} < 0.4$
  - $h_j^2 = 0$
  - $\alpha = 0.05$

n	pSSU	pSSUw	pScore	pSum	pUminP	LaSPU	LaSPUw	LaSPU.sco	LaSPU.omni
500	0.047	0.048	0.047	0.052	0.048	0.060	0.058	0.058	0.058
1000	0.047	0.046	0.044	0.058	0.048	0.057	0.057	0.057	0.057
2000	0.049	0.047	0.051	0.048	0.048	0.061	0.058	0.058	0.058
3000	0.051	0.052	0.052	0.052	0.050	0.063	0.060	0.059	0.059

Table: Empirical Type I Error Table in the simulation setting A

Simulation Results II

Journal Article 1

- **Simulation setting B**
  - $h_j^2 = 0.001$

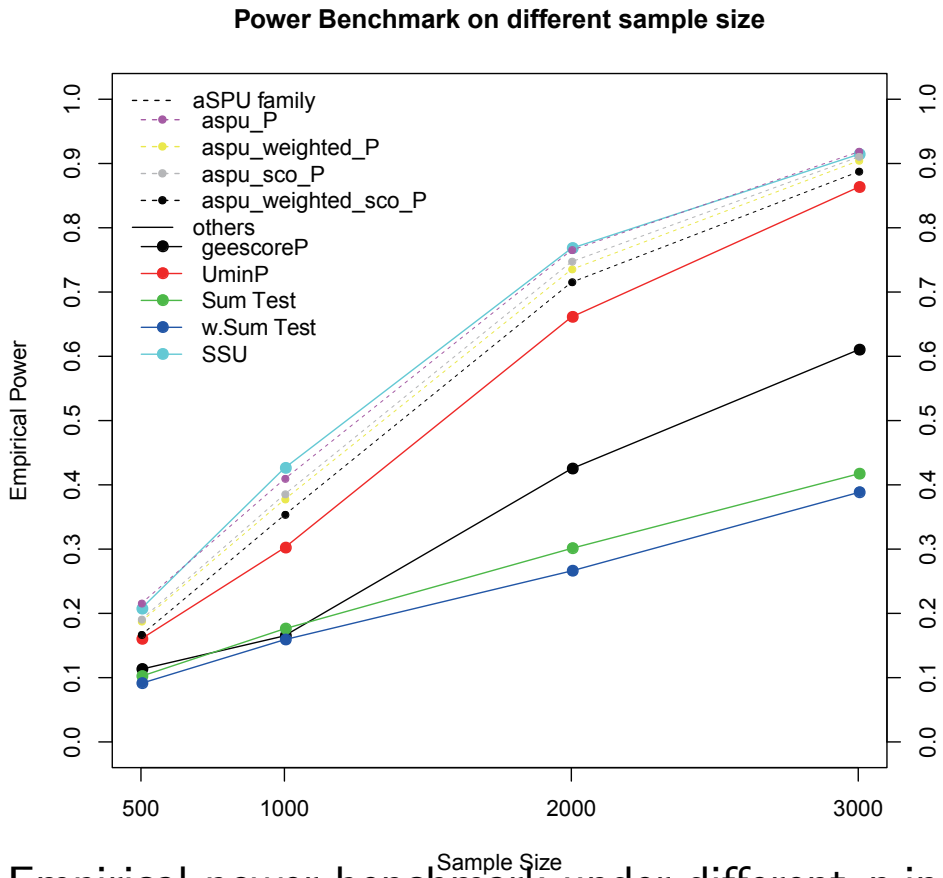


Figure: Empirical power benchmark under different n in the simulation setting B.

Simulation Results III

Journal Article 1

- **Simulation setting C: Testing with a growing number of null SNPs (in the second block)**

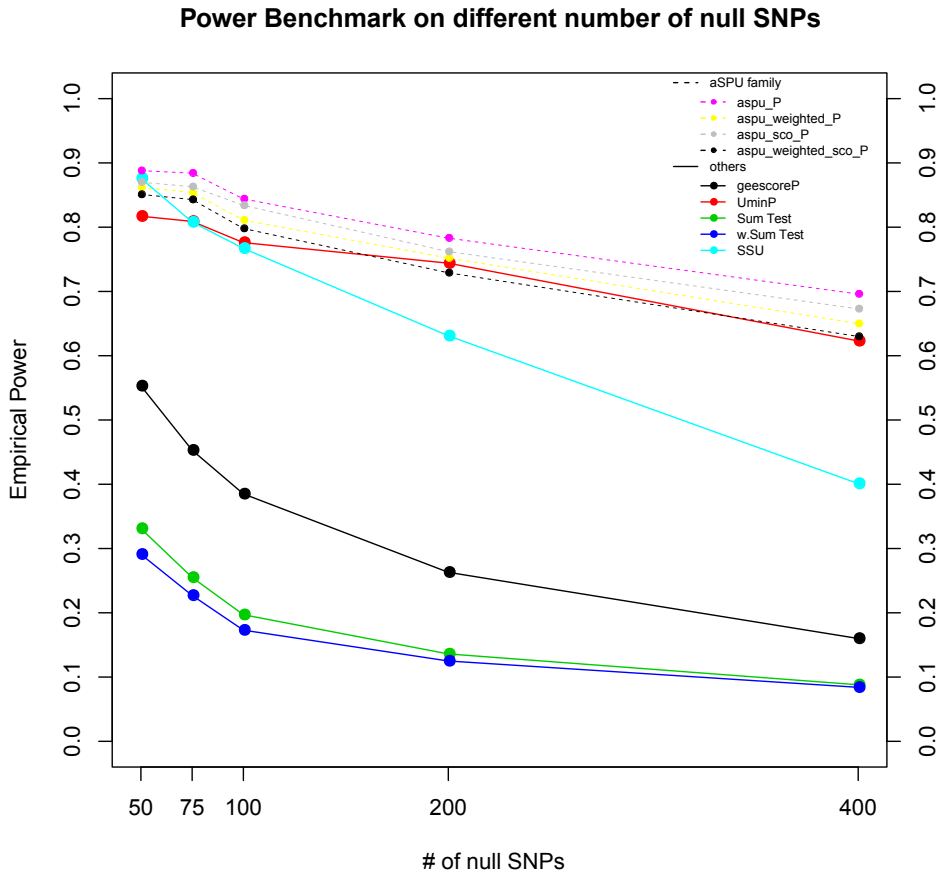


Figure: Empirical power benchmark under an increased number of null SNPs in the simulation setting C.

