DATA-ADAPTIVE SNP-SET-BASED ASSOCIATION TESTS OF LONGITUDINAL TRAITS

by

YANG YANG, B.S, M.S

APPROVED:

PENG WEI, PHD

ALANNA C. MORRISON, PHD

YUN-XIN FU, PHD

HAN LIANG, PHD

DEAN, THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH

DEDICATION

To my family members:

Nainan Hei

&

Tianpeng Yang and Qi Lu

DATA-ADAPTIVE SNP-SET-BASED ASSOCIATION TESTS OF LONGITUDINAL TRAITS

by

YANG YANG

B.S, Nanjing Agricultural University, 2006

M.S, Indiana University, 2010

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
December, 2015

ACKNOWLEDGMENTS

Great gratitude to my dissertation advisor Dr. Peng Wei, as he mentored me ever from 2011, put countless efforts in training me to be an accountable person, and then a qualified Ph.D. He taught me with his solid background in statistical theory, to make me an as well solid statistician to qualify for future career challenges; he corrected me many times to let me not bypass but instead overcome the difficulty in speaking and writing in a native English style; he also taught me the spirit of persistence, either in research or in life, which is indispensable to every kind of definition of success. I also want to appreciate the great helps from my dissertation committee members: Dr. Alanna C. Morrison, Dr. Yun-Xin Fu and Dr. Han Liang. They are talented experts in their fields and provided me with enormous valuable advice towards my research and scientific writings. I also want to express my special gratitude to Dr. Han Liang. As I have been a Graduate Research Assistant at MD Anderson Cancer Center under his supervision and mentoring between 2012 and 2013, he inspired me to be a bioinformatics researcher rather than a proficient analyst, ignited me the passion in cancer genomics, influenced me to have innovative thinking and meticulous altitude in pursuing science. In the end, I want to express my gratitude to the NIH/NHLBI grant "Association analysis of rare variants with sequencing data" (R01 HL116720), which has supported my dissertation research.

DATA-ADAPTIVE SNP-SET-BASED ASSOCIATION TESTS OF LONGITUDINAL TRAITS

Yang Yang, M.S, PhD
The University of Texas
School of Public Health, 2015

Dissertation Chair: Peng Wei, PhD

Abstract

Genome-wide association studies (GWASs) have been largely limited to investigating traits with a single time measurement. However, many prospective cohort studies and electronic health record (EHR)-based cohorts with GWAS data have collected traits with repeated measurements across follow-up time. Effectively utilizing the information embedded in the time trajectory of measurements could greatly increase the power of association testing. Local association signal patterns across the whole genome are usually variable, complicated and unpredictable, which underscores the need for a data adaptive test capable of maintaining high statistical power across different genetic architecture and association signal patterns. Furthermore, complex diseases are usually affected by multiple variants in a gene and multiple genes in a biological pathway. In addition, traditional single SNP-based association testing has very limited statistical power for variants of low to rare minor allele frequency (MAF $< 5\%$ or 1%). A SNP-set-based association test, e.g., based on genes or pathways, could boost the statistical power by aggregating individual weak to moderate association signals across a region of interest. In this dissertation, I have developed such powerful data-adaptive tests that address these analytical challenges facing association testing between longitudinal traits and rare or common variants. I implemented extensive simulation studies to evaluate the perfor-

mance of the proposed tests, and illustrated their applications in the Atherosclerosis Risk in Communities (ARIC) study. I also produced a software package with documentation to implement the proposed tests in high performance computing platform. In conclusion, this dissertation paves a new path in extending the traditional association tests to longitudinal traits, helps identify novel genes and explain the missing heritability in human complex diseases.

Contents

1 Background **1**

1.1 Literature Review 1

1.1.1 Gene-based association tests 2

1.1.2 Longitudinal study design and analysis strategy in GWAS 7

1.1.3 Gene-set/Pathway based association tests 14

1.2 Public Health Significance 23

1.3 Specific Aims 24

2 Methods **26**

2.1 Overall Study Design 26

2.1.1 Simulation studies 26

2.1.2 Real data application 26

2.2 Methods for Aim 1(a): To develop data-adaptive SNP-set based association tests for longitudinal data analysis within the Generalized Estimating Equations (GEE) framework for common variants 30

2.2.1 Statistical Modeling 30

2.2.2 Methods for Simulation Settings 41

2.2.3 Plan for Simulation Studies 45

2.3 Methods for Aim 1(b): To develop data-adaptive SNP-set based association tests for longitudinal data analysis within the Generalized Estimating Equations (GEE) framework for rare variants 49

2.3.1 Statistical Modeling 49
2.3.2 Methods for Simulation Settings 50
2.3.3 Plan for Simulation Studies 51
2.4 Methods for Aim 2: To develop the pathway-based data-adaptive association tests for longitudinal data analysis 54
2.4.1 Statistical Modeling 54
2.4.2 Methods for Simulation Settings 56
2.4.3 Plan for Simulation Studies 56
2.5 Methods for Aim 3: To develop the software package for method implementation 57
2.6 Methods for Real Data Application for Proposed Aims 58
2.7 Declaration on Human Subjects 60

3 Journal Article 1 **61**

3.1 Introduction 64
3.2 Material and Methods 67
3.2.1 Data and Notation 67
3.2.2 Several Existing Association Tests 71
3.2.3 The Data-Adaptive LaSPU Test 73
3.2.4 Other Modifications 75
3.2.5 Extension to Rare Variant Analysis 77
3.2.6 Simulation Set-ups 79

3.3 Results 83
3.3.1 Simulation Results for CVs 83
3.3.2 Simulation results for RVs 91
3.3.3 Application to the ARIC study 95
3.4 Discussion 103

4 Journal Article 2 **107**

4.1 Introduction 110
4.2 Material and Methods 115
4.2.1 Data and Notation 115
4.2.2 Several Existing Set-based Association Tests 119
4.2.3 Review: The Data-Adaptive aSPU Test 121
4.2.4 A Data-Adaptive Pathway-Based longitudinal test: LaSPUpath . 123
4.2.5 Extension to Rare Variant Analysis 126
4.2.6 Simulation Set-ups 128
4.3 Results 132
4.3.1 Simulation Results for CVs 132
4.3.2 Application to the ARIC study 134
4.4 Discussion 137

5 Journal Article 3 **141**

5.1 Introduction 143

5.2 Methods 143

5.2.1 Features 143

5.2.2 Workflow 144

5.3 Conclusion 147

5.4 Appendix for Article 3 - LaSPU Manual 148

6 Conclusion and Future Directions **152**

References **156**

List of Tables

1 Sample Table of Type I error Benchmark among tests 47

2 Sample Table of Type I error Benchmark among tests using simulation-based method in RV analysis. mvn.UminP: UminP calculated by approximating a MVN distribution; UminP: UminP method calculated by simulation-based method. 52

3 Empirical Type I Error Table in the simulation setting A. 84

4 Empirical Type I Error Table in the simulation setting A'. 93

5 Top Gene-Based Association Results Based on Level of statistical Significance 100

6 Empirical Type I Error Table in the simulation set-up A. 132

7 Results of the ARIC Data Application: KEGG Pathways with p Value $<$ 0.00025 137

List of Figures

1 **Examples of competitive approach and self-contained approach based testings using Fisher's exact test as a demonstration** (A). Example of competitive approach; (B). Example of self-contained approach. This figure is adopted from [Fridley and Biernacka, 2011]. . . . 18

2 **Types of pathway association tests in GWAS.** (a). Categorization based on data input type; (b). Categorization based on hypothesis testing. This figure is adopted from Wang et al (2010) [Wang et al., 2010]. . . . 20

3 ARIC Cohort Characteristics by Gender or Race. Table adopted from the ARIC website 29

4 Empirical power benchmark under different n in the simulation setting B. 85

5 Empirical power benchmark under a heterogeneous SNP effects in the simulation setting C. 86

6 Empirical power benchmark under an increased number of null SNPs in the simulation setting D. 87

7 Empirical power benchmark with more tests in the simulation setting B. 89

8 Power increases from repeated measurements in the simulation setting E. 91

9 Empirical power benchmark under different n in the simulation setting B'. 95

10 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. 99

11 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 102

12 Aggregation of SNPs in a Pathway 115

13 Empirical power benchmark under different heritability (h^2) in the simulation set-up B. 133

14 LaSPU Workflow Chart. 146

15 Example Output by LaSPU in Manhattan Plot. 147

1 Background

1.1 Literature Review

Genome-wide association studies (GWASs) have been popular since 2007. Hundreds of GWASs have been published already (A Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/26525384). The most popular approach in GWAS is to test the association with complex traits on single nucleotide polymorphism (SNP), also known as single nucleotide variant (SNV), one by one, then select the the SNVs that meet a stringent significance level after multiple testing correction, such as the Bonferroni and false discovery rate (FDR) methods [McCarthy et al., 2008, Hirschhorn and Daly, 2005]. However, this strategy will suffer from low power when the minor allele frequency (MAF) of the SNV is low (between 1% and 5%), and as a result the signals contained within the low MAF SNVs are hard to detect [Sham and Purcell, 2014]. In addition, the usual regression coefficient estimate of SNV becomes unstable due to the small number of minor allele counts and the coefficient estimate's variance becomes very large [Sham and Purcell, 2014]. It will become an even more severe problem for rare variants (RVs) analysis. RVs are usually defined as SNVs with MAF below 1% [Bansal et al., 2010]. In spite of their extremely low MAF, RVs' important role in conferring disease risk cannot be underestimated. Due to the constraint of purifying selection, causal and functionally deleterious variants are often RVs. In turn, they typically have larger effect sizes than common variants [Fu et al., 2013, Bansal et al., 2010, Sham and Purcell, 2014, McCarthy et al., 2008]. Therefore, developing new association tests tailored to low MAF SNVs and RVs has been a very active research area in recent years. Due to the nature of low MAF, either increasing the total sample size or aggregating information across multiple variants in an analysis set (for example gene) is expected to achieve a practically acceptable power [Capanu et al., 2011, Basu and Pan, 2011, Bansal et al., 2010, Sham and Purcell, 2014].

As increasing the sample size is usually expensive and demanding, SNP-set or gene-set based association tests pooling together information have been the major research directions [Ye and Engelman, 2011, Pinto et al., 2010, Sham and Purcell, 2014]. Sets of SNVs can be defined by gene boundaries (i.e., gene-based) or sliding windows; sets of genes can be defined by Gene Ontology terms, protein-protein interactions, canonical genetic signaling pathways or gene expression networks as examples. [Sham and Purcell, 2014, De la Cruz et al., 2010, Weng et al., 2011, Wang et al., 2010, Wei et al., 2012].

1.1.1 Gene-based association tests

A large number of gene-based association tests (mainly designed for RVs) have been proposed in recent years. The earliest methods include the cohort allelic sums test (CAST)[Morgenthaler and Thilly, 2007] and the combined multivariate and collapsing (CMC) method [Li and Leal, 2008]. Afterward, more advanced tests were proposed. Those methods can be classified into major groups as follows.

A very famous category of these methods is the so-called "burden test" or "sum test", such as a weighted sum statistic (WSS) [Madsen and Browning, 2009], which uses MAF based weighting scheme to combine the test statistics from multiple SNVs in a region, with the assumption that all the alleles to be deleterious. WSS is also known as Madsen and Browning test (MB test). Many other tests within the "burden test" category inherited and improved the WSS performance in some scenarios [Hoffmann et al., 2010, Zhang et al., 2010b, Ionita-Laza et al., 2011, Feng et al., 2011]. Such improved "burden tests" includes the sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS) [Zhu et al., 2010, Feng et al., 2011]; the replication-based test (RBT) [Ionita-Laza et al., 2011] which is built on WSS with the aim to be less sensitive to the presence of both risk and protective effects in a genetic region of interest; the yet another weighted-sum test with a "step-

up" approach to choose the "best" combination of rare variants into a single aggregated group [Hoffmann et al., 2010]; the MB test with approximately optimal collapsing (AOC) method [Zhang et al., 2010b]; a data-driven P-value Weighted Sum Test (PWST) [Zhang et al., 2011] which used both significance and direction of individual variant effect from single-variant analysis to calculate a single weighted sum score.

Another major category of gene-based association tests is the so-called "variance-component test", which can be formulated as testing on a variance component in a random-effects (R-E) model. These tests include the Sum of Squared U-statistics test (SSU) [Pan, 2009], which is equivalent to a variance-component test; the C-alpha test [Neale et al., 2011], which handles RVs with mixed effect directions well but is not able to adjust for co-variates (such as the principal components used to correct for population stratification confounders); the kernel machine regression (KMR) method [Wu et al., 2010, Kwee et al., 2008], which provides the flexibility of choosing different kernel functions $h(.)$ to measure the genomic similarity between the genotypes of subject i and j. It then regresses response on the specified kernel functions (if linear kernel, it is equivalent to the SSU test [Pan, 2009]); the widely used sequence kernel association test (SKAT)[Wu et al., 2011], which up-weights the SNVs with lower MAFs and assumes the effect of variants are independently and identically distributed with an arbitrary distribution of mean 0 and variance τ^2; the SKAT-O [Lee et al., 2012b, Lee et al., 2012a], which is a weighted linear combination of a burden test and the SKAT variance component test; the adjusted-SKAT [Oualkacha et al., 2013], which allows the variant effects to have an equal correlation ρ besides the usual assumption in SKAT; the GEE-based linear kernel machine SNP set association test [Wang et al., 2013] which is very closed to the SSU test.

The collapsing-based test inherited the idea from CMC/CAST method, and this type of tests is actually closely related to "sum test". Here are a few most representative methods: the RARECOVER algorithm [Bhatia et al., 2010], which is a model-free method, col-

lapses only a subset of the variants in a region to achieve the strongest association with a phenotype; the kernel-based adaptive cluster (KBAC) method [Liu and Leal, 2010], comparing the difference of weighted multi-site genotype frequencies between cases and controls; the rare variant weighted aggregate statistic (RWAS) method [Sul et al., 2011], which groups rare variants and computes a weighted sum of differences between case and control mutation counts.

Lasso and group-penalized regression based methods incorporated a mixture of group Euclidean penalties and single-predictor penalties (lasso) into linear or logistic regression [Zhou et al., 2010, Kim et al., 2014]. Group penalties are applied to SNVs within a single gene or within several genes in a pathway, while single-predictor penalties are applied at the single SNV level. The authors developed the coordinate descent algorithms that allow exceptionally fast computation and permit the optimal tuning of the penalty constant by cross-validation method.

Functional linear models (FLM) and (smoothed) functional principal component analysis (FPCA) based association tests [Luo et al., 2011, Luo et al., 2012b, Luo et al., 2012a, Fan et al., 2013] treated a chromosome as a continuum, on which variants identified from next-generation sequencing platform approximately evenly distributed. For FLM methods, the authors incorporated the genomic position t into the penalized regression equation for both genotype function and coefficient function, and then used basis function expansion method to solve the FLM. For FPCA methods, the authors incorporated the genomic position t into the eigenfunction for both genotype function and weight function, and then used either discretization method or basis function expansion method to solve the eigenfunction. If smoothing was used, the smoothing parameter λ was chosen by cross validation. The statistics from both FLM and FPCA follow the central χ^2 distribution.