Simulation Results IV

Journal Article 1

- **Simulation setting D: assessing power gain from longitudinal measurements by decreasing the residual variance via repeated measures**

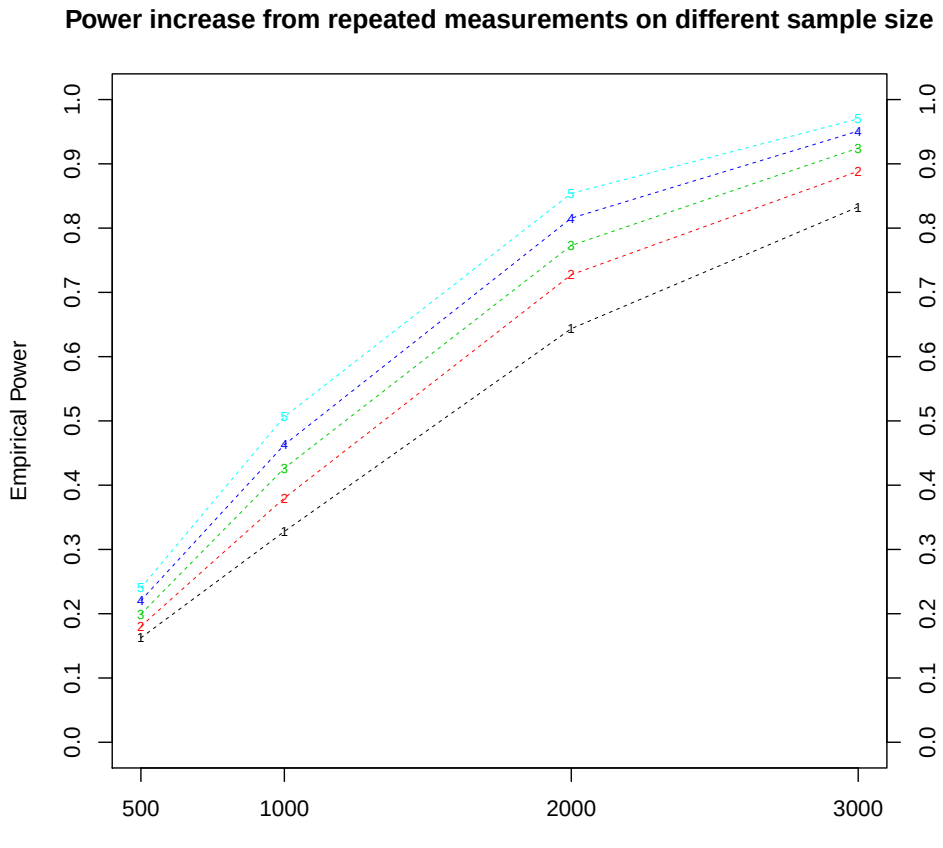


Figure: Power increases from repeated measurements in the simulation setting D.

Simulation Results for Rare Variants I

Journal Article 1

- **Simulation setting A'**
  - 5/15 with 5 included in the test
  - $0.001 < \text{MAF} < 0.01$
  - permutation based method to calculate the P-value

n	pSSU	pSSUw	pScore	pSum	pUminP	LaSPU	LaSPUw	LaSPU.sco	LaSPU.omni
500	0.053	0.054	0.052	0.049	0.047	0.054	0.053	0.060	0.058
1000	0.055	0.040	0.042	0.048	0.054	0.047	0.045	0.052	0.051
2000	0.054	0.050	0.048	0.049	0.046	0.063	0.057	0.058	0.057
3000	0.045	0.044	0.039	0.060	0.053	0.049	0.053	0.049	0.053

Table: Empirical Type I Error Table in the simulation setting A'.