Adaptive or hybrid tests combined the advantages from at least two major categories

above to make the new test more data adaptive and more powerful. A few most representative methods are as follows: The EREC method [Lin and Tang, 2011] builds a general framework for association testing, which combines strength from MB test and variable-threshold (VT) test [Price et al., 2010] to form the most powerful test by setting the weight function ϵ proportional to the set of estimated regression coefficients $\hat{\beta}$ in the test statistic. A data adaptive test combines the score test, SSU test and Sum test's advantages [Pan and Shen, 2011]. An exponential combination (EC) framework for set-based association tests [Chen et al., 2012] features with the sum of exponential statistics (statistics should follow either independent normal or independent chi-square distribution). The sum of exponential statistics are parametric and standardized from previous MB test and C-alpha test. A robust and powerful test uses Fisher's method to combine linear and quadratic statistics [Derkach et al., 2013]. A unified mixed-effect model [Sun et al., 2013] tests both group effect equal to 0 and variance component equal to 0. It includes both burden and SKAT tests as special cases by embedding the variant functional information and allowing a variant specific random effect in the model.

There are other miscellaneous tests. Some of them can be classified into more than one category mentioned above, thus I include them here as well as other miscellaneous tests. A variable-threshold (VT test) approach [Price et al., 2010] computes z-score $z(T)$ for each different MAF threshold T, defines z_{max} as the maximum z-score across values of $z(T)$, and finally assesses the statistical significance of z_{max} by permutations on phenotypes. A data-adaptive sum test (aSum) is capable of handling both deleterious and protective effects and allowing collapsing common variants (CV) into the test [Han and Pan, 2010]. A probabilistic disease-gene finder employs an aggregative variant association test that combines both amino acid substitution and allele frequencies as implemented in VAAST [Yandell et al., 2011] and the later improved version VAAST 2 [Hu et al., 2013]. The weighted score test [Cai et al., 2012] up- or down-weights the contribution from each

member of the marker-set based on the Z-scores of their effects.

For a detailed comparison among and discussion of some of these tests, Basu and Pan have done a very comprehensive review and simulation-based benchmark [Basu and Pan, 2011]. Another comprehensive review can be found in [Bansal et al., 2010]. Recently Pan et al. also did a comparison of several latest methods including the PWST, EREC, aSSU, SKAT-O and their newly proposed aSPU method [Pan et al., 2014].

Due to the complexity of genetics association with a phenotype, for example, specific association effect direction and size, a given test favoring one scenario may or may not perform well in other scenarios [Pan, 2009, Derkach et al., 2013, Pan et al., 2014, Sun et al., 2013]. In other words, there is no single test that is the most powerful in all testing scenarios. Therefore, there have been a lot of efforts in developing adaptive/hybrid tests for RVs (for example, [Derkach et al., 2013, Chen et al., 2012, Han and Pan, 2010, Lee et al., 2012a, Lin and Tang, 2011, Pan and Shen, 2011, Sun et al., 2013, Zhang et al., 2011]). However, due to limited adaptability, for example, with a fixed set or pre-determined weights on individual RVs, these tests that combined some earlier tests' advantages (for example, the MB test, burden test and SKAT) are still not flexible enough to avoid power loss under some situations. Recently, a very prominent novel data adaptive test named aSPU has been proposed by [Pan et al., 2014]. It features the ability to achieve quasi-optimal power in all scenarios, such as varying number of SNVs within the genetic region of interest, varying ratio of signal SNVs to noise ones, same directional effect alleles or a mixed directional effect of both protective and deleterious alleles, varying allele frequencies and varying effect sizes. It maintains the most power as compared to other state-of-the-art tests in the presence of a large number of RVs within a genetic region that only contains a small portion of signals. This is usually the case in association studies based on whole-exome or whole-genome sequencing data [Pan et al., 2014]. In summary, the data-adaptive test is in general more powerful and

robust than non-adaptive tests, and thus preferred in the future development of novel association tests. Among current data-adaptive association tests, the aSPU method is more adaptive than its predecessors. Hereby, I propose to extend the aSPU framework from the cross-sectional data scenario to the longitudinal data scenario. I also propose a few aSPU "variant" tests within the aSPU tests family. These aSPU 'variant' tests combine strength from the Score test and hence they are more robust in maintaining a higher power in almost all scenarios.

1.1.2 Longitudinal study design and analysis strategy in GWAS

Comparison between longitudinal studies and cross-sectional studies

I first introduce two linear models for cross-sectional studies and longitudinal studies respectively. In a cross-sectional study ($n_i = 1$) we are restricted to the model

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

where Y represents the quantitative trait, x represents the covariate, i represents the ith subject, j represents the jth measurement, n_i represents total measurement number for ith subject. Therefore, Y_{i1} represents the ith subject's trait measured at baseline while x_{i1} represents the ith subject's covariate measured at baseline. Furthermore, 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

$$Y_{ij} = \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1}) + \epsilon_{ij}, \quad i = 1, \ldots, m;\ j = 1, \ldots, n_i \tag{2}$$

[WARE et al., 1990]. Now β_C still represents the time-averaged cross-sectional difference while β_L is interpreted as the expected change in Y over time per unit change in x for a

given subject. The inference about β_C is a comparison of individuals with a particular value of x to other individuals with a different value of x at baseline. In contrast, the parameter estimation of β_L is by comparing a person's responses at two times, assuming that a person's x changes with time.

Based on above formula, we can better explain the merits of longitudinal studies over cross-sectional studies. Longitudinal studies allow us to estimate both the cross-sectional difference (β_C) and the rate change over time (β_L), the latter of which cross-sectional studies cannot estimate. Even when $\beta_C = \beta_L$, that is only time-averaged cross-sectional difference exists, 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. For most outcomes Y, there is considerable variability across individuals due to the influence of unmeasured characteristics such as genetic make-up, environmental exposures, personal behaviors/habits, and so forth. While these factors tend to persist over time for the same individual, their influences are canceled in the estimation of the β_L or equivalently here the β_C, and thus lead to more accurate estimate (with smaller variance). Another merit of the longitudinal study is its ability to distinguish the degree of variation in Y across time for one subject from the variation in Y across subjects. With repeated measurements, we can borrow strength across time for the same person of interest as well as across individuals. If there is little variation across subjects, one subject's estimate can rely on data from others as in the cross-sectional case. However, if the variation across individuals is large, we might prefer to use more data for the same individual across time. Last but not least, with longitudinal studies, we can estimate a person's current and future outcomes.

In my dissertation, I will mainly study the scenario under $\beta_C = \beta_L$ for longitudinal association test. I assume the SNPs in a region contribute to the outcome Y as the main effect only and the fixed effect remains the same across time ($\beta_C = \beta_L$). There is more

to explain here about the efficiency of the longitudinal study. Let $e = Var(\hat{\beta}_L)/Var(\hat{\beta}_C)$ be the specific measure of efficiency. Apparently, the smaller the value of e, the greater is the information gained by taking additional measurements across time on each person. The value of e depends on a number of factors, including correlation structure R (for example, compound symmetry or auto-regression), number of measurements (n_i), magnitude of within-subject correlation (ρ) and the ratio δ of within-subject variation in x to between-subjects variation in x at baseline. In general, increasing n_i (for example, more measurements) and increasing δ (for example, uneven measurement intervals) will lead to a smaller e under the scenario $\beta_C = \beta_L$. Besides, except when δ is small and ρ is high at the same time, there is much to be gained by conducting longitudinal studies even when the number of repeated observations n_i is as small as two according to [Diggle et al., 2002].

Under different hypothesis testing scenarios, the identified significant signal loci from a longitudinal study may be the **same** or **different** from a comparable cross-sectional study. In the GWAS settings, the cross-sectional study always tests the SNP main effect (β_{main}), and this will equate the longitudinal study with **only time-averaged SNP main effect** (i.e., $\beta_C = \beta_L$ in Equation 2). However, when the longitudinal study includes **the additional SNP $\times$ time interaction term**, either joint testing both of the main effect and interaction effect equal to 0 or individual testing either effect equal to 0 will possibly lead to different significant loci from the corresponding cross-sectional study.

Factors that influence the statistical power in longitudinal studies

In any study, investigators must provide the following quantities to determine the power P, including the Type I error rate (α), smallest meaningful difference (d) to be detected, sample size (n), variance (σ^2) in response variable. In longitudinal studies, there are several additional factors to consider, including the number of repeated observations per

person (n_i) and the correlation among the repeated observations within the same person (ρ). Let us briefly describe the relationship between these quantities and the power P: increasing α will increase P; increasing d will increase P; increasing n will increase P; reducing σ^2 will increase P; increasing n_i will increase P. For ρ, the relationship with P is not fixed but depends on which hypothesis is tested. In the $\beta_C = \beta_L$ scenario, we are testing the time-averaged (group) main effect. An **decreasing** ρ will lead to a larger power. In contrast, in the $\beta_C \neq \beta_L$ scenario when we are testing the slope effect β_L, an **increasing** ρ will lead to a larger power in testing the $\beta_L = 0$, that is the rate change over time equal to 0. This, at the first glance, seems counter-intuitive but is indeed reasonable. In the $\beta_C = \beta_L$ scenario, the parameter of interest is the expected average of the Y's for individuals in a group (i.e., the β_C). A decreased ρ leads to an effectively larger sample size (within-subject measurements are more distinct), which in turn results in a smaller variance of β_C estimate. On the contrary, in the $\beta_C \neq \beta_L$ scenario when we are testing $\beta_L = 0$, the rate of the change in Y, the contribution from each subject to the estimation of β_L is a linear contrast of the Y_{ij}. The Y_{ij}'s variance is decreasing as ρ increases, i.e., within-subject measurements are more alike. Thus, an increasing ρ will lead to a larger power of testing the significant deviation from $\beta_L = 0$.

Longitudinal studies in GWAS

Many GWASs have been performed in cohort studies, where phenotypes have been collected across multiple time points for each individual [Aulchenko et al., 2009, Ionita-Laza et al., 2007, Kamatani et al., 2010, Kathiresan et al., 2007, Sabatti et al., 2008]. However, the longitudinal information has not been fully utilized as the majority of the current association tests only considered either the baseline measurement or average measurement for each individual [Sabatti et al., 2008, Ionita-Laza et al., 2007, Kamatani et al., 2010, Kathiresan et al., 2007]. Compared with large number of GWASs, only very few studies involved longitudinal data analysis. One such study on smoking and nicotine de-

pendence by [Belsky et al., 2013] has data from a four-decade longitudinal study. They used generalized estimating equation model to analyze the longitudinal data while accounting for correlation within subject. There are also several GWASs on Alzheimer's Disease involving the analyses of longitudinal phenotypic information collected at multiple time points [Wang et al., 2012, Melville et al., 2012, Silver et al., 2012]. Increased power from longitudinal study has been elucidated herein before, and recently this fact has been studied in depth by either simulation study and/or real data analysis in the GWAS settings [Xu et al., 2014, Furlotte et al., 2012]. Depending on the specific parameters settings in simulation studies (for example, correlation among repeated measurements, genetic variance, and environmental variance) and case by case for real data analysis (for example, the sample size and the SNV effect size), the power gain from longitudinal data analysis as compared to baseline data analysis can range from a moderate to a significant amount [Xu et al., 2014, Furlotte et al., 2012]. For example, the increased power was demonstrated from 0 to 0.4 with the maximal possible power value being 1 ideally, or the longitudinal analysis has as much as an eightfold gain in power over the baseline analysis. Therefore, a longitudinal study design and performing the longitudinal data analysis when data are available are appealing in the GWAS settings.

Classical longitudinal data analysis methods

Existing methods in longitudinal data analysis can be categorized into three main categories: 1, mixed-effect models; 2, marginal models with regression coefficients estimated by generalized estimating equations (GEE); 3, transition (Markov) models.

The mixed-effect model was first proposed in 1982 [Laird and Ware, 1982]. Mixed-effect model is an extension of a regression model to model longitudinal (correlated) data. It contains fixed effects and random effects, where random effects are subject-specific and are used to model between-subject variation and the correlation induced by this variation. Mixed-effect model is a two-stage method, which treats 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. Parameter estimation is usually done by restricted maximum likelihood (REML) and expectation-maximization (EM) algorithm [Laird and Ware, 1982].

Another major method, the marginal model with GEE was first proposed in 1986 [Zeger and Liang, 1986, Liang and Zeger, 1986]. It is an extension of the quasi-likelihood methods by Wedderburn [Wedderburn, 1974]. Rather than giving subject-specific (SS) estimates as in mixed-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 of the outcome variable as a known function of the marginal expectation. By specifying a "working" correlation matrix, GEE method accounts for the correlation among the repeated observations for a given subject. Another appealing property of GEE is, by using the so-called sandwich variance estimator, the "working" correlation matrix does not need to be correctly specified in order to achieve consistent estimates. The generalized estimating equations are thus derived without specifying the joint likelihood function of a subject's observations as needed in the SS model. The covariance structure across time is treated as a nuisance parameter. GEE can finally give consistent estimators of the regression coefficients by simply solving the score equations and doing iteratively reweighted linear regression.

The last major method, 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. With different link functions, Markov models can be applied to a range of GLMs as dealt in mixed models and marginal models. A few examples are linear link [Tsay, 1984], logit link [Cox and Snell, 1989, Zeger et al., 1985, Korn and Whittemore, 1979] and log link [Zeger and Qaqish,

1988]. Model fitting is straightforward for linear link, as in Gaussian autoregressive models, the full maximum likelihood estimation is available [Tsay, 1984]. For logistic and log-linear cases, the full likelihood is unavailable and the alternative is to maximize the conditional likelihood with GEE-like iterative weighted least square algorithm to solve the conditional score function and obtain consistent estimates [Cox and Snell, 1989, Zeger et al., 1985, Korn and Whittemore, 1979, Zeger and Qaqish, 1988].

Since transitional models are not widely used in the genetics association studies, we will omit its further discussion. Here we focus on the comparison between the mixed-effect and marginal-effect models. The application of GEE may be less appropriate when the time course of the response variable for each individual (subject-specific slope), for example, body mass index (BMI) measurements across several time points, is of primary interest, so as to the correlation parameters within the same subject [Zeger et al., 1988, Liang and Zeger, 1986]. The mixed-effect model could handle such inference of interests [Laird and Ware, 1982]. However, in the genetic association studies, subject-specific time course effect and/or within-subject correlation parameters are usually not of major interests. In other words, they are often treated as nuisance parameters. On the other hand, for association tests of a set of SNPs, the increased number of explanatory variables, for example, the SNVs on the right hand side (RHS) of the regression-like equation tend to consume a lot of degrees of freedoms (dfs) and increase the difficulty of the optimization algorithm convergence. Large consumption of the dfs also leads to power loss and Type I error inflation, for example, excessive inflation in the Wald test [Guo et al., 2005, Pan, 2001, Shete et al., 2004]. Algorithm convergence difficulty is very often encountered in mixed model when it has a lot of covariates. For some extreme scenarios, for example, with a binary trait, the maximum likelihood estimator (MLE) of a regression coefficient of a RV does not exist if the minor alleles of this RV only appear in case or control, resulting in the convergence failure with an iterative

algorithm to obtain the MLE [Zhang et al., 2014, Pan et al., 2014]. Another caveat of the mixed model is that mis-specification of the random-effects distribution and/or omitting part of the random-effects (for example, keeping only the random intercept in the mixed model when the random slope is also needed) will lead to excessive Type I error inflation [Litière et al., 2007, Xu et al., 2014]. Compared with the mixed-effect models, GEE models suffer much less from these problems. In particular, the GEE Score test is proved to be robust to Type I error inflation in the presence of a large number of covariates; upon usage of the so-called sandwich or robust covariance matrix, GEE will give consistent estimates even when the working correlation is misspecified, in contrast to the misspecified random effect in the mixed-effect models. In addition, GEE model fitting requires only evaluation under the null hypothesis, which greatly accelerates the computation. With regard to the power loss in the presence of an increased number of covariates (SNVs), a recently proposed data-adaptive association test within the GEE framework demonstrated convincing capability in maintaining a high power [Zhang et al., 2014, Pan et al., 2014]. Although this work is designed for a single cross-sectional trait or multiple cross-sectional traits, it can be extended to the longitudinal design scenario as detailed in Aim One below. Extending the gene-based association test to sets of multiple related genes could return more biological meaningful inference.

1.1.3 Gene-set/Pathway based association tests

In general, a gene set is a set of genes. A genetic pathway is a special gene set, which includes multiple genes interacting to form an aggregate biological function. Extending the gene-based association tests to sets of related genes could lead to more biological meaningful inference. By analyzing functional related genes together with the phenotype of interest, we are more likely to identify those signals hidden from or attenuated in single-gene based tests [for Blood Pressure Genome-Wide Association Studies et al.,

2011, Hirschhorn, 2009, Zhong et al., 2010, Wang et al., 2010]. Complex diseases are known to be influenced by a combination of genetic factors in addition to environmental factors, lifestyle factors, and their interactions [Hirschhorn and Daly, 2005, McCarthy et al., 2008]. Thus, by investigating the gene sets, more evidence contributing to a specific disease could be found. Another advantage of pathway-based association tests is similar to that of gene-based association tests: aggregating multiple genes/RVs, in contrast to testing each gene/RV separately, may boost the statistical power by combining moderate/weak signals. One convincing evidence is from the Cancer Genome Atlas project (TCGA: http://cancergenome.nih.gov/) in tumor sequencing studies. While only few oncogenes (for example, TP53 and EGFR) harbor many mutations, most others harbor few mutations in a tumor-dependent manner. Single gene-based association tests still suffer from low aggregated mutation frequency, whereas collectively, they have a much higher aggregated mutation frequency in a gene-set/pathway. Therefore, for some diseases such as cancer, a gene-set/pathway analysis by aggregating the somatic mutation information across genes will boost the statistical power, and is thus preferred.

Among association tests for sets of functional related genes, pathway-based association test is probably the most popular one [De la Cruz et al., 2010, Wang et al., 2010]. Other types include Gene Ontology terms, protein-protein interaction, canonical genetic signaling pathways, gene expression networks as examples [Sham and Purcell, 2014, De la Cruz et al., 2010, Weng et al., 2011, Wang et al., 2010]. A "pathway" in the GWAS setting usually means a set of genes involved in the same biological function or process, for example, apoptosis. Some commonly used public pathway and gene-set databases/repositories include Kyoto Encyclopedia of Genes and Genomes (KEGG) [Ogata et al., 1999], BioCarta [Nishimura, 2001] and Gene Ontology [Ashburner et al., 2000]. KEGG and BioCarta provide manually curated pathways in different biological processes, whereas Gene Ontology mainly contains computational annotations for human genes. Several commer-

cialized databases are also available including Ingenuity Pathway Analysis (IPA) and MetaCore from GeneGo. They combine the manually curated evidence, literature review and algorithm predicted results. There are also other specialized pathway databases, such as Science Signal Transduction Knowledge Environment [Gough, 2002] and Nature Pathway Interaction Database [Schaefer et al., 2009], both of which manually curated the cell signaling pathways; the MetaCyc database [Karp et al., 2002] and BioCyc database [Caspi et al., 2008], both of which contain metabolic pathways. In summary, there are abundant existing biological pathway databases, which will facilitate the pathway-based association study analysis.

Major classes of pathway-based association tests

Depending on the null hypothesis to be tested, pathway-based association tests can be categorized into two major classes: self-contained approach and competitive approach [Goeman and Bühlmann, 2007, Liu et al., 2007, Nam and Kim, 2008, Wang et al., 2010, Fridley et al., 2010, Fridley and Biernacka, 2011]. Self-contained (a.k.a. Constrained) approach hypothesizes there is no gene in the gene set associated with the phenotype, while competitive approach hypothesizes the same level of association of a gene set with the given phenotype as the complement of the gene set. Figure 1 illustrates the hypothesis testing difference between the two classes of approaches.

Competitive methods usually start with identifying SNPs/genes that are significantly associated with a phenotype, and then evaluate whether the significantly associated SNPs tend to be enriched in a predefined gene-set/pathway. These methods are called 'competitive' because they compare the frequency of significantly associated SNPs in a particular set of genes/pathway with the frequency of significant associations among all genes not in the set [Fridley and Biernacka, 2011]. Representatives of competitive approach are gene set enrichment analysis (GSEA) [Subramanian et al., 2005], which is based on the Kolmogorov-Smirnov test and DAVID [Dennis Jr et al., 2003], which uses a modified

Fisher's exact test.

In contrast, self-contained approach considers the null hypothesis that no SNPs/genes in the gene-set of interest are associated with the trait versus the alternative hypothesis that some SNPs/genes in the gene-set are associated with the trait. Methods in the self-contained class are more flexible. Their statistical significance can be assessed (1) by the deviation from the expected number of significant SNPs under the null hypothesis of no association between the phenotype and the gene-set/pathway, (2) by computing the association p-values for each SNP in a gene-set/pathway, followed by testing whether the difference between the observed distribution of the SNP-level p-values and the expected distribution under the null hypothesis is significant, (3) by modeling the effect of gene via aggregating multiple SNPs, followed by modeling the effect of gene-set via aggregating multiple relevant genes, or (4) by directly modeling the effect of gene-set by aggregating the SNPs within the gene-set, skipping the gene-level statistics.

A. Competitive Approach

Example A:

	Significant	Not Significant	
SNP in gene set *G*	20	80	100
SNP outside gene set *G*	100	400	500
	120	480	600 SNPs

- 20% of SNPs within *G* significant
- 20% of SNPs outside of *G* significant
- P = 0.55 for Fisher's exact test of the competitive hypothesis
- No evidence of enrichment

Example B:

	Significant	Not Significant	
SNP in gene set *G*	40	60	100
SNP outside gene set *G*	100	400	500
	140	460	600 SNPs

- 40% of SNPs within *G* significant
- 20% of SNPs outside *G* significant
- P < 0.001 for Fisher's exact test of the competitive hypothesis
- Evidence of enrichment

**Note that the statistical test applied here assumes independence of p-values, which is an invalid assumption in the presence of LD. Here this simple test is only used to illustrate the competitive hypothesis.*

B. Self-contained Approach

Number of SNPs in gene set *G* significant with p < 0.05		
	Significant	Not Significant
Observed	20	80
Expected	5	95

- 20% of SNPs within *G* significant.
- Under the null hypothesis, expect 5% of the SNPs to be significant.
- P = 0.002 for Fisher's exact test of the self-contained hypothesis.
- Evidence of association of the gene set with the trait.

**Note that the statistical test applied here assumes independence of p-values, which is an invalid assumption in the presence of LD. Here this simple test is only used to illustrate the self-contained hypothesis.*

Figure 1: **Examples of competitive approach and self-contained approach based testings using Fisher's exact test as a demonstration** (A). Example of competitive approach; (B). Example of self-contained approach. This figure is adopted from [Fridley and Biernacka, 2011].

The self-contained approaches have a few edges over the competitive approach. A limitation of the competitive approaches is that they cannot be applied to studies of candidate gene-sets for which only SNPs in the candidate gene-sets have been genotyped but not

in the complemented ones. The reason is straightforward: competitive approaches require a comparison between many different pathways. On the other hand, self-contained approaches require only genotypes from a collection of candidate genes, which in turn enable the genome-wide studies, candidate gene studies, pathway studies or specific disease gene group studies. Specific disease gene group studies are very popular, for example, the cardiovascular diseases, the metabolic traits and the autoimmune diseases. These studies usually come with the disease-specific genotyping platforms, for example, the ImmunoChip [Cortes and Brown, 2011], the metabochip [Voight et al., 2012] and the CVD35/cardiovascular-IBC-array [Cheng et al., 1999, Keating et al., 2008]. The self-contained approaches have also been reported to be more powerful than the competitive approaches [Goeman and Bühlmann, 2007]. This follows immediately from the fact that the self-contained null hypothesis is more restrictive than the competitive null hypothesis, as noted before. As a result, a self-contained test will almost always reject the null hypothesis for more gene-sets than a competitive null. Nevertheless, some drawbacks of the self-contained approach have been reported, for example, the global inflation of test statistics is often observed or not adequately adjusted, finally leading to an inflated Type I error [Wang et al., 2007, Goeman and Bühlmann, 2007, Fridley et al., 2010].

Additionally, based on the input data type, the pathway-based tests can be broadly classified into two categories: those that require raw genotypes and those that require a list of SNP p-values. The first approach, 'raw genotype approach', requires raw SNP genotypes as input to derive gene-level and pathway-level test statistics, whereas the second approach, 'p-value enrichment approach', requires a list of pre-calculated SNP p-values to determine whether a specific group of p-values for SNPs (or genes) is enriched with association signals. The 'p-value enrichment approach' only requires pre-computed SNP pvalues and it greatly saves the labor in coordinating data analysis and data sharing, however, the 'raw genotype approach' provides more flexible solutions such as multi-SNP

tests which requires individual genotype data to derive gene-level test statistics (some of these methods pool all the SNPs in a pathway together without calculating test statistics for pathway gene members). Another example is those methods based on single-SNP p-values but require raw genotype data to execute phenotype permutation-based test. In this way, those methods can come up with a more unbiased pathway enrichment score. The 'raw genotype approach' is also less biased, such as it can adjust for gene length, the distance threshold to assign SNPs to nearby genes and the way to summarize gene-level test statistics. The graphic demonstration of the method categorization is shown in Figure 2.

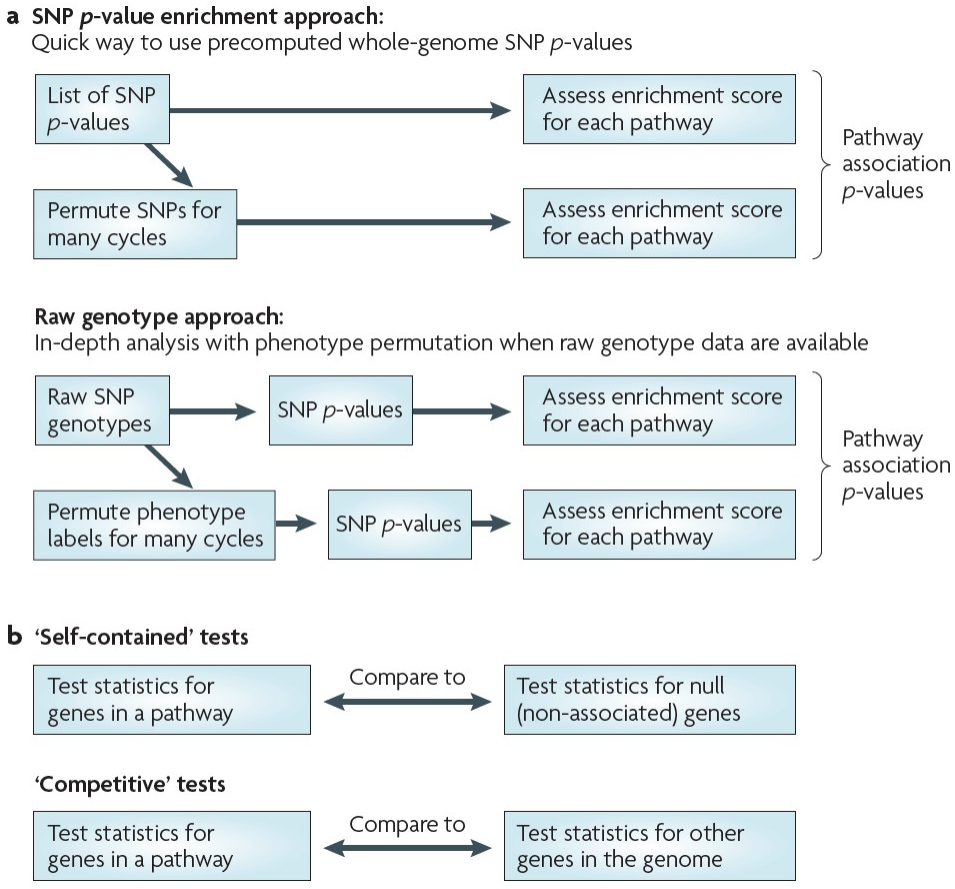

Figure 2: **Types of pathway association tests in GWAS.** (a). Categorization based on data input type; (b). Categorization based on hypothesis testing. This figure is adopted from Wang et al (2010) [Wang et al., 2010].

Existing pathway-based association tests

There are several popular existing methods for pathway-based association tests. The earliest method is the gene-set enrichment analysis (GSEA) algorithm, a method adapted for pathway-based analysis of GWAS data. It calculates a weighted Kolmogorov-Smirnov-like running-sum statistics and uses a permutation-based procedure to evaluate the statistical significance [Wang et al., 2007]. The GSEA-SNP, a modification of Wang et al's GSEA [Wang et al., 2007], uses the max-test and all SNPs in a gene [Holden et al., 2008]. The i-GSEA4GWAS, first permutes the SNP labels (for example, the SNP rs id), then assigns SNPs to genes, and finally calculates the modified GSEA enrichment score[Zhang et al., 2010a]. The Gene Set Analysis (GSA-SNP), computes the gene-level test statistic based on the SNP with the minimum P-value (or the second minimal), followed by gene-set-level test using either a Z-test, maxmean test, or GSEA [Nam et al., 2010]. The Gene Set-based Analysis of Polymorphisms (GeSBAP), first calculates enrichment score using ranked gene list, then assigns the best SNP p-value to a gene, and finally uses Fisher's exact test for the gene-set association [Medina et al., 2009]. De la Cruz et al (2010) proposed A modification of Fisher's method for combing SNP P-values for gene-level or gene-set-level association [De la Cruz et al., 2010]. The gene set ridge regression in association studies (GRASS), 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 the shrinkage of gene effect (for example, disease odds ratio) estimates simultaneously [Chen et al., 2010a]. PLINK, a widely used software package in GWAS data analysis, provides an option to execute gene-set-level association analysis [Purcell et al., 2007]. The association list go annotator (ALIGATOR) method, a 'p-value enrichment approach' requiring only pre-computed SNP p-values, uses Fisher's exact test on SNP with the minimum p-value for the gene-level association. It can correct for linkage disequilibrium (LD) between SNPs, various gene size, and multiple testing of

nonindependent pathways [Holmans et al., 2009]. The SNP ratio test (SRT) method, tests the ratio of significant SNPs in a pathway and computes the empirical p-value based on permutation [O'Dushlaine et al., 2009]. The supervised principal component analysis, uses the Gumbel extreme value mixture distribution as test statistic's null distribution. The test statistic is standardized for pathway size using a simulation procedure [Chen et al., 2010b]. The Prioritizing Risk Pathways fusing SNPs and pathways (PRP) method, executes the gene-level association test based on maximum risk statistic, followed by mean risk approach to obtain gene-set-level risk statistic, then weights this statistic by specific pathway degree (that is, total edges in a pathway) and standardizes it to a zero dimension (that is, the minimal value is 0) [Chen et al., 2009]. Luo et al (2010) proposed three statistics to combine a set of dependent p-values of SNPs into an overall significance level for a gene, and then combined a set of dependent p-values of genes into an overall significance level for a pathway. The three statistics, which take into account the LD among SNPs or correlation among genes in the specific pathway, are linear combination test (LCT) asymptotically following normal distribution under null hypothesis, Quadratic test (QT) asymptotically following central Chi-square distribution under null hypothesis, and decorrelateion test (DT) combineing decorrelated individual statistics by Fisher's combination test and asymptotically following a central Chi-square distribution under null hypothesis [Luo et al., 2010]. In addition, Peng et al (2009) developed four methods to combine a list of SNP p-values or gene-level p-values with the assumption that individual SNPs/genes are independent. These four methods are Fisher's, Sidak's, Simes' and the FDR method [Peng et al., 2009]. The Gene-loci Set Analysis (GLOSSI) method, first uses the Cochran-Armitage trend test at single-SNP level assuming an additive SNP effect, then uses Fisher's combination test to combine individual p-values of SNPs, and finally corrects the test statistics by Brown's approximation to better control the Type I error [Chai et al., 2009]. An adaptive rank truncated product (ARTP) statis-

tic, combines permutation-based SNP-level p-value to derive gene-level significance level and/or combines gene-level p-values to derive pathway-level significance level [Yu et al., 2009]. Detailed reviews about these and other pathway-based association tests can be found in [Tintle et al., 2011, Wang et al., 2010, Fridley and Biernacka, 2011].