Simulation Results for Rare Variants II

Journal Article 1

- **Simulation setting B'**

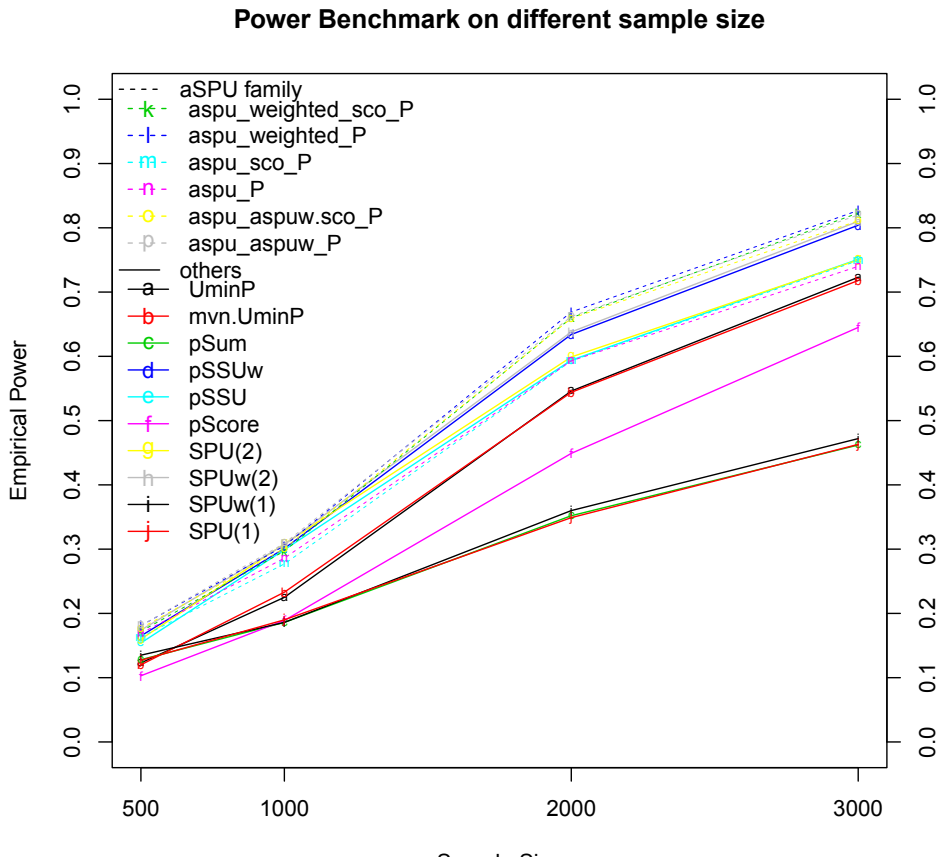


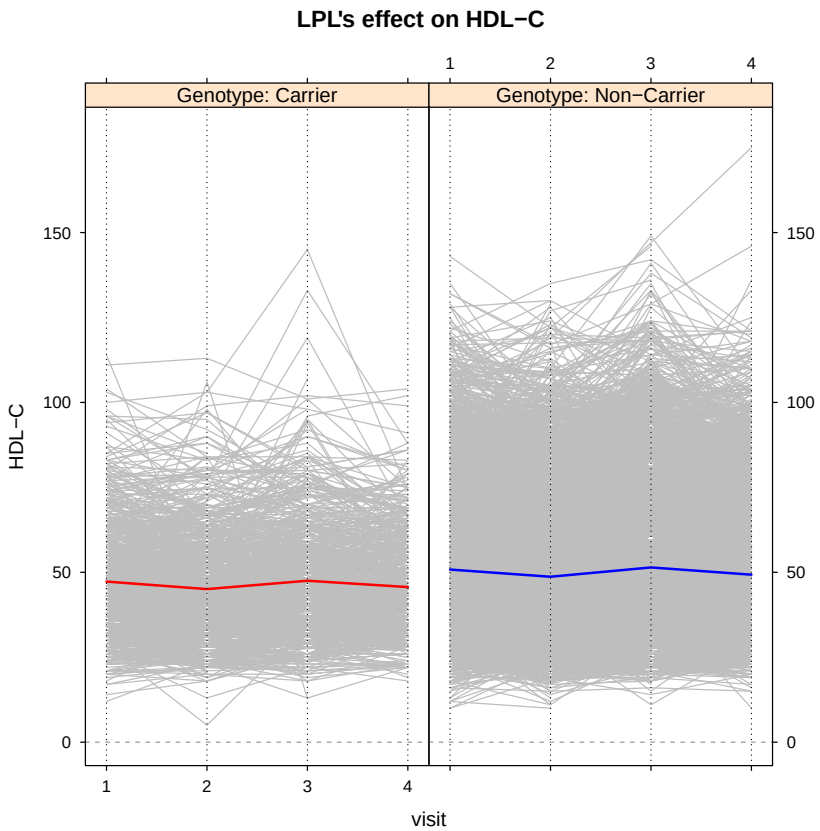
Figure: Empirical power benchmark under different n in the simulation setting B'.

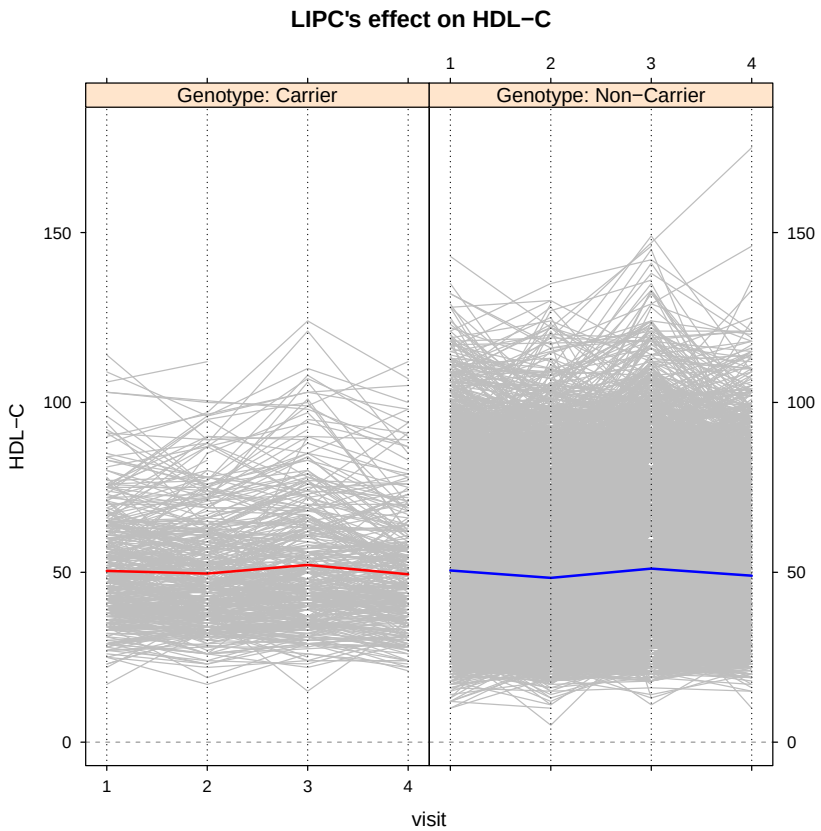
Table of Contents

Application to the ARIC study I

Journal Article 1

Recap the ARIC data





Application to the ARIC study II

Journal Article 1

- Phenotype: HDL-C levels measured at four visits in each of n=11,478 ARIC EA subjects
- Genotype: rare/low frequency variants (MAF $< 5\%$) on the ExomeChip
- Statistical methods: LaSPU proposed here and its extensions
- Baseline (aSPU) vs all four measures (LaSPU)
- Gene-based association analysis

Application to the ARIC study III

Journal Article 1

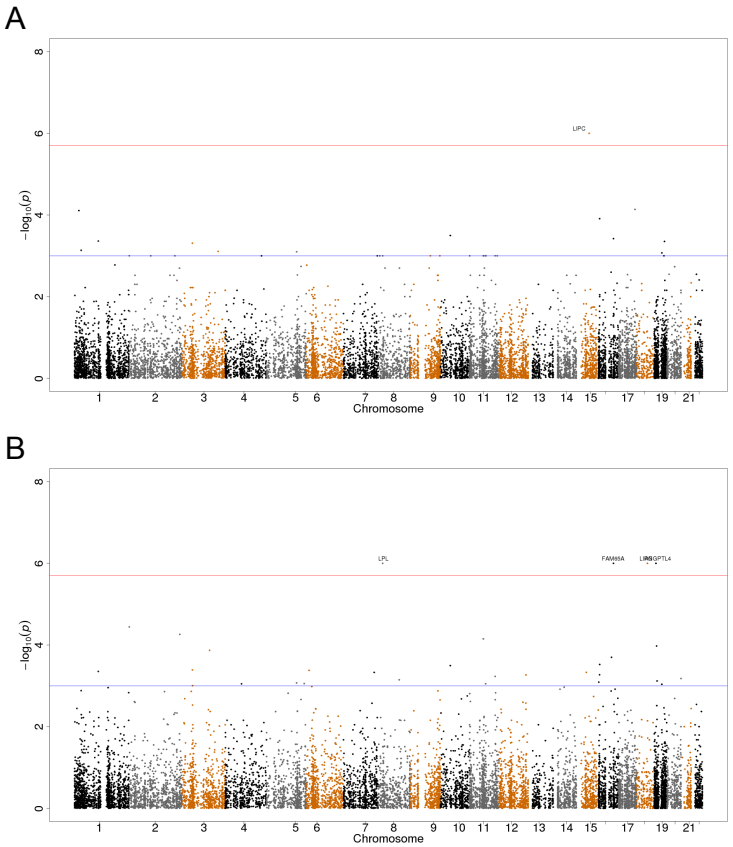
Methods in data application

- RV defined by MAF < 0.05
- Inclusion of only nonsynonymous and splice site variants
- QC procedures as in [PAB+14]
- Additive genetic model
- Covariates include age, age^2, gender, BMI, time of measurement and top two PCs
- Gene boundary defined in [PAB+14]
- Significant: 0.05/25,000 = 2e-06; marginal significant: 1e-03

Application to the ARIC study IV

Journal Article 1

Results in data application: compare longitudinal method with baseline method



LIPC: previously not reported

LPL, LIPG, ANGPTL4: previously reported;

FAM65B: previously not reported

Figure: Manhattan Plot Comparison between baseline study and longitudinal study by LaSPU test on the association between HDL-C and Rare Variants in the ARIC study. A. baseline study; B. longitudinal study using total four measurements.

Application to the ARIC study V

Journal Article 1

Results in data application: compare longitudinal method with baseline method

Table: Top Gene-Based Association Results Based on Level of Statistical Significance

Gene	Chr	p Value	No.Variants[a]	CMAC[b]	CMAF[c]	p Value of Baseline[d]
LPL	8	1.00E-06	10	879	0.00807	9.99E-04
*FAM65A**	16	1.00E-06	11	751	0.00627	3.79E-04
LIPG	18	1.00E-06	11	369	0.00308	3.13E-02
ANGPTL4	19	1.00E-06	9	579	0.00591	2.89E-01
ANGPTL8	19	1.06E-04	5	64	0.00118	2.07E-01
APOC3	11	5.87E-04	3	21	0.00064	9.99E-04
PAFAH1B2	11	2.19E-03	3	287	0.00879	1.50E-02

[a] number of variants contributing to the test

[b] cumulative minor allele count of the variants contributing to the test

[c] cumulative minor allele frequency of the variants contributing to the test

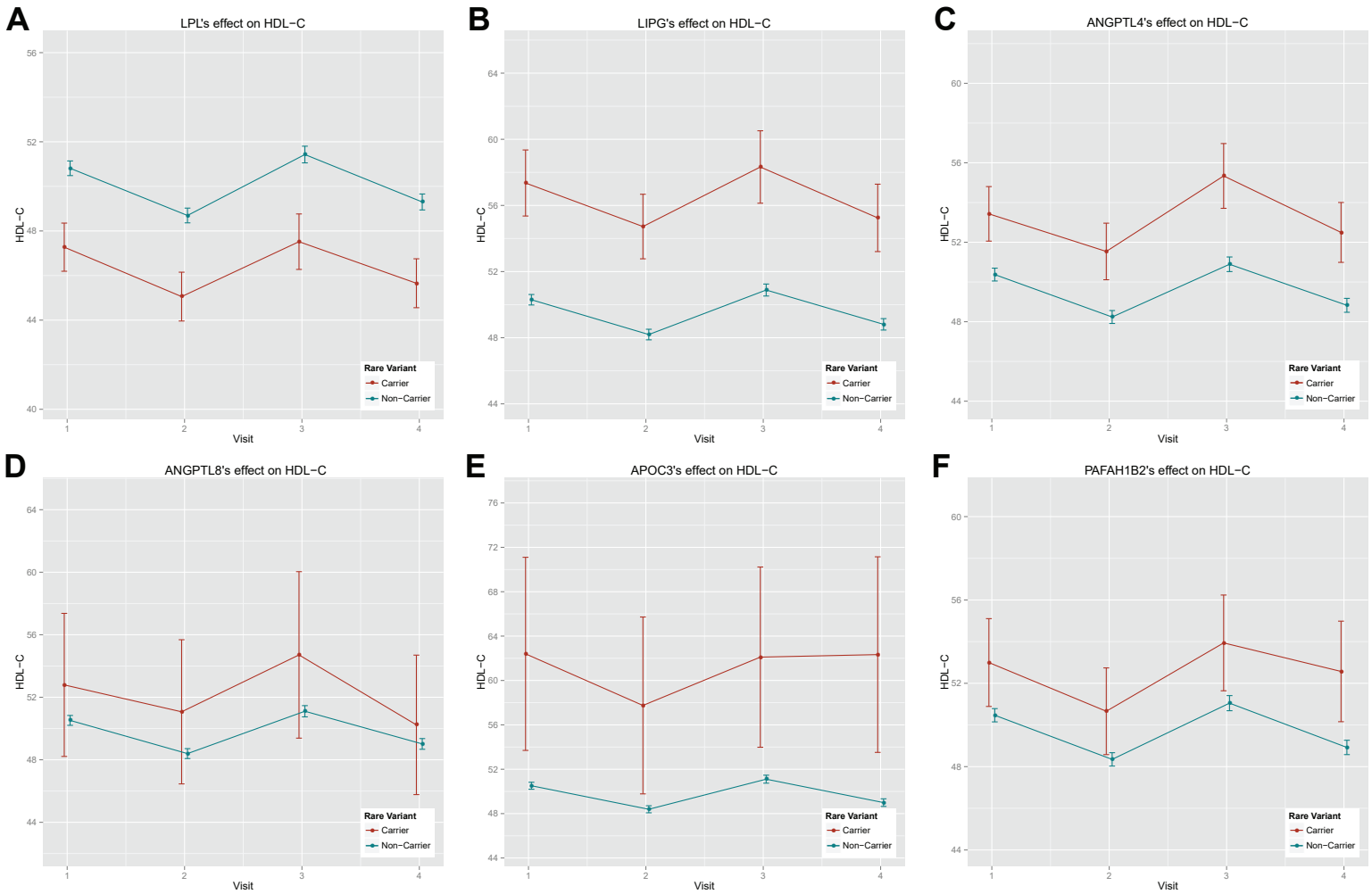
[d] aSPU test using baseline only measurement of HDL-C