1.2 Public Health Significance

The majority of human diseases are complex diseases, for example, cardiovascular disease, type 2 diabetes, Alzheimer's disease and autoimmune disease. These diseases have high incidence rate in the US and worldwide [Craig et al., 2008, Inzucchi et al., 2012, Go et al., 2013, Association et al., 2014, Benros et al., 2014]. The development of complex diseases involves genetic factors, environmental factors, behavior factors, and the interactions among them. In public health research, identification of the causal factors and the heritability of complex disease has always been a frontier topic. Researchers often first look for the genetic factors, followed by gene-gene and gene-environment interaction analyses. The GWASs have already identified more than 1000 genetic loci associated with many human diseases and traits [Hindorff et al., 2009]. These genetic loci have been validated by some validation procedures, such as replicate studies, meta analysis and wet lab experiments [Wang et al., 2005, Hirschhorn and Daly, 2005, McCarthy et al., 2008, Hindorff et al., 2009, Cantor et al., 2010].

The advent of the Next-Generation Sequencing (NGS) technique has brought human genetics to a new era [Ansorge, 2009, Metzker, 2009, Mardis, 2008, Shendure and Ji, 2008], and has the potential to explain some of the missing heritability via disease/trait-associated rare variants [Eichler et al., 2010]. Researchers have delivered tremendous efforts in developing powerful association tests either for common variants or rare variants, in gene-based and/or pathway-based manner as aforementioned. These tests are

mainly designed for cross-sectional data analysis, which utilizes less information and is thus less powerful than longitudinal data analysis. Although some of the existing methods have the potential to be extended for longitudinal data scenario, the work has not been done yet. As the association pattern between variants and disease/trait is subtle and largely unpredictable, more and more novel "data-adaptive" association tests have been developed. The so-called "data-adaptive" test can maintain a high power for various real world data sets. In the dissertation, I developed statistical methods that provided researchers with a powerful and robust data-adaptive association test for either common variants (CVs) or rare variants (RVs) in the longitudinal data settings, grouping variants in a gene-based or pathway-based way. Furthermore, I developed an independent Linux command-line based software implementing these methods, which will greatly facilitate the research community to apply the new methods on real data analysis. In conclusion, my dissertation work will provide useful methods and tools to identify the underlying genetic factors and explain the heritability of human complex diseases; in the long run, it may contribute to the prevention, diagnosis and cure of complex diseases.

1.3 Specific Aims

As reviewed before, current association testing methods are mainly designed for cross-sectional data analysis, while many cohort studies have the longitudinal measurements which have not been fully utilized. For instance, the Atherosclerosis Risk in Communities (ARIC) study [Heiss, 1989] has multiple follow-up measurements across almost 30 years. Association tests that fully utilize the information across time points tend to achieve a higher power and identify more disease-associated loci [Furlotte et al., 2012, Xu et al., 2014]. As both the common variants (CV) and rare variants (RV) are important in identifying the disease attributing genetic factors, a well-rounded association test should

have the flexibility to work with either of them. It should also maintain a relatively high power in almost all scenarios encountered in real data analysis. In practice, it is hard to predict which specific method has the highest power and which methods suffer from a large power loss for a specific data set. To meet these urgent needs, I propose a powerful data-adaptive SNP-set based association test for the longitudinal data analysis, applicable to either CVs or RVs. The specific aims are as follows.

Aim 1: To develop data-adaptive SNP-set based association tests for longitudinal data analysis within the Generalized Estimating Equations (GEE) framework. The proposed test, namely aSPU test, will have relatively high power in most data scenarios and avoid drastic power loss in any single data scenario, as compared to existing methods. This will be the first data-adaptive association test method for longitudinal data analysis. There are two **sub-aims**: **1(a).** model development for common variants; **1(b).** model development for rare variants.

Aim 2: To develop the pathway-based data-adaptive association tests for longitudinal data analysis. I propose to extend the proposed method in Aim 1 to the pathway-based association test version, namely Path-aSPU. It will work for either common variants or rare variants in a gene-set/pathway-based manner. Currently, there are no statistical models designed for pathway-based association test in longitudinal data settings, let alone the data-adaptive merit.

Aim 3: To develop the software package for method implementation. I will provide an R package or a Linux command-line based software program to enable convenient implementation of proposed methods. The package/software will be released to public (for example, CRAN) eventually.

2 Methods

2.1 Overall Study Design

2.1.1 Simulation studies

I will first generate the simulated dataset for testing the performance of our proposed novel methods in Aim 1(a), Aim 1(b) and Aim 2. Specifically, I will generate the simulated genotypes within a gene, that is simulated CVs for Aim 1(a) and simulated RVs for Aim 1(b) and Aim 2. I will also generate the simulated longitudinal phenotype mimicking the real data, the ARIC cohort data, used for Aim 1 and 2. Additionally, for Aim 2, I will simulate the pathway containing multiple genes for testing the pathway-based association. For all simulations, I will refer to several literature [Pan et al., 2014, Basu and Pan, 2011, Pan and Shen, 2011, Han and Pan, 2010, Pan, 2009] and the ARIC data to set up the simulation parameters, thereby the simulated dataset will be more close to the real dataset. In the simulation studies for Aim 1 and 2, I will evaluate the proposed methods' performance on maintaining the nominal Type I error and a higher empirical power under different simulation scenarios. I will compare the proposed tests to a few existing methods to demonstrate the advantages.

2.1.2 Real data application

After simulation studies, I will apply the proposed methods in Aim 1 and Aim 2 on the real dataset. A brief introduction to the real dataset is as follows.

The Atherosclerosis Risk in Communities Study (ARIC), sponsored by the National Heart, Lung, and Blood Institute (NHLBI), is a prospective epidemiological study conducted in four U.S. communities. ARIC is designed to investigate the causes of atheroscle-

rosis and its clinical outcomes, the variation in cardiovascular risk factors, medical care, and disease by race, gender, location, and time. ARIC data include two parts: the Cohort Component and the Community Surveillance Component.

The Cohort Component of the ARIC study, on which I will apply our proposed methods, began in 1987. Each of the four ARIC field centers (Washington County, MD; Forsyth County, NC; Jackson, MS; and Minneapolis, MN) randomly selected and recruited a cohort sample of approximately 4,000 individuals aged 45-64 from a defined population in their community. 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 examination (baseline) occurring in 1987-89, the second in 1990-92, the third in 1993-95, and the fourth exam was in 1996-98. The fifth examination is farther apart from the previous screens and was finished during 2011-2013. A detailed description of the ARIC study design and methods was published elsewhere [Investigators et al., 1989].

As for real data application purpose for both Aim 1 and Aim 2, I will exclusively use the Caucasian samples ($n = 11478$) in the ARIC cohort dataset. I will use lipid traits in the ARIC cohort as the response variables with longitudinal measurements. Candidate traits include four closely cardiovascular disease-related traits, which are total cholesterol (TC), High-density lipoprotein cholesterol(HDL-C), Low-density lipoprotein cholesterol (LDL-C) and triglycerides (TRG). For genotype data part, I will use the common variants for Aim 1(a) and the rare variants for Aim 1(b) and Aim 2. Both CVs and RVs are genotyped by ExomeChip platform [Grove et al., 2013] in the ARIC study. Through real data application, I will try to validate the reported risky loci for cardiovascular disease and identify potential novel loci, on a specific gene (Aim 1) or within a specific pathway (Aim 2). The method details for real data application is put below in Section 2.6.

A demographic introduction of the ARIC cohort data is shown in Figure 3

Cohort Characteristics

Characteristics of ARIC Cohort at Baseline by Sex or Race

Variable	WOMEN (n=8710) Percent or Mean (SD)	MEN (7082) Percent or Mean (SD)
White	69%	77%
Age 45-54	56%	49%
55-64	44%	51%
Family Income > $25,000	53%	67%
Glucose Diabetes (cut point=126)	12%	12%
Current Smoker	25%	28%
Usually Have Cough	12%	13%
Hypertension (140/90 or meds)	35%	35%
Rose Angina	6%	4%
Major Q-wave	0.3%	0.6%
Prior MI Reported	2%	8%
Ever Exercise	60%	66%
BMI (kg/m^2)	28(6)	27(4)
Cholesterol (mg/dl)	218(43)	211(40)
HDL Cholesterol (mg/dl)	57(17)	44(14)
Triglycerides (mg/dl)	124(82)	142(99)
Fibrinogen (mg/dl)	308(66)	298(65)
Factor VIIc	125(31)	112(26)

Variable	White (n=11478) Percent or Mean (SD)	Non-White (4314) Percent or Mean (SD)
Women	53%	62%
Age 45-54	51%	58%
55-64	49%	42%
Family Income > $25,000	72%	27%
Glucose Diabetes (cut point=126)	9%	20%
Current Smoker	25%	30%
Usually Have Cough	13%	11%
Hypertension (140/90 or meds)	27%	56%
Rose Angina	5%	4%
Major Q-wave	0.4%	0.4%
Prior MI Reported	5%	4%
Ever Exercise	70%	44%
BMI (kg/m^2)	27(5)	30(6)
Cholesterol (mg/dl)	215(41)	215(45)
HDL Cholesterol (mg/dl)	50(17)	55(18)
Triglycerides (mg/dl)	138(93)	114(80)
Fibrinogen (mg/dl)	298(62)	320(72)
Factor VIIc	119(29)	118(31)

Figure 3: ARIC Cohort Characteristics by Gender or Race. Table adopted from the ARIC website

2.2 Methods for Aim 1(a): To develop data-adaptive SNP-set based association tests for longitudinal data analysis within the Generalized Estimating Equations (GEE) framework for common variants

2.2.1 Statistical Modeling

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 for SNP $j = 1, \ldots, p$, and $z_i = (z_{i1}, z_{i2}, \ldots, z_{iq})$ is a row vector for q variates. We assume common effect sizes (a.k.a., time-averaged group effect) 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 $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ and $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_{q+1})'$ for X_i and Z_i respectively. The marginal mean of each measurement, $E(y_{im}|x_i, z_i) = \mu_{im}$ where $m = 1, 2, \ldots, k$ 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):

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

with $H_i = (Z_i, X_i), \theta = (\varphi', \beta')'$ and $g(.)$ 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 [Liang and Zeger, 1986]:

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

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

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

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. For convenience, a working independence model with $R_w = I$ is often used. 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:

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

$$\widetilde{\Sigma} = \widehat{\mathrm{Cov}}(U) = \sum_i (Z_i, X_i)' \widehat{\mathrm{var}(Y_i)} (Z_i, X_i) = \sum_i (Z_i, X_i)' (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' (Z_i, X_i) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \tag{3}$$

where $\hat{\mu}_i$ is an estimator of μ_i, $\widetilde{\Sigma}$ is an estimate of the covariance of score (U) vector. $\widetilde{\Sigma}$ is partitioned with the dimensions according to the score vector component $U_{.1}$ and $U_{.2}$ for φ and β respectively.

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:

$$\begin{aligned} U &= \sum_i (Z_i, X_i)' R_w^{-1} (Y_i - \mu_i) \\ \widetilde{\Sigma} &= \sum_i (Z_i, X_i)' R_w^{-1} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' R_w^{-1} (Z_i, X_i) \end{aligned} \tag{4}$$

if the assumption of a common covariance matrices across Y_i for i is valid, for example for quantitative continuous traits study [Pan, 2001], 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 [Pan, 2001].

In my dissertation, I will **focus on** the case with quantitative traits, since they are most typical traits used as response variable in the longitudinal data analysis. Nevertheless,

I introduce the strategy for binary traits below. In general, the only difference is the canonical link function, with all other equations/formulas keep the same.

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

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

Several Existing Association Tests

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 13.

- **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 even worse, the type I error tends to inflate in such case [Pan et al., 2014, Zhang et al., 2014].

- **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 [Pan et al., 2014, Zhang et al., 2014].

- **The UminP Test:** $T = \underset{j}{max}\frac{U_{.2,j}^2}{\Sigma_{.2,jj}}$ 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. An asymptotic multivariate normal distribution numerical integration based method provided a fast way to calculate its p-value [Pan et al., 2009, Pan, 2009]; alternatively, a simulation based method relying on the asymptotic normal distribution of the score vector can be used to calculate its p-value [Pan et al., 2014, Zhang et al., 2014]. Specifically, we first simulate the score vector $U_{(b)} = (U_{(b).1}, U_{(b).2}, \ldots, U_{(b).p})'$ from its null distribution $U_{(b)} \sim N(0, \Sigma_{.2})$ for $b = 1, 2, \ldots, B$, then calculate a total number of B null statistics: $T^{(b)} = \max_{j=1,\ldots,p} \frac{U^2_{(b).j}}{\Sigma_{.2,jj}}$, 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 $\frac{U^2_{.2,j}}{\Sigma_{.2,jj}}$ is equivalent to running the model on each single SNP (for example jth) one by one and obtain 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.

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.

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

These two tests are called Sum test and SSU test [Pan, 2009] respectively. The former is closely related to other burden tests such like those in [Morgenthaler and Thilly, 2007, Li and Leal, 2008, Madsen and Browning, 2009]. If there is a common association either in direction or strength for causal SNVs with no or few non-associated SNVs, then Sum test and the likes will be most powerful; otherwise, the SSU test and its closely relatives, such as kernel machine regression (KMR or SKAT) [Lee et al., 2012a, Ionita-Laza et al.,

2013, Oualkacha et al., 2013, Lee et al., 2012b, Wu et al., 2011] and C-alpha test [Neale et al., 2011], will be most powerful.

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

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

where $W = (W_1, \ldots, W_p)'$ is a vector of weights for the p SNVs [Lin and Tang, 2011]. Different researchers proposed different weighting schemes to pool the information of all SNVs in a region of interest, such as those used in [Madsen and Browning, 2009, Sul et al., 2011, Pan and Shen, 2011, Han and Pan, 2010, Li and Leal, 2008, Zhang et al., 2011, Lin and Tang, 2011, Basu and Pan, 2011]. However, all of these weighting schema have used fixed weights, for example, their weights were chosen to be proportional to the MAFs of SNPs, to the standard deviations of SNPs, to the regression coefficients, or to the single SNP p-value. There is no uniformly best weighting scheme as discussed in [Pan et al., 2014, Basu and Pan, 2011, Pan and Shen, 2011].

As a complement to SNPs weighted average, SNPs selection is preferred in the case that there are many non-associated SNPs among the group of SNPs to be tested. Such methods include aSum+ and aSSU which are based on Neyman-type tests [Neyman, 1937]. However, variable selection will also omit those variables with mild to moderate information. In our context, due to extremely low MAF of RVs, even underlying fact is that the individual RV is strongly associated with trait, there is only limited information stored in this single RV. Dumping seemingly non-informative RVs may actually omit the signals within the group of SNPs. Therefore, we expect the model averaging based tests will outperform the model selection based tests in above settings.

The SPU test

Our goal is to specify a whole class of weights which can cover a wide range of association patterns: for any given data with unknown association pattern, we hope at least one member of the whole class of weights can render a powerful test. We reason that, since association information is largely maintained in the score vector itself as comparable to regression coefficient, score vector is not only the basis in GEE and Score test based methods aforementioned, but may be an informative and simple weight! Specifically, we propose a class of weights

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

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 $\gamma = 1$, the SPU(1) test uses $\mathbf{1}$ as weight and sums up the information contained in all the SNPs in the region of interest, equivalent to Sum test or burden test; when $\gamma = 2$, the SPU(2) test uses U as weight to itself and is equivalent to SSU test and other variance-component test such as SKAT; when γ keeps increasing, the SPU(γ) test puts higher weights on the jth SNV with larger $|U_{.2,j}|$, while gradually decreasing the weights of other SNVs with smaller $|U_{.2,j}|$. As the large value of $|U_{.2,j}|$ indicates strong association information stored in SNV j and small value of $|U_{.2,j}|$ indicates weak or none association information stored in SNV j, a higher γ tends to put more and more weights on those informative SNVs. When $\gamma \to \infty$ as an extreme situation, where ∞ is assumed to be an even number, we have

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

which takes only the largest element (in absolute value) of score vector. Apparently,

SPU(∞) 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(γ) test with a large $\gamma > 8$ usually gave similar results as that of SPU(∞) test [Pan et al., 2014], thus I will only use $\gamma \in \Gamma = \{1, 2, \ldots, 8, \infty\}$ for the whole dissertation work. Suppose the sample size is large enough or MAF of SNP is large enough, thus the theory that the asymptotic normal distribution of the score vector can hold under the null hypothesis, I will use a simulation based method to calculate the p-value from each $T_{SPU(\gamma)}$ [Lin, 2005, Seaman and Müller-Myhsok, 2005]. Specifically, suppose T is a 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 the original data (see Equation 13). 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 $P_{SPU(\gamma)} = \sum_{b=1}^{B} \frac{I(T^{(b)} \geq T^{obs})+1}{B+1}$.

The aSPU test

Given we have a list of SPU(γ) statistics and p-values, we are still 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). I hereby propose an adaptive SPU (aSPU) test to achieve such a purpose. There is a number of combining methods, such as the exponential combine [Chen et al., 2012], linear combine, quadratic combine and fisher's combine methods [Luo et al., 2010, Peng et al., 2009, Derkach et al., 2013]. In this dissertation work, I will use the minimum-P-value combining method exclusively with room left for other combining methods. For different γ, it is difficult to characterize the power curve of an SPU test in real data situation. Thus, I 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(γ) test.

Similarly as the above simulation based method to obtain p-value of $T_{SPU(\gamma)}$, the *same strategy* can be applied to obtain the p-value of T_{aSPU}. Since the previous simulated $U_{.2}^{(b)}$ and $T_{SPU(\gamma)}^{(b)}$ can be reused here, another simulation work becomes *unnecessary*. Specifically, at the SPU test stage we already have the $U_{.2}^{(b)}$ for $b = 1, 2, \ldots, B$. 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}.

In practice for genome wide scan purpose, we can use a stage-wise aSPU test strategy: we first start with a smaller B, for example, use $B = 1000$ to scan the genomes, then gradually increase B to, for example, 10^6 for a few selected groups of SNPs. For example, we could choose specific genes or windows which passed a pre-determined significance cutoff (for example, p-value $\leq 5/B$) in the previous stage; we then repeat this process until the pre-determined significance level is reached. For example, a p-value of $\leq 10^{-7}$ requires we increase $B \geq 10^7$. In this stage-wise way, we will be able to apply the aSPU

test to GWAS data.

Other versions of aSPU test

- **aSPUw test**

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

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

  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 the above **aSPU** test. Finally, aSPUw p-value can be obtained by:

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

  It is worth noting that **aSPU** and **aSPUw** test can be implemented in the meantime using the same simulated score vector, which makes the computation more efficient.

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

  Although the **GEE Score test** will lose power in some scenario of the gene-based GWAS analysis as aforementioned, it still has the unique advantage in some scenarios when the correlation structure among SNPs matters. GEE Score test in the form of $T = U_{.2}' \Sigma_{.2}^{-1} U_{.2}^{-1}$ will keep the covariance matrix in the denominator, which preserves the information of possible linkage disequilibrium among SNPs. To combine the strengths from GEE Score test and aSPU(w) test, I propose to adopt

the minimum-P-value combining strategy again, yielding the aSPU(w).Score test statistic:

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

where P_{Score} is the p-value of the Score test.

2.2.2 Methods for Simulation Settings

I will generate the simulated genotypes following [Wang and Elston, 2007, Pan, 2009, Basu and Pan, 2011]. In brief, I will generate two independent blocks of SNPs for each subject: the first block will included causal SNPs and null SNPs in linkage disequilibrium (LD); the second block will include only null SNPs in LD. I will use the first-order auto-regression (AR(1)) correlation structure to imitate real-world LD among SNPs. I will simulate the longitudinal response variables using AR(1) following [Song et al., 2013]. Then I will take into account the SNPs main effect and time course main effect as fixed effects on the longitudinal response variables **without consideration of SNP $\times$ time interaction**. I will not consider other covariate effects (such as demographic) in the simulation studies, though they can be simply added without any change to the method. I will refer to several literatures [Pan et al., 2014, Basu and Pan, 2011, Pan and Shen, 2011, Han and Pan, 2010, Pan, 2009] and the Atherosclerosis Risk in Communities (ARIC) data (https://www2.cscc.unc.edu/aric/) to set up the simulation parameters, for example, ρ_y across longitudinal measurements and ρ_x across SNVs as used in AR(1) correlation structure model.

I did notice that there are other strategies in simulating the genotype data, such as the forward time simulation method to generate population genetic data, which includes co-alescence models like two-epoch model and six parameter complex bottleneck model, and allows for simulation of purifying selection effect and scaled fitness effect as well [Boyko

et al., 2008, Hernandez, 2008]. Compared with the populational genetics simulation method, my simulation strategy does not take into account the population coalescence theory and assumes each sampled individual genotype is independent to the others. The lack of these properties may be a limit in my simulation studies. However, my simulation strategy takes the edges on the flexible control over the correlation magnitude among SNPs, the desired MAF of SNPs and the proportion of causal SNPs. Such advantages were proved and utilized in developing new association tests in a number of past researches [Wang and Elston, 2007, Pan, 2009, Han and Pan, 2010, Pan and Shen, 2011, Basu and Pan, 2011, Pan et al., 2014, Zhang et al., 2014].

Methods for simulation of genotype data

To construct a block of SNPs for subject i, at first, a latent vector $G_i = (G_{i1}, \ldots, G_{ip})'$ will be 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 $f \neq g$. The default ρ will be set at 0.8 to mimic the real data. Secondly, the latent vector will be dichotomized to yield a haplotype with each latent element G_{ij} dichotomized to 0 or 1 with probability $\text{Prob}(G_{ij} = 1) = \text{MAF of } j\text{th SNP}$; the MAFs will be randomly drawn from a uniform distribution: for causal SNPs, the MAFs will be set between 0.3 and 0.4; for null SNPs, the MAFs will be set between 0.1 and 0.5. Thirdly, we will combine two independent haplotypes to form the genotype $X_i = (X_{i1}, \ldots, X_{ip})'$ for subject i. The haplotypes for different subject will be generated **independently**, that is, the subjects will be simulated as perfectly **independent** subjects with no identity by descent (IBD).

Using this strategy, I will placed 35 SNPs in the first block with AR(1) correlation structure to imitate the real LD structure among these SNPs; out of 35 SNPs, I will randomly select 5 SNPs to be causal (that is, they will have non-zero coefficients). To mimic the case in the SNP genotyping array platforms, for example, tag SNPs are genotyped but

not the causal SNPs. I will excluded the 5 causal SNPs from the later test procedure. Therefore, in the first block, only null SNPs in LD with these 5 causal SNPs will be retained. I will further place 15 null SNPs in the second block as I planed for the first block. Note that the first block and second block are independent to each other. All the 15 null SNPs from the second block will participate in the test.

Methods for simulation of phenotype data

To simulate longitudinal quantitative phenotype, I will follow the strategy used in [Song et al., 2013]. Specifically, I will first implement an exploratory analysis, that is the generalized least square estimation with AR(1) correlation structure, on the candidate lipid traits from the ARIC study, to get an approximate estimate of the correlation coefficient between traits across time points. For example, I will obtain $\rho_{data} = 0.7$ on average for different lipid trait candidates.

Secondly, I will setup the mixed effect model to achieve the AR(1) correlation structure as:

$$y_{im} = \mu_i + b_i + \underbrace{\rho e_{i,m-1} + s_{i,m}}_{e_{i,m}}, \tag{5}$$

with $m = 1, \ldots, k$ indexes the longitudinal measurements within subject i as stated in section 2.2.1; $\mu_i = Z_i\varphi + X_i\beta = H_i\theta$ as in quantitative trait case (see section 2.2.1); 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. I can simply plugin the estimate from the real data here, by setting up $\rho = 0.7$. $e_{i,m}$ is the total residual, which can be divided into two parts: the first part depends on $e_{i,m-1}$ and the second part is an independent term.

I assume the following distribution:

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

It's not difficult to see the term $\rho e_{i,m-1} + s_{i,m}$'s variance is equal to the variance of $e_{i,m}$ by algebraically summing up two parts. Under this assumption, the variance-covariance matrix across longitudinal measurements becomes (assuming $k = 4$ for the number of longitudinal measurements):

$$\Sigma_{4\times 4} = Var \begin{pmatrix} b_i + e_{i1} \\ b_i + \rho e_{i1} + s_{i2} \\ b_i + \rho e_{i2} + s_{i3} \\ b_i + \rho e_{i3} + s_{i4} \end{pmatrix} = \sigma_b^2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \tag{6}$$

Among the rightmost two parts summed together in Equation (28), the first part defines the inter-subject variances, and the second part allows the measurements with a k-interval lag to have a correlation coefficient of ρ^k. This is closer to reality in some cases for longitudinal data.

Methods for tuning simulated genetic effect

As noted in association tests, different SNPs contribute to the phenotype with different effect sizes. However, the SNP effect magnitude tuning in the simulation study is not a trivial task. Instead of assigning a β_d coefficient to a SNP with a arbitrary numerical value, for example, 0.1 or 10000, there is a way to use genetic heritability to control the association magnitude from the jth SNP [Lynch et al., 1998]. Let I first introduce the

formula of the variance of the phenotype :

$$Var(y_{im}) = Var(X_{ij})\beta_j^2 + \sigma_{oth}^2 = 2f(1-f)\beta_j^2 + \sigma_{oth}^2 \tag{7}$$

where the Hard-Weinberg equilibrium (HWE) is assumed to hold. f is the MAF of the SNP; σ_{oth}^2 is the residual variance after removing the effect of jth SNP. Obviously we can see σ_b^2 and σ_e^2 are contained in σ_{oth}^2 (see equation (27)), 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 (29):

$$h^2 = \frac{Var(A)}{Var(P)} \tag{8}$$

This is the classical formula of narrow-sense heritability, with $Var(A)$ represents the variance due to the additive effects of the alleles, and $Var(P)$ represents the total variance in the phenotype. In our situation for jth SNP, this can be expanded to:

$$h_j^2 = \frac{Var_j(A)}{Var(P)} = \frac{Var(X_{ij})\beta_j^2}{Var(y_{im})} = \frac{Var(y_{im}) - \sigma_{oth}^2}{Var(y_{im})} \approx \frac{Var(y_{im}) - \sigma_b^2 - \sigma_e^2}{Var(y_{im})} \tag{9}$$

By systematically solving the equations (29) and (31), we can easily calculate the β_j for jth SNP once we have pre-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 will use $h_j^2 = 0.001$ as default value in the simulation studies to control β_j.

2.2.3 Plan for Simulation Studies

I will evaluate the proposed method's performance using simulated data. Specifically, I will simulate genotype data and longitudinal phenotype data mimicking real data. For example, the simulated genotype data will allow LD structure among SNPs and the

simulated longitudinal trait will allow flexible control over the between-subject variance, the within-subject variance, the number of measurements and the correlation structure among measurements. I will benchmark the new test against several existing methods, such as Sum test, UminP test and Score test. Specifically, I will evaluate whether the Type I error can be controlled at the nominal level (neither inflated nor conservative), and compare the empirical power under different simulation scenarios.

I plan to perform the following simulation studies:

1. **Power comparison between longitudinal study and cross-sectional study**
   I will evaluate the power gain from longitudinal study over cross-sectional study by estimating the empirical powers as a function of the number of visits (starting from one, that is actually the cross-sectional study, to k, for example, four as the maximum measurement number). We will also test the power gain magnitude under different levels of within-subject correlation coefficient.

   The quantities of interested include:

   (a) the magnitude of power gain at different levels of ρ, the within-subject correlation coefficient as used in the simulation of the AR(1) correlation structure. For example, $\rho = 0.3$ represents a weak correlation between measurements of the same subject, while $\rho = 0.7$ represents a strong correlation;

   (b) the empirical powers as a function of the number of visits. We want to confirm the magnitude of the power gain coming from each extra follow-up measurement. There may be the case when k increases to a specific level, for example three, the power gain after it will be negligible as compared to previous power gains. This is the so called "elbow point", which is quite meaningful in deciding a sufficient point to stop. In our settings, we do not want to infinitely increasing the k, which will lead to a larger and unnecessary cost. A sufficient

k will achieve a relatively higher power to meet the study requirement, for example, a power of 0.9 in longitudinal studies, while avoiding unnecessary cost from pursuing an even larger k.

2. **Type I error benchmark under the default simulation settings with varying sample sizes**

   I will evaluate the Type I error of the aSPU test and its extensions (we will call them aSPU family hereinafter) as compared with several existing tests: Score GEE, UminP, Sum Test, weighted Sum Test and SSU test. I will set the significance level at 0.05. I provide a sample table below to show format of the future result presentation (dummy numbers shown in each cell).

n	Score	UminP	SumP	SumP.w	SSU	aSPU	aSPUw	aSPU.sco	aSPUw.sco
500	0.038	0.059	0.048	0.033	0.034	0.052	0.045	0.040	0.058
1000	0.048	0.054	0.049	0.059	0.045	0.035	0.044	0.049	0.047
2000	0.056	0.042	0.043	0.033	0.049	0.062	0.045	0.048	0.048
3000	0.055	0.053	0.067	0.050	0.055	0.033	0.054	0.046	0.049

Table 1: Sample Table of Type I error Benchmark among tests

3. **Empirical power benchmark under the default simulation settings with varying sample sizes**

   I will benchmark the empirical power among aSPU family tests and several existing tests. I will keep the significance level α at 0.05, and evaluate the empirical power for sample size n equal to 500, 1000, 2000 and 3000. We will present either a figure plotting power curve as a function of n for each participating test or a power table in the similar format as the empirical Type I error table above.