* novelly identified gene(s)

Application to the ARIC study VI

Journal Article 1

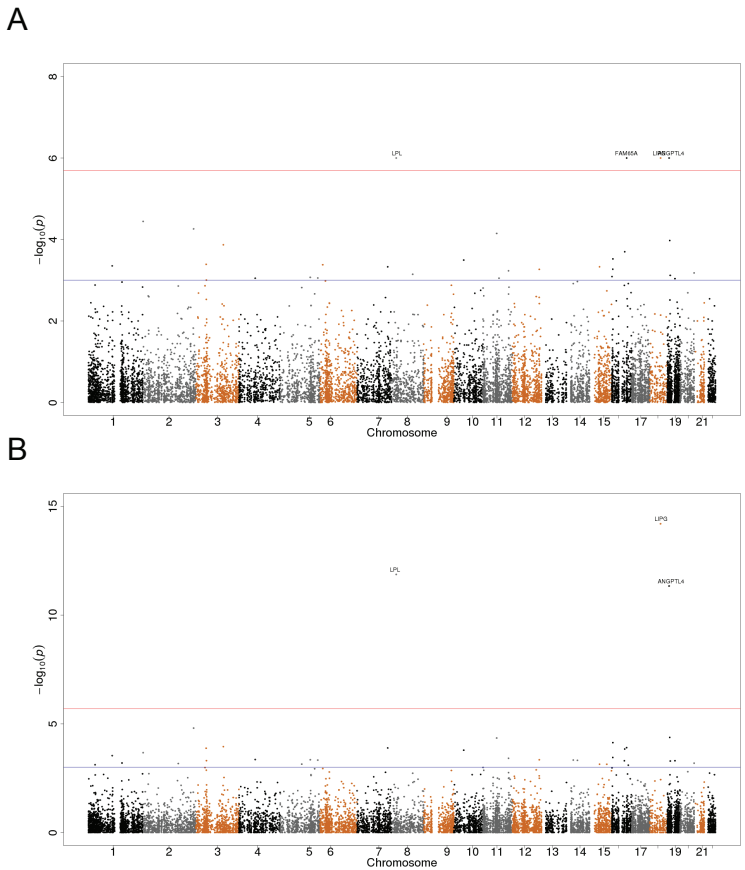
Gene effect visualization of previously reported genes



Application to the ARIC study VII

Journal Article 1

Results in data application: compare LaSPU with LSSU method



LPL, LIPG, ANGPTL4: previously reported;

FAM65B: previously not reported

LPL, LIPG, ANGPTL4: previously reported;

Figure: Manhattan Plot Comparison between LaSPU test and LSSU test on the association between HDL-C and Rare Variants in the ARIC study. A. result by LaSPU; B. result by LSSU .

Application to the ARIC study VIII

Journal Article 1

Gene effect visualization of previously unreported genes
A: *FAM65A* identified by only LaSPU and LSum tests; **B:** *LIPC* identified by only aSPU (baseline analysis)

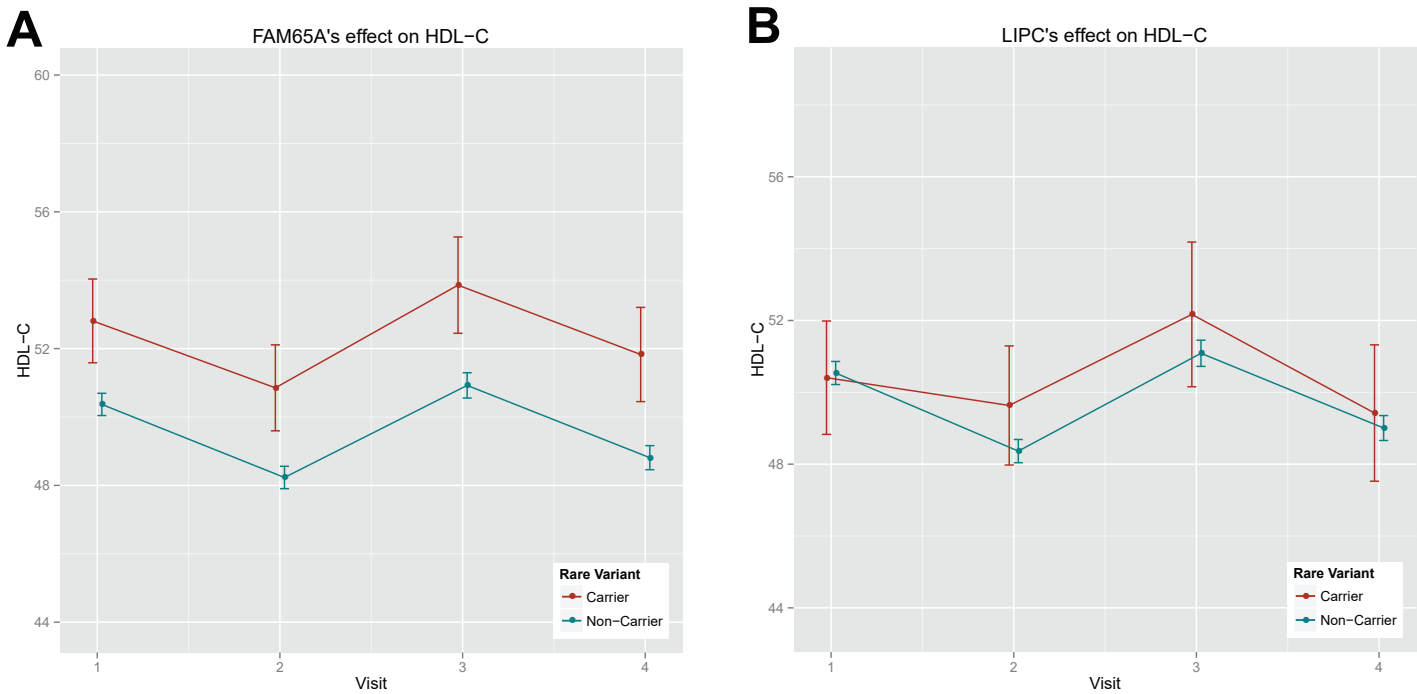


Table of Contents

Discussion I

Journal Article 1

- LaSPU could efficiently use the **longitudinal** information to increase testing power
- LaSPU could **adaptively** select the best test and achieve a greater power
- LaSPU proved its ability in real data analysis by identifying the **same** reported genes associated with HDL-C using the ARIC data set only
- LaSPU identified one **novel** gene associated with HDL-C: *FAM65B*, which warrants further validation

Table of Contents

Journal Article 2

Title of Journal Article

Pathway-based data-adaptive association tests for longitudinal phenotypes.