4. **Empirical power benchmark under the simulation settings where half of causal SNPs are in the opposite effect direction**

   Out of the 5 causal SNPs (simulated in the region with all other SNPs but excluded from the tests assuming that the causal SNPs are not genotyped), I will set 2 of them to have opposite effect direction copmared with the rest 3 SNPs by flipping the signs of SNP main effects. Other settings will remain the same as above. I will keep the significance level α at 0.05, and evaluate the empirical power for sample size n equal to 500, 1000, 2000 and 3000. I will present either a figure plotting power curve as a function of n for each test under consideration or a power table in the similar format as the empirical Type I error table above.

5. **Empirical power benchmark under the simulation settings where the number of null SNPs number grows**

   In association testing, there is sometimes the case that in a region of interest, causal SNP signals are very sparse. Namely, there exists many null SNPs. I hence want to investigate the performance of the aSPU family tests in the presence of a larger proportion of null SNPs in the region. I will gradually increase the number of null SNPs from 50 to 75, 100, 200, and then finally 400. I will only consider $n = 3000$ as the sample size in this scenario. I will keep all other settings the same as Scenario 3 (the default simulation settings) above. I will present either a figure plotting power curve as a function of number of null SNPs for each test under consideration or a power table in the similar format as the empirical Type I error table above.

6. **Empirical power benchmark under the simulation settings where working correlation structure varies**

   I will investigate the aSPU family tests' performance under other working correlation structures than the working independence, such as AR(1), compound sym-

metry, and unstructured. Note that, as I simulated the longitudinal trait using the AR(1) correlation structure as described in Section 2.2.2, fitting GEE with the AR(1) working correlation matrix is actually using the true correlation matrix. I will keep all other settings the same as Scenario 3 (the default simulation settings) above. I will present either a figure plotting power curve as a function of n for each test under comparison under a specific working correlation matrix or a power table in the similar format as the empirical Type I error table above. It is of interest to investigate the effect of combining a specific working correlation matrix and a specific n for each test.

2.3 Methods for Aim 1(b): To develop data-adaptive SNP-set based association tests for longitudinal data analysis within the Generalized Estimating Equations (GEE) framework for rare variants

2.3.1 Statistical Modeling

In the previous section 2.2.1, I have discussed the method development of the aSPU family tests on common variants with a longitudinal trait. In this section, I will discuss the extension of the proposed methods to rare variants.

While MAF of RVs are usually low, for example, between 0.001 to 0.01, the property of asymptotically normal distribution of either the regression coefficient or score vector may not hold. The simulation-based p-value calculating method as proposed in Aim 1(a) for CVs may not be sufficient for RV analysis. Specifically, in last section, 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 the RV case by large, except that the $U_{.2}^{(b)}$ may not follow the multivariate normal distribution any longer. As a remedy, I propose a permutation based method to generates the empirical null distribution of $U_{.2}^{(b)}$ instead of the previous simulation based method. The permutation strategy will maintain the relationship between longitudinal traits and possible covariates such as age and gender for subject i, in other words, the permutation strategy will only permute the genotype part in the regression model. The algorithm will be also robust to missing data as this is usually the case in longitudinal data settings. After we obtain enough $U_{.2}^{*(b)}$ from permutation strategy to form an empirical null distribution, the left work of the aSPU tests for RVs will be exactly the same as we did on CVs. This is because, the only difference between the CV analysis and RV analysis is, the null distribution of score vectors for CVs is obtained by simulation based method, while for RVs, the null distribution of score vectors is obtained by permutation based method.

2.3.2 Methods for Simulation Settings

The simulation strategy of RV genotype data is almost the same as previous strategy for generating CV genotype data (see section 2.2.2), except that:

1. the MAF of RVs, regardless of causal one or null one, are set between 0.001 and 0.01.

2. the causal RVs are not excluded from later test stage as we expect the whole-genome sequencing or exome sequencing platform will identify high density variants including the causal ones.

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

2.3.3 Plan for Simulation Studies

I will test the proposed methods' performance using simulated data. Specifically, we will simulate RV genotype data and longitudinal phenotype data mimicking real data. For example, the simulated genotype data will allow LD structure among SNPs and the simulated longitudinal trait will allow flexible control over the between-subject variance, the within-subject variance, the number of measurements and the correlation structure among measurements. I will benchmark the new test against several existing methods, such as Sum test, UminP test and Score test. Specifically, I will evaluate whether the Type I error could be controlled at the nominal level (neither inflated nor conservative), and compare the empirical power under different simulation scenarios. I will evaluate, on simulation data, the effect of implementing permutation or parametric bootstrap strategy. We expect to see such procedures can provide a better control of type I error and a more unbiased estimate of the real power.

I plan to perform the following simulation studies:

1. **Type I error benchmark using simulation-based P-value calculating method under the default settings with varying sample sizes**

   Similarly as I planed for CVs in Aim 1(a), I will evaluate the aSPU family tests' Type I error performance as compared to a few existing tests for RVs. I will still use the simulation-based P-value calculating method as we planned for CVs before. I will compare the aSPU family tests with SSU, SSUw, Score, Sum, UminP (calculated by simulation-based method) and mvn.UminP (calculated by approximating a multivariate normal distribution) test. I will set the significance level α at 0.05. I provide a sample table below to show the format of future presentation (dummy

numbers shown in each cell).

n	pSSU	pSSUw	pScore	pSum	mvn.UminP	UminP	SPU(1)	SPUw(1)	SPU(2)	SPUw(2)	aSPU	aSPUw	aSPU.sco	aSPUw.sco
500	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.021	0.055	0.035
1000	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055
2000	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.066	0.062	0.062	0.062
3000	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055

Table 2: Sample Table of Type I error Benchmark among tests using simulation-based method in RV analysis. mvn.UminP: UminP calculated by approximating a MVN distribution; UminP: UminP method calculated by simulation-based method.

2. **Empirical power benchmark using simulation-based P-value calculating method under the default settings with varying sample sizes**

   I will benchmark the empirical power among aSPU family tests and several existing tests in RV analysis using simulation-based P-value calculating method. I will keep the significance level α at 0.05, and evaluate the empirical power for sample size n equal to 500, 1000, 2000 and 3000. I will present either a figure plotting power curve as a function of n for each test considered or a power table in the similar format as the empirical Type I error table above.

3. **Type I error benchmark using permutation-based P-value calculating method under the default settings with varying sample sizes**

   I plan to compare the result from previous Scenario 1 to the Type I error benchmark result from using **permutation-based** P-value calculating method. I will compare the aSPU family tests with SSU, SSUw, Score, Sum, UminP (calculated by simulation-based method) and mvn.UminP (calculated by approximating a multivariate normal distribution) tests. I will set the significance level α at 0.05. I will present the result using a similar table as the previous example table 2.

4. **Empirical power benchmark using permutation-based P-value calculating method under the default settings with varying sample sizes**

   I will benchmark the empirical power among aSPU family tests and several existing

tests in RV analysis using **permutation-based** P-value calculating method. I will also compare them to the one computed by **simulation-based** P-value calculating method in Scenario 2. I will discuss the observed difference between these two methods for RV analysis. I will keep the significance level α at 0.05, and evaluate the empirical power for sample size n equal to 500, 1000, 2000 and 3000. I will present either a figure plotting power curve as a function of n for each test under comparison or a power table in the similar format as the previous example table 2.

5. **Evaluation of the performance of proposed aSPU.aSPUw.Score test**

   Within the aSPU, aSPUw, and Score test, there may be at least one test having a high power in a specific data scenario, depending on the association pattern and the correlation structure within SNPs. I plan to combine the three tests to the aSPU.aSPUw.Score test.

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

   Afterwards, I will evaluate the performance of the proposed test with regard to empirical Type I error and power. We are interested in whether the proposed new test, that have combined advantages from two adaptive tests and the score test, can control the Type I error well and maintain a higher power in all scenarios (regardless of the variance homogeneity of RVs). I will present the result using a similar table or power curve plot as above.

2.4 Methods for Aim 2: To develop the pathway-based data-adaptive association tests for longitudinal data analysis

In the previous sections, I have discussed the method development of the aSPU family tests in a gene-based or region-based manner for CVs and RVs. Here, I will discuss the extension of the new methods to the pathway-based manner, the so called **Path-aSPU**. Path-aSPU is proposed mainly for analyzing RVs. Since RVs have extreme low MAFs, they need more aggregation to increase the test power. Like we aggregate the RVs in a gene, we further aggregate the genes in a pathway. Additionally, there may be a large number of non-associated RVs, a preferred case for the aSPU family. Thus I expect Path-aSPU will perform well when comparing to existing methods, such as SSU, Sum and UminP test.

2.4.1 Statistical Modeling

A pathway analysis will involve multiple genes (for example, 20 or 50). Too few or too many genes in a pathway will make the pathway difficult to interpret in the biological perspective. For example, a pathway with only two genes and another pathway with 2000 genes are both difficult to interpret. Therefore, I will consider the pathway with a reasonable number of genes, for example, 20 to 200. For each gene within a pathway, it may contain different numbers of RVs. For example, a gene has 10 RVs while another gene has 400 RVs. This bring over a new problem that a larger gene (a gene with more RVs) may dominate a smaller gene (a gene with fewer RVs). Hereby, I propose to modify the aSPU test to adjust for various gene length, thus avoid the dominant influence from a large (or small) gene.

Let the short notation $U_{g.}$ represent $U_{.2}$ for genotype data , and $U_{g.} = (U_{g,1}, U_{g,1}, \ldots, U_{g,p_g})'$ represent the score vector for gene g with p_g RVs from the GEE fitting. Given a pathway

(or a gene set) S, the gene-specific SPU statistic is as follows:

$$T_{SPU(\gamma;g)} \propto ||Ug.||_\gamma = \left(\frac{\sum_{j=1}^{p_g} |U_{g,j}|^\gamma}{p_g} \right)^{\frac{1}{\gamma}} \tag{10}$$

Then accordingly, the pathway-based SPU statistic is

$$T_{Path-SPU(\gamma,\gamma 2;S)} = \sum_{g \in S} (T_{SPU(\gamma;g)})^{\gamma 2} \tag{11}$$

Note the $T_{SPU(\gamma;g)}$ is now standardized by the 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. Again, for any given $(\gamma, \gamma 2)$, I will recourse to the same simulation or permutation based 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:

$$T_{Path-aSPU(S)} = min_{\gamma,\gamma 2} P_{Path-SPU(\gamma,\gamma 2;S)} \tag{12}$$

We again adopt the same strategy as previous, that is we will utilize the same set of simulated U generated in the previous step for calculating $P_{Path-SPU(\gamma,\gamma 2;S)}$) to calculate the final **pathway-based aSPU** p-value $P_{Path-aSPU(S)}$.

The intuition of $\gamma 2$ is like that of γ: If we treat the pathway as the gene and the gene as the RVs. A larger $\gamma 2$ (γ) put more weights on heavily associated genes (RVs), when gradually ignoring the less associated genes (RVs) in a pathway (gene). An extreme case is $\gamma 2 = \infty$, as I already explained $\gamma = \infty$'s interpretation in section 2.2.1, it indicates the pathway-based analysis actually used only one gene - the most heavily associated gene with the trait. Since the goal of pathway-based analysis is to take advantage of multiple "co-working" genes, and aggregate more RVs, it is less meaningful to consider the use

of a $\gamma 2 = \infty$. Instead, I propose to use $\gamma 2 \in \Gamma 2 = \{1, 2, 4, 8\}$. The reason is that, at the pathway level, the statistic $T_{SPU(\gamma;g)}$ is always a positive number, not like that $U_{.2,j}$ from the GEE model fitting for variants can have different signs (SNP effect directions). Thus, deliberately assigning both odd and even number of $\gamma 2$ becomes unnecessary, and I can actually use most representative $\gamma 2$ values and expect them to have most distinct effects from each other. Fewer $\gamma 2$ candidates will also expedite the computation. To these purposes, $\gamma 2 \in \Gamma 2 = \{1, 2, 4, 8\}$ will cover Sum-like test, SSU-like test, and two more tests preferring the sparse-causal-gene situation (for example, only 2 or 3 genes are associated with the longitudinal trait in a pathway with 20 genes).

2.4.2 Methods for Simulation Settings

The simulation strategy of RVs within a gene is the the same as section 2.3.2. I will use the same simulated longitudinal phenotype data as in Aim 1(a).

With regard to the pathway consisting of multiple genes, I will simulate a pathway with 20 genes. Each gene g will contain p_g RVs with p_g randomly draw from a uniform distribution $U(5, 20)$. 5 of the 20 genes will be randomly selected to be causal, with each causal gene containing $U(1, 3)$ causal RVs.

2.4.3 Plan for Simulation Studies

I will test the Path-aSPU tests' performance on the simulated data to evaluate the empirical Type I error and power, with comparison to several existing tests like SSU, Sum, Score and UminP test. I will fix the sample size $n = 1000$; I will set the single SNP heritability $h_j^2 \in \{0.001, 0.0025, 0.005, 0.0075, 0.01\}$ to control the effect size. All other settings will be the same as Scenario 2 of Aim 1(a) simulation study plan (see section 2.2.3).

Optionally, I may consider more extensive simulation studies, such as tuning the causal RV number within a causal gene or tuning the number of causal genes within a pathway, using independent RVs within a gene instead of correlated RVs in AR(1), and using different working correlation matrix. In these ways, I can evaluate the robustness of the Path-aSPU performance.

2.5 Methods for Aim 3: To develop the software package for method implementation

I will develop a software package using R language and Linux shell script mainly. The package/software will require a few existing R packages, such as "data.table" enables big data manipulation and "geepack" enables GEE estimation, to make itself functional and more efficient. All dependent R packages can be freely downloaded from CRAN (http://cran.r-project.org/). The software package will install the dependent R packages automatically upon first time software installation/setup.

The software package will have a list of nice properties. It will be straightforward to install and use for 1st-time user. The software package will allow user to run the program in a very flexible parallel computation framework. For example, user can choose to use a single node with multiple cores or use multiple nodes with multiple cores. I will employ either SOCKET or MPI as the parallel computing protocol. I will use R packages "SNOW", "doSNOW" and "doMC" to fulfill the parallel function. The software package will also have the state-of-the-arts technique to enable efficient implementation of the aSPU methods, such as the hash table, radix sort, memory-efficient task send and collect among nodes, and calling C++ code for some intensive loops.

The software package will finally have a clear help document with demonstration examples in addition to on-screen help brief (triggered by command, for example, "Exe-

cutable.r -h"). The software can be operated though Linux command line arguments. For example, the command line that "Executable.r -i inputname -o outputname -p FALSE -S SNOW -m FALSE", will use the executable file "Executable.r" to process the input file "inputname" and later output the file "outputname". It disables the permutation-based method, instead uses the simulation-based method by specifying "-p FALSE". It uses parallel computing schema set by "SNOW" and executes on single node with multiple cores as indicated by "-m FALSE". This is a short example, the real arguments could be more complicated to allow more flexible control of the software package.

I will test our software package on the simulated dataset for the purpose of debugging and optimization. Since we know the "real answer" for simulated dataset, testing the software package on such dataset and evaluate the result can help us confirm the scientific function of the software package is correct.

I will test our software package on the ARIC data for the purpose of debugging and optimization. I expect the real dataset can give us more useful feedback from, for example, the innate data complexity, which simulated dataset usually lack, to help us improve the robustness, convenience and efficiency of the software package.

2.6 Methods for Real Data Application for Proposed Aims

Here I will summarize the methods of real data application for proposed Aims 1 and 2. For Aim 3, since it is for software package development, I will use the package, test it, and improve its performance and robustness through the whole real data application processes for Aim 1 and 2. As shared in common, I will apply the developed novel methods for both Aims 1 and 2 to the Atherosclerosis Risk in Communities (ARIC) data (https://www2.cscc.unc.edu/aric/). I will exclusively use the Caucasian samples ($n = 11478$). Specifically, I will select one or several traits from ARIC cohort data as

the response variable(s) with longitudinal measurements. Candidate traits include four cardiovascular disease-related lipid traits, which are total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C) and triglycerides (TRG). I will **take cautions before using these lipid traits**, such as accounting for lipid-lowering therapy in TC and LDL-C traits, and natural log transformation on the TRG trait according to the procedures described in [Peloso et al., 2014].

For genotype data part, in Aim 1(a), I will use the common variants genotyped by the ExomeChip [Grove et al., 2013] platform in the ARIC study. I will follow conventional quality control (QC) criteria for GWAS. For example, the MAF of any SNP should be greater than 5%, missing rate of both single SNP and single subject should be less than 5% and hardy-weinberg-equilibrium (HWE) test p-value should be greater than 0.001. In Aim 1(b), I will use the rare variants genotyped by the same ExomeChip platform in the AIRC study. I will follow conventional QC criteria for rare variants analysis. For example, the MAF of any SNP should be less than 5%, missing rate of both single SNP and single subject should be less than 5%, HWE test p-value should be greater than 0.001 and region aggregate counts of minor alleles should be greater than 20 or 40 as previously done in [Lange et al., 2014, Peloso et al., 2014]. In Aim 2, I will define the gene pathway by public pathway resources like KEGG [Ogata et al., 1999] and BioCarta [Nishimura, 2001]. I will consider the medium size pathways in selected database, for example, the pathway with 20 to 100 genes. I will use the rare variants from the ExomeChip platform [Grove et al., 2013] in the ARIC study. I will adopt the same QC criteria for rare variant analysis as in Aim 1(b).

With regard to covariates, I will include top two principal components eigenvectors (PCs) produced by EIGENSTRAT [Price et al., 2006] in the longitudinal regression model to adjust for the potential population structure within the ARIC Caucasian subjects. Additionally, I will include subject's demographic information such as age, gender and

BMI. I will also include the fixed time course effect as a covariate.

I will run gene-based test with gene boundary defined in the analysis of Peloso et.al.[Peloso et al., 2014] We only included nonsynonymous and splice site variants. Optionally, if sliding-window based test is chosen for Aim 1(a), I will set the window size to 40 consecutive SNPs in a window, while neighboring windows share 10 SNPs as the step size so that SNPs signals in the gap between two windows would not be omitted.

By implementing the real data application, I expect to verify known risk genes and/or pathways related to either cardiovascular disease or associated trait level, therefore I can validate the functionality of our proposed methods. I also look forward to identify the novel risk genes and/or pathways, which will provide new valuable information to the disease research consortium. Since I will provide a software package to implement our new methods, I will test its correctness and robustness through the whole real data application.

2.7 Declaration on Human Subjects

This dissertation study will focus on statistical method development. I will use the ARIC cohort data for method demonstration purpose. I will use the blood lipid phenotypes, covariates such as demographic variables, and genotype (GWAS and ExomeChip) in the ARIC data set. All data are pre-existing and de-identified. The IRB approval for the use of ARIC data set in my dissertation research has been obtained by my dissertation advisor, Dr. Peng Wei, under UTHealth IRB approval (HSC-SPH-13-0492).

3 Journal Article 1

Title of Journal Article

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

Name of Journal Proposed for Article Submission

American Journal of Human Genetics

Abstract

Many prospective cohort studies and electronic health record (EHR)-based cohorts have collected phenotypes with repeated measures; however, genetic association studies have thus far ignored the rich longitudinal information and simply used the baseline measurements, leading to reduced statistical power. In addition, since complex phenotypes are likely affected by multiple genetic variants, a set-based analysis of these variants may help explain additional phenotypic variations and boost the statistical power. Due to the unknown and varying complexity in the association patterns between genetic variants and phenotypes, a single association testing method may not perform well under all possible genetic architectures. Thus, a data-adaptive association testing method is preferred as it accommodates different association patterns and adaptively selects a test from a class of tests to maintain the high statistical power. In this article, we proposed a family of data-adaptive SNP-set based association tests for longitudinal phenotypes within the Generalized Estimating Equations (GEE) framework. We developed different strategies to appropriately handle common or rare variants in the association studies. We evaluated the performance of proposed methods by extensive simulation studies. We demonstrated that the proposed methods are robust to the misspecification of within-subject correlation, which is a desirable feature for analysis of longitudinal phenotypes. We also showed that the proposed methods can handle subjects with missing measurements effectively. Our simulation studies demonstrated substantial power gain by: (i) using the proposed methods over several existing set-based methods for longitudinal phenotypes, and (ii) using the proposed methods compared with analysis of the baseline measurement only. We illustrated the utility and efficiency gains of proposed methods using the Exome Chip data from the Atherosclerosis Risk in Communities (ARIC) study. The real data application demonstrated that our proposed method was able to identify the same genetic variants

associated with high-density lipoprotein cholesterol (HDL-C) using repeated measurements from the ARIC dataset only as compared with a previous study with much larger sample size and multiple cohorts but only using the baseline measurement. In addition, we identified a novel gene *FAM65A* associated with HDL-C, which is worth follow up in future studies. Finally, we have built an efficient software package to implement the proposed methods and make it scalable for large-scale genome-wide association studies with longitudinal phenotypes.

3.1 Introduction

Genome-wide association studies (GWAS) have been popular since 2007. Hundreds of GWAS have been published already (A Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/26525384). For genetic studies on cardiovascular disease risk factors, for example, the Atherosclerosis Risk in Communities (ARIC) study, repeated measurements of traits across study time are available for each individual [Heiss, 1989]. Association tests that fully utilize the information across time points tend to achieve a higher power and identify more disease-associated loci [Furlotte et al., 2012, Xu et al., 2014]. However, current statistical methods for testing the association in longitudinal settings, even for one single nucleotide polymorphism (SNP) at a time, are limited [Fan et al., 2012, Furlotte et al., 2012]. Investigators often take simplified approaches that collapse the repeated measurements into a single value, such as use baseline measurement and use average measurement, to handle the longitudinal data in GWAS. These approaches cannot fully harness the power of the complete information contained in the longitudinal trajectory. On the other hand, researchers can also apply the standard methods for correlated outcome model to longitudinal GWAS settings, namely, random effect models [Laird and Ware, 1982] and marginal models estimated by generalized estimating equations (GEE) [Liang and Zeger, 1986, Zeger et al., 1988]. However, these standard methods are not optimized for testing a large number of SNPs, a common scenario in set-based association test, for example, in a gene or in a region of interest.

Recent studies showed the advantages set-based association tests brought over the single SNP based analyses. First, the recent prevalence of next-generation sequencing technology enables investigators to access exome sequencing data or even whole-genome sequencing data. Compared to the previous microarray or chip based genotyping platform,

sequencing platform could catch much more dense markers as in base pair accuracy. More markers inevitably lead to heavier computation burden and multiple tests correction burden using the single SNP based analyses, however, they are greatly alleviated by set-based association tests. Second, set-based tests enhance the power of identifying rare variant association in next-generation sequencing studies due to aggregated information [Morris and Zeggini, 2010]. Third, in a region of interest, the genetic markers in linkage disequilibrium (LD) with the causal SNP(s) may bring additional information so that it increases the power of set-based tests to identify the true effect. Last, set-based tests have the edge in analysis of multi-ethnic cohorts because of the differences in LD structures across ethnic groups and thus meta-analysis of a set-based statistic is likely to be more consistent than meta-analysis of a single SNP based statistic across ethnic groups. There are many existing popular set-based methods, such as Uniformly min-Pvalue (UminP), GEE-Score, the weighted sum statistic (WSS) [Madsen and Browning, 2009], the the sum of squared U-statistics test (SSU) [Pan, 2009], variable-threshold (VT) test [Price et al., 2010], and sequence kernel association test (SKAT) [Wu et al., 2011]. All these methods could drastically reduce the number of regression parameters, capable of handling high-dimensional variants. Nevertheless, they differ in their underlying framework and general idea. For example, UminP test uses the only SNP in a region with the most extreme statistic. GEE-Score test applies the traditional score test to test the effect of SNPs after fitting the GEE model. WSS, as a representative of burden test, uses minor allele frequency (MAF) based weighting scheme to combine the test statistics from multiple SNPs in a region. VT test computes z-score $z(T)$ for each different MAF threshold T, defines z_{max} by permutation on phenotypes. SSU test and SKAT test are both variance component score tests which treat the effect of SNPs as random effect and test the variance equal to 0. These tests have their own edges in different testing circumstances, however, no such test can dominate the others in all scenarios [Han and Pan, 2010, Sun

et al., 2013, Pan et al., 2014, Derkach et al., 2013].

Recently, several "Data Adaptive" association tests emerged which relieved the above issue. Such tests combined the advantages from at least two previous methods to make the new test more well-rounded in different scenarios. The EREC method [Lin and Tang, 2011] builds a general framework for association testing which combines strength from burden test and VT test to form the most powerful test by setting the weight function ϵ proportional to the set of estimated regression coefficients. A data adaptive test combines the score test, SSU test and Sum test's advantages [Pan and Shen, 2011]. An exponential combination (EC) framework for set-based association tests [Chen et al., 2012] features with the sum of exponential statistics (statistics should follow either independent normal or independent chi-square distribution). The sum of exponential statistics are parametric and standardized from previous MB test and C-alpha test. A robust and powerful test uses Fisher's method to combine linear and quadratic statistics [Derkach et al., 2013]. A unified mixed-effect model [Sun et al., 2013] tests both group effect equal to 0 and variance component equal to 0. It includes both burden and SKAT tests as special cases by embedding the variant functional information and allowing a variant specific random effect in the model. More recently, the adaptive sum of powered score vector (aSPU) method [Pan et al., 2014] was proposed. The main idea of the aSPU test is that, it treats the score vector U from GEE estimation as the weight for SNPs. Since U intrinsically contains large association information of a SNP in a region contributing to a trait, use U as the weight is easy to interpret and efficient. By assigning different power on the weight, on could construct a class of SPU tests overweighting a sequence of increasingly smaller sets of the top-ranked (i.e., most statistically significant) SNPs, then select the test with the most significant result (with a proper adjustment for multiple testing). For relatively small sets of rare variants, the aSPU test often outperform other tests. [Pan et al., 2014] So far, however, extensions of these methods are not available for longitudinal data.

It is desirable to have a set-based and adaptive association test for longitudinal data. Extending the newly developed aSPU method to longitudinal settings, we propose the longitudinal aSPU (LaSPU). LaSPU models the repeated measurements through GEE method by incorporating within-subject correlation. It works on common variants (CVs) or rare variants (RVs), for the latter a special permutation procedure is optional for obtaining a more precise p-value. Within the LaSPU, we also propose a few more generalized methods to combine the SPU statistics and other statistics, for example, the score test statistic and the weighted version SPU statistics (SPUw). These methods complement the performance of the original aSPU method in some scenarios as from our extensive simulation studies. At last, we demonstrate the application of LaSPU by analyzing the association between high density lipoprotein cholesterol (HDL-C) and genetic variants data from ExomeChip platform [Grove et al., 2013] in the ARIC study [Heiss, 1989].

3.2 Material and Methods

3.2.1 Data and Notation

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 genotype score row vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ with x_{ij} coded as 0,1 or 2 (i.e., additively) for the count of the minor alleles of SNP $j = 1, \ldots, p$, and $z_i = (z_{i1}, z_{i2}, \ldots, z_{iq})$ is a row vector for q covariates including measurement time. We assume common effect sizes (i.e., time-averaged group effect) of the SNPs and covariates (other than measurement time) on the repeated 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 $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ and $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_{q+1})'$ for X_i and Z_i respectively. The marginal mean of each measurement, $E(y_{im}|x_i, z_i) = \mu_{im}$ where $m = 1, 2, \ldots, k$ 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):

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

with $H_i = (Z_i, X_i), \theta = (\varphi', \beta')'$ and $g(.)$ 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 [Liang and Zeger, 1986]:

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

with

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

and

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

ϕ in V_i is the dispersion parameter in GEE and is usually treated as nuisance parameter. $v(\mu_{im})=\phi\mathrm{Var}(y_{im}|x_i,z_i)$. $R_w(\alpha)$ is a working correlation matrix depending on some unknown parameter α. For convenience, a working independence model with $R_w=I$ is often used. 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:

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

$$\widetilde{\Sigma}=\widehat{\mathrm{Cov}}(U)=\sum_i(Z_i,X_i)'\widehat{\mathrm{var}(Y_i)}(Z_i,X_i)=\sum_i(Z_i,X_i)'(Y_i-\hat{\mu_i})(Y_i-\hat{\mu_i})'(Z_i,X_i)=\begin{pmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix} \tag{13}$$

where $\hat{\mu}_i$ is an estimator of μ_i, $\widetilde{\Sigma}$ is an estimate of the covariance of score (U) vector. $\widetilde{\Sigma}$ is partitioned with the dimensions according to the score vector component $U_{.1}$ and $U_{.2}$

for φ and β respectively.

As we use the identity link for longitudinal continuous trait, i.e. $g(\mu_{im}) = \mu_{im}$ and $v(\mu_{im}) = \phi \times 1 = \phi$, we will have:

$$
\begin{aligned}
U &= \sum_i (Z_i, X_i)' R_w^{-1} (Y_i - \mu_i) \\
\widetilde{\Sigma} &= \sum_i (Z_i, X_i)' R_w^{-1} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' R_w^{-1} (Z_i, X_i)
\end{aligned} \tag{14}
$$

if the assumption of a common covariance matrices across Y_i for i is valid, for example for quantitative continuous traits study [Pan, 2001], 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 [Pan, 2001].

Although repeated measurements are usually for continuous traits, our method can be easily adapted to longitudinal binary trait. In brief, the only difference between dealing with continuous trait and dichotomous trait is the canonical link function. 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

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

3.2.2 Several Existing Association Tests

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 13.

The classical score test statistic is $T = U_{.2}'\Sigma_{.2}^{-1}U_{.2}$, which, however, will lose power when SNP number p increases with a fixed sample size n. As shown theoretically [Fan, 1996], as the dimension p increases, the power of the score test will diminish, tending to the type I error rate α. The popular UminP test statistic is $T = \underset{j}{max}\frac{U_{.2,j}^2}{\Sigma_{.2,jj}}$ with $\Sigma_{.2,jj}$ is the jth entry on the diagonal of $\Sigma_{.2}$. It might also has low power if we have many small $|\beta_j| \neq 0$. Two alternatives, called the Sum and SSU tests, are

$$T_{Sum} = 1'U/\sqrt{1'V1} = \sum_{j=1}^{p} U_j/\sqrt{1'V1}, \qquad T_{SSU} = U'U = \sum_{j=1}^{p} U_j^2.$$

The Sum test is closely related to other burden tests such like those in [Morgenthaler and Thilly, 2007, Li and Leal, 2008, Madsen and Browning, 2009]. If there is a common association either in direction or strength for causal SNVs with no or few non-associated SNVs, then Sum test and the likes will be most powerful; otherwise, the SSU test and its closely relatives, such as kernel machine regression (KMR or SKAT) [Lee et al., 2012a, Ionita-Laza et al., 2013, Oualkacha et al., 2013, Lee et al., 2012b, Wu et al., 2011] and C-alpha test [Neale et al., 2011], will be more powerful. Nevertheless, as shown in Pan et al., [Pan et al., 2015a, Pan et al., 2014] a variance-component test is not adaptive and may lose power in the presence of many non-associated SNPs. Accordingly, a more powerful and adaptive test (aSPU) extended to the longitudinal version (LaSPU) is proposed next.