Table of Contents

Methods I

Journal Article 2

Pathway idea illustration

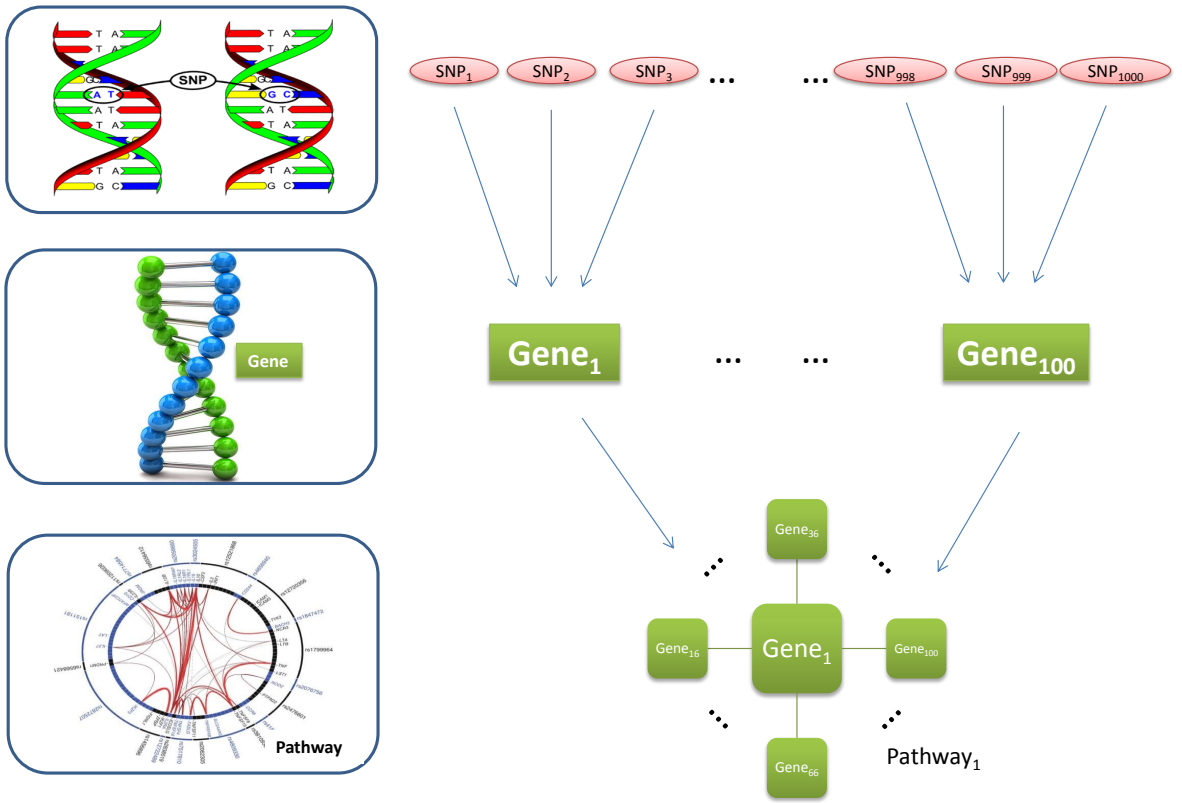


Figure: Aggregation of SNPs in a Pathway

Methods II

Journal Article 2

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 $U_{g.}$ represent $U_{.2}$ for the RVs X_i'part in the whole score vector, and $U_{g.} = (U_{g,1}, U_{g,1}, \ldots, U_{g,p_g})'$ 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:

$$T_{SPU(\gamma;g)} \propto ||Ug.||_\gamma = \left(\frac{\sum_{j=1}^{p_g} |U_{g,j}|^\gamma}{p_g} \right)^{\frac{1}{\gamma}} \tag{5}$$

Then accordingly, the pathway-based SPU statistic is

$$T_{Path-SPU(\gamma,\gamma 2;S)} = \sum_{g \in S} (T_{SPU(\gamma;g)})^{\gamma 2} \tag{6}$$

Methods III

Journal Article 2

A data-adaptive pathway-based longitudinal association test: LaSPUpath

The pathway-based aSPU statistic is thus

$$T_{Path-aSPU(S)} = min_{\gamma,\gamma 2} P_{Path-SPU(\gamma,\gamma 2;S)} \tag{7}$$

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

For any given $\gamma, \gamma 2$, We recourse to the similar simulation or permutation based strategy as in LaSPU test to calculate the P-values of a class of SPUpath tests, and then calculate the final P-value of the LaSPUpath test.

Table of Contents

Simulation Methods I

Journal Article 2

- We simulated a pathway with 20 independent genes. Each gene g contained p_g SNPs with p_g randomly draw from a uniform distribution $U(3, 20)$; 5 of the 20 genes will be randomly selected to be causal, with each causal gene containing $U(1, 3)$ causal SNPs.
- The other parts of the simulation are the same as Article 1.

Simulation Methods II

Journal Article 2

Summary of parameter setup in simulation studies

- $h_j^2 = 0$ or $h_j^2 = 0.001$, 0.0025, 0.005, 0.0075 and 0.010
- $\sigma_b = 7$
- $\sigma_e = 27$
- $k = 5$
- 1000 replicates of simulated dataset
- B = 1000
- $\alpha = 0.05$
- $\rho_y = 0.7$
- $\rho_x = 0.8$
- $R = AR(1)$
- $Rw = I$
- $0.05 <$ MAF < 0.4 for CVs; $0.001 <$ MAF < 0.01 for RVs
- causal SNPs are excluded from testing for CVs but retained for RVs

Table of Contents

Simulation Results

Journal Article 2

- **Simulation setting A**
  - $0.05 < \text{MAF} < 0.4$
  - $h_j^2 = 0$
  - $\alpha = 0.05$

pSSU	pSSUw	pScore	pSum	pUminP	LaSPUpath
0.037	0.042	0.025	0.049	0.050	0.053

Table: Empirical Type I Error Table in simulation set-up A.

Simulation Results

Journal Article 2

- **Simulation setting B**

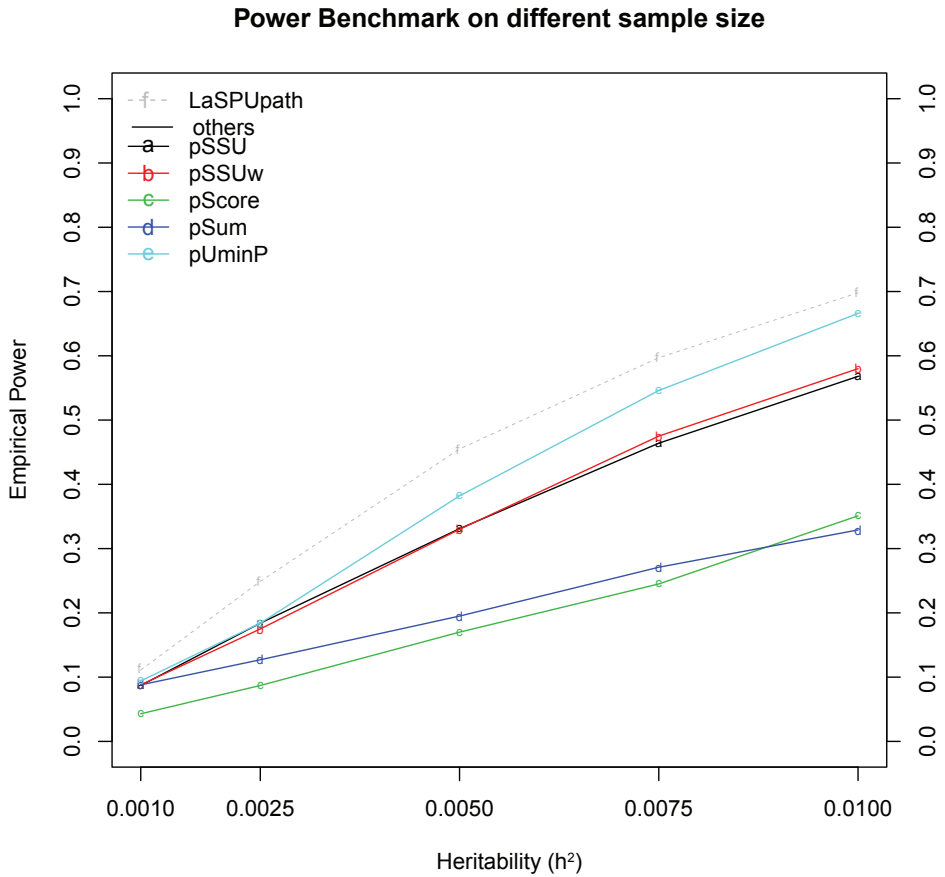


Figure: Empirical power benchmark under different heritability (h^2) in simulation set-up B.

Table of Contents

Application to the ARIC study I

Journal Article 2

Methods in data application

- RV defined by MAF < 0.05
- Inclusion of only nonsynonymous and splice site variants
- QC procedures as in [PAB+14]
- Additive genetic model
- Covariates include age, age^2, gender, BMI, time of measurement and top two PCs
- obtain biological pathways from the KEGG database [OGS+99]
- trim pathway to be of moderate size (between 10 and 500 genes)
- Gene boundary defined in [PAB+14]
- Significant: $0.05/197 = 0.00025$; marginal significant: 1e-03

Application to the ARIC study II

Journal Article 2

Results in data application: significant pathways

Table: Results of the ARIC Data Application: KEGG Pathways with p Value < 0.00025

KEGG ID	Pathway Name	No. of Genes	No. of SNPs	p Value	Contributing Genes[a]
hsa00561	Glycerolipid metabolism	44	311	1.00E-06	**LPL**,**LIPG**,*LIPC*, *DGKQ*,*PPAP2A*,*PNLIPRP3*
hsa03320	PPAR signaling pathway	55	465	1.00E-06	**LPL**,**ANGPTL4**,*NR1H3*, *CD36*,*APOA1*
hsa05010	Alzheimers disease	98	747	1.00E-06	**LPL**,*APOE*,*BID*,*PLCB3*,*NDUFS8*, *NDUFS3*,*NDUFB6*,*RYR3*,*NCSTN*

[a] p Value of the gene < 0.05

- hsa00561: It belongs to lipid metabolism class; associated with Hyperlipoproteinemia, type I
- hsa03320: It was reported to play a role in the clearance of circulating or cellular lipids via the regulation of gene expression; associated with multiple lipid-related diseases like Hyperalphalipoproteinemia
- hsa05010: There are a number of studies reporting a reduced level of HDL-C is highly correlated with the severity of AD; atherosclerosis also increases the risk of AD.

Table of Contents

Discussion I

Journal Article 1

- LaSPUpath could efficiently use the **longitudinal** information to increase testing power
- LaSPUpath could efficiently use the **biological** information (such as pathway) to increase the testing power
- LaSPUpath could **adaptively** select the best test and achieve a greater power
- LaSPUpath identified **three** pathways significantly associated with HDL-C. The annotated functions of the pathways are closely related to HDL-C, either as the regulator of lipids (including HDL-C) and/or lead to the lipid-related diseases like Hyperlipoproteinemia and Alzheimer's disease.
- LaSPUpath could be used in **more general** gene-set based testings, for example, annotations of functional variation available in ENCODE (https://www.encodeproject.org/) and epigenome activities collected in the Roadmap Epigenomic Data at NCBI/GEO (http://www.roadmapepigenomics.org/).

Table of Contents

Journal Article 3

Title of Journal Article

LaSPU: a suite of powerful data-adaptive SNP-set and pathway-based association testing tools for longitudinal traits.

Features

Journal Article 3

- implement the LaSPU method
- compatible with Unix-like system
- command line operated program
- optimized and possess flexible parallel computing schema
- user-friendly manual and examples

Workflow

Journal Article 3

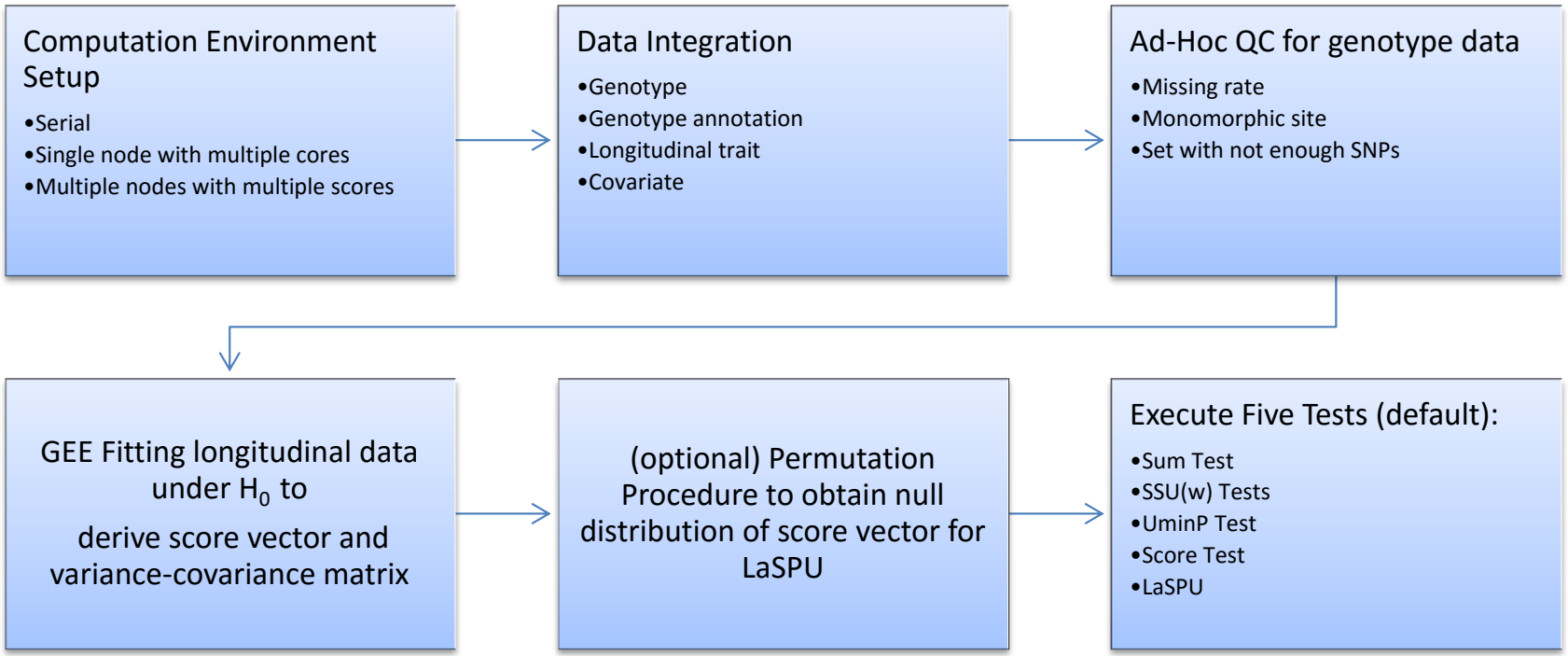


Figure: LaSPU Workflow Chart.

Result

Journal Article 3

Table of Contents

Conclusion I

In brief, our developed association test(s) utilized

- the longitudinal information
- the data adaptivity to select the most powerful test
- the biological information such as biological pathways

to boost the statistical power.

Conclusion II

More elaborately, in this dissertation we have

- developed the LaSPU method.
- developed the LaSPUpath method.
- performed simulation studies that demonstrated that both LaSPU and LaSPUpath could efficiently utilize the **longitudinal** information and **adaptively** select the best test to increase the statistical power as compared to non-adaptive and/or baseline testing methods.
- performed simulation studies that also demonstrated that the LaSPUpath could efficiently utilize the additional **biological** information when available, such as the biological pathway information, to further boost the statistical power.
- Real data application of LaSPU demonstrated that LaSPU can **replicate** the previously reported genes with a much **smaller sample size** (1/4). On the other hand, LaSPU identified a **novel** gene *FAM65B* associated with HDL-C.
- Real data application of LaSPUpath identified **three** significant pathways, the functional annotations of which all closely relate the regulation of lipid metabolism and/or the onset of lipid-related diseases.
- The development of the "LaSPU" software package enables the research community to implement the novel methods with ease.

Future Work

- **Joint test** of the SNPs' main effect and the SNPs' interaction effect with time in the LaSPU/LaSPUpath framework
- Setup of appropriate **weighting schema** for SNPs (and genes) to incorporate functional annotations of SNPs (and genes)
- Setup of appropriate **weighting schema** for SNPs to enable the analysis of CVs and RVs in the same region
- Extension from testing the genetic pathway to other functional related gene sets.
- Inclusion of the aSPUpath method into the "LaSPU" software package.

Table of Contents

Acknowledgement

Advisors:

- Peng Wei, Ph.D
- Alanna C. Morrison, Ph.D
- Yunxin Fu, Ph.D
- Han Liang, Ph.D, Associate Professor, Department of Bioinformatics and Computational Biology, The University of Texas M.D Anderson Cancer Center

External Reviewer:
Xiaoming Liu, Ph.D
Collaborator
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)
My classmates, colleges and friends at UTSPH and MDACC
...
My Parents
Tianpeng Yang and Qi Lu
My Wife
Nainan Hei

Thanks to Audience

Thank you for your participation!

References I


Alan P Boyle, Eurie L Hong, Manoj Hariharan, Yong Cheng, Marc A Schaub, Maya Kasowski, Konrad J Karczewski, Julie Park, Benjamin C Hitz, Shuai Weng, et al., *Annotation of functional variation in personal genomes using regulomedb*, Genome research **22** (2012), no. 9, 1790–1797.


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.


Kung-Yee Liang and Scott L Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika **73** (1986), no. 1, 13–22.


Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa, *Kegg: Kyoto encyclopedia of genes and genomes*, Nucleic acids research **27** (1999), no. 1, 29–34.


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.


Tao Wang and Robert C Elston, *Improved power by use of a weighted score test for linkage disequilibrium mapping*, The american journal of human genetics **80** (2007), no. 2, 353–360.

References II

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.