3.2.3 The Data-Adaptive LaSPU Test

Is is straightforward to see both the Sum test statistic and the SSU test statistic are based on score vector but using different weights. A more general form of the score-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 SNPs [Lin and Tang, 2011]. The aSPU test proposed $W_j = U_{.2,j}^{\gamma-1}$ 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 $\gamma = 1$, the SPU(1) test uses $\mathbf{1}$ as weight and sums up the information contained in all the SNPs in the region of interest, equivalent to Sum test or burden test; when $\gamma = 2$, the SPU(2) test uses U as weight to itself and is equivalent to SSU test and other variance-component test such as SKAT; when γ keeps increasing, the SPU(γ) test puts higher weights on the jth SNP with larger $|U_{.2,j}|$, while gradually decreasing the weights of other SNPs with smaller $|U_{.2,j}|$. As the large value of $|U_{.2,j}|$ indicates strong association information stored in SNP j and small value of $|U_{.2,j}|$ indicates weak or none association information stored in SNP j, a higher γ tends to put more and more weights on those informative SNPs. When $\gamma \to \infty$ as an extreme situation, where ∞ is assumed to be an even number, we have

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

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

A simulation based method [Lin, 2005, Seaman and Müller-Myhsok, 2005] could be used to calculate the p value for SPU test. Specifically, suppose T is a short notation of $T_{SPU(\gamma)}$ for a specific γ and $\hat{\Sigma}_{.2}$ is the covariance matrix of the score vector $U_{.2}$ from the original data (Equation 13). 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 $P_{SPU(\gamma)} = \sum_{b=1}^{B} \frac{I(T^{(b)} \geq T^{obs})+1}{B+1}$. We used $B = 1000$ in our simulations for a nominal significance level at 5%.

There is no uniformly most powerful test in set-based association testing; on the other hand, it has been found empirically that the Sum, SSU and UminP tests are preferred over different scenarios. For a given dataset, to adaptively choose the value of γ for the SPU tests, Pan et al. [Pan et al., 2014] proposed an adaptive SPU (aSPU) test that simply combines the results of a series of SPU tests: suppose we have some candidate value of $\gamma \in \Gamma = \{1, 2, \ldots, 8, \infty\}$ as used in our later experiments, and suppose the p vaule of the SPU(γ) test is $P_{SPU(\gamma)}$, then the aSPU test simply takes the minimum p value:

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

Of course, T_{aSPU} is no longer a genuine p value; we recourse to the previous simulation based strategy to obtain its p value. Specifically, since we have the previous simulated $U_{.2}^{(b)}$ and computed $T_{SPU(\gamma)}^{(b)}$ for $b = 1, 2, \ldots, B$ and $\gamma \in \Gamma$, we can directly calculate the p

value for $T^{(b)}_{SPU(\gamma)}$:

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

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:

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

Because we have calculated U and $\hat{\Sigma}$ at the beginning by solving the GEE of the generalized linear model already taking into account the repeated measurements, the LaSPU method will smoothly inherit the aSPU essence. In practice for genome wide scan purpose, we can use a stage-wise LaSPU test strategy: we first start with a smaller B, for example, use $B = 1000$ to scan the genomes, then gradually increase B to, for example, 10^6 for a few selected groups of SNPs. For example, we could choose specific genes or windows which passed a pre-determined significance cutoff (for example, p-value $\leq 5/B$) in the previous stage; we then repeat this process until the pre-determined significance level is reached. For example, a p-value of $\leq 10^{-6}$ requires we increase $B \geq 10^6$. In this stage-wise way, we will be able to apply the LaSPU test to GWAS data.

3.2.4 Other Modifications

LaSPU weighted (LaSPUw) test

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

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

Accordingly, the aSPUw test statistic is defined as

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

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

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

LaSPUw test retains the information of SNP's variance and is sometimes more powerful than LaSPU test. It automatically adjusts the weight of each SNP in a region by its variance, which is usually determined by SNP's MAF. It is worth noting that LaSPU and LaSPUw test can be implemented in the meantime using the same simulated score vector, which makes the computation efficient.

LaSPU(w).Score test

Although the GEE Score test will lose power in some scenario of the gene-based GWAS analysis as aforementioned, it still has the unique advantage in some scenarios when the correlation structure among SNPs matters and dimension p is not too large. GEE Score test in the form of $T = U'_{.2}\Sigma^{-1}_{.2}U^{-1}_{.2}$ will keep the covariance matrix in the denominator, which preserves the information of possible linkage disequilibrium among SNPs. To combine the strengths of GEE Score test and the aSPU(w) test, we propose to adopt the minimum P-value combining strategy, yielding the aSPU(w).Score test statistic:

$$T_{aSPU.Score} = \min\Big\{\underset{\gamma\in\Gamma}{\min} P_{SPU(\gamma)}, P_{Score}\Big\}.$$

where P_{Score} is the p-value of the Score test.

LaSPU omnibus (LaSPU.o) test

Among aSPU, aSPUw, and the Score test, there may be at least one test having a higher power in a specific data scenario, depending on the association pattern and the correlation structure within SNPs. We thus propose to combine these three tests into the aSPU omnibus test.

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

3.2.5 Extension to Rare Variant Analysis

While MAF of RVs are usually low, for example, between 0.001 to 0.01, the property of asymptotically normal distribution of either the regression coefficient or score vector may not hold. The simulation based strategy in LaSPU (mainly for CVs) may not be sufficient for RV analysis. It will be nice if we can provide an alternative to estimate the p value more precisely for RV analysis in LaSPU method. Thereby, we propose a permutation based method to generates the empirical null distribution of $U_{.2}^{(b)}$ instead of the previous simulation based method. The permutation strategy will not break the time-dependent measurement order or the association between the longitudinal trait and possible covariates, but only permute the genotype codes across individuals. It will be also robust to missing data as this usually happens in the longitudinal data analysis.

The permutation strategy can be implemented as follows:

1. identify the max k across all n subjects, which is the number of longitudinal measurements. For example, $k = 4$ to illustrate.

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 $Y_i = (y_{i,1}, \text{NA}, \text{NA}, y_{i,4})'$). 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., $k \times (q+p+1)$, for covariates and SNPs. Now we should have $\begin{pmatrix} Y_i & H_i \end{pmatrix}$ as an augmented matrix of dimension $k \times (q+p+2)$ 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 $nk \times (q+p+2)$.

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

After we obtain $U_{.2}^{*(b)}$ pool from permutation strategy, we could have an empirical null distribution as we did for simulation based strategy. Thus, the left work of the LaSPU test becomes the same for both strategies.

In the real data application involving RVs, we recommend the combined strategy that users could first run simulation based LaSPU test to obtain the putative significant

genetic loci, and then apply the more time-consuming permutation based LaSPU test on those selected loci as validation.

3.2.6 Simulation Set-ups

We conducted extensive simulation studies to evaluate and compare the performance of the LaSPU tests with several alternative methods. As a summary, we simulated genotypes following Wang and Elston. [Wang and Elston, 2007] In brief, we generated two independent blocks of SNPs for each subject with each block has the first-order auto-regression (AR(1)) correlation structure to imitate the real-world LD among SNPs: the first block will included causal SNPs and null SNPs in linkage disequilibrium (LD); the second block will include only null SNPs in LD. We simulated the longitudinal trait using AR(1) correlation structure following [Song et al., 2013]. We took into account the SNP main effect and time course main effect as fixed effects on the longitudinal trait without consideration of SNP by time interaction. For simplicity, we did not include other covariate effects (such as demographic variables) in the simulation studies, though they can be simply added. We referred to previous studies [Pan et al., 2014, Basu and Pan, 2011, Pan and Shen, 2011, Han and Pan, 2010, Pan, 2009] and the ARIC data to set up the simulation parameters, for example, ρ_y across longitudinal measurements and ρ_x across SNPs as used in AR(1) correlation structure model, so that our simulation can better approximate real data analysis in the later.

Methods for simulation of genotype data

First, a latent vector $G_i = (G_{i1}, \ldots, G_{ip})'$ was drawn from a multivariate normal distribution $MVN(0, R)$, 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 $f \neq g$. The default ρ was set at 0.8 to mimic the real data.

Second, the latent vector was dichotomized to yield a haplotype with each latent element G_{ij} dichotomized to 0 or 1 with probability Prob($G_{ij} = 1$) = MAF of jth SNP; the MAFs were randomly drawn from a uniform distribution between 0.05 and 0.4 for CVs or between 0.001 and 0.01 for RVs. Third, we combined two independent haplotypes to obtain the genotype $X_i = (X_{i1}, \ldots, X_{ip})'$ for subject i. Using this strategy, we placed 35 SNPs in the first block with AR(1). Out of 35 SNPs, we randomly choose 5 SNPs to be causal. To mimic the case of GWAS, for example, tag SNPs are genotyped but not the causal SNPs, we excluded the 5 causal SNPs from the later test procedure for CV simulation study. We further placed 15 null SNPs in the second block, of which the number could be further increased to increase the signal sparsity of the region. Note that the first block and second block are independent to each other.

Methods for simulation of longitudinal trait

We first obtained the estimates of the parameters used in this simulation from a preliminary analysis of ARIC dataset. For example, we estimate $\rho = 0.7$ for repeated measurements of HDL-C trait within the same subject. Secondly, we set-up the mixed effect model to achieve the AR(1) correlation structure as:

$$y_{im} = \mu_i + b_i + \underbrace{\rho e_{i,m-1} + s_{i,m}}_{e_{i,m}}, \tag{15}$$

with $m = 1, \ldots, k.$ indexes the longitudinal measurements within subject i; $\mu_i = Z_i\varphi + X_i\beta = H_i\theta$, where time as a covariate and time effect as a parameter are included in Z_i and φ respectively; 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. $e_{i,m}$ is the total residual, which can be divided into two parts: the first part depends on $e_{i,m-1}$ and the second part is an inde-

pendent term. We assume they follow:

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

It is straightforward to see the sum $\rho e_{i,m-1} + s_{i,m}$'s variance is equal to the variance of $e_{i,m}$ by algebraically summing up two parts. Under this assumption, the variance-covariance matrix across longitudinal measurements becomes (for $k = 4$ as the number of longitudinal measurements):

$$
\Sigma_{4\times 4} = Var \begin{pmatrix} b_i + e_{i1} \\ b_i + \rho e_{i1} + s_{i2} \\ b_i + \rho e_{i2} + s_{i3} \\ b_i + \rho e_{i3} + s_{i4} \end{pmatrix} = \sigma_b^2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \tag{16}
$$

On the rightmost end of the equation, the first part defines the within-subject variances, and the second part allows the measurements with a k-interval lag to have a correlation coefficient of ρ^k. This is closer to reality in some cases for longitudinal data.

Methods for tuning simulated genetic effect

As noted in association tests, different SNPs contribute to the trait with different effect sizes. However, the SNP effect magnitude tuning in the simulation study is not trivial. Instead of assigning a β_d coefficient to a SNP with an arbitrary numerical value, for example, 0.1 or 10000, there is a way to use genetic heritability to control the association magnitude of the jth SNP [Lynch et al., 1998]. We first introduce the formula of the

variance of the phenotype :

$$Var(y_{im}) = Var(X_{ij})\beta_j^2 + \sigma_{oth}^2 = 2f(1-f)\beta_j^2 + \sigma_{oth}^2 \tag{17}$$

where the Hard-Weinberg equilibrium (HWE) is assumed to hold. f is the MAF of the SNP; σ_{oth}^2 is the residual variance after removing the effect of jth SNP. Obviously we can see σ_b^2 and σ_e^2 are contained in σ_{oth}^2 (see equation (27)). If other SNPs' effects 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 (29):

$$h^2 = \frac{Var(A)}{Var(P)} \tag{18}$$

This is the classical formula of narrow-sense heritability, with $Var(A)$ represents the variance due to the additive effects of the alleles, and $Var(P)$ represents the total variance in the phenotype. In our situation for jth SNP, this can be expanded to:

$$h_j^2 = \frac{Var_j(A)}{Var(P)} = \frac{Var(X_{ij})\beta_j^2}{Var(y_{im})} = \frac{Var(y_{im}) - \sigma_{oth}^2}{Var(y_{im})} \approx \frac{Var(y_{im}) - \sigma_b^2 - \sigma_e^2}{Var(y_{im})} \tag{19}$$

By systematically solving the equations (29) and (31), we can easily calculate the β_j for jth SNP once we have pre-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 be small for complex disease, and we will use $h_j^2 = 0.001$ as default in the simulation studies to obtain β_j unless otherwise specified.

Overall Settings in Simulation

We fixed the test significance level at $\alpha = 0.05$. We ran 1000 replicates for all participating tests, including LaSPU, GEE-Score, UminP, SSU, SSU weighted (SSUw) and Sum tests. We used $B = 1000$ for either simulation or permutation strategy whenever encountered. To investigate the type I error and the power performances under different sample sizes,

we used total four measurements simulated on sample sizes of 500, 1000, 2000 and 3000. We put 15 SNPs in block one with 5 out of them are causal SNPs; we put 35 SNPs in block two, all of which are null SNPs. causal SNPs are excluded from the final test for CVs to mimic the genotyping array while still kept in the final test for RVs to mimic the sequencing platform. We set the MAF of SNPs between 0.05 and 0.4 for CVs and between 0.001 and 0.01 for RVs. In the setting A, we let $h_j^2 = 0$ to set $\beta_j = 0$ (i.e., no causal SNPs) when we were testing the type I error of those methods; in the setting B, we set causal SNPs' $h_j^2 = 0.001$ to derive their $\beta_j \neq 0$ when we were testing the empirical power of those methods. In the setting C, to investigate the power performance under heterogeneous SNP effects in a region, we flipped the signs of 3 causal SNPs out of 5 total. In the setting D, to investigate the power performance under different signal sparsenesses of a region, we increased the null SNPs in second block (as independent to first block) gradually from 50 to 75, 100, 200 and 400, where the first block remained 15 SNPs and 5 out of 15 are causal SNPs as default. In the setting E, to investigate the power performance due to adding repeated measurements, we benchmarked the LaSPU.omni test starting from the baseline measurement to total five measurements with a step size of one measurement. We allowed one more measurement from previous simulated four measurements for better illustration purpose.

3.3 Results

3.3.1 Simulation Results for CVs

Type I Error

In the setting A, as shown in Table 3, it appears that all the tests maintained the nominal level ($\alpha = 0.05$) of Type I Error rate well.

n	pSSU	pSSUw	pScore	pSum	pUminP	SPU(1)	SPUw(1)	SPU(2)	SPUw(2)	SPU(∞)	SPUw(∞)	LaSPU	LaSPUw	LaSPU.sco	LaSPUw.sco	LaSPU_LaSPUw	LaSPU.omni
500	0.047	0.048	0.047	0.052	0.048	0.052	0.051	0.058	0.059	0.054	0.053	0.060	0.058	0.058	0.055	0.060	0.058
1000	0.047	0.046	0.044	0.058	0.048	0.057	0.057	0.059	0.056	0.051	0.051	0.057	0.057	0.057	0.054	0.057	0.057
2000	0.049	0.047	0.051	0.048	0.048	0.048	0.049	0.058	0.055	0.052	0.052	0.061	0.058	0.058	0.058	0.058	0.058
3000	0.051	0.052	0.052	0.052	0.050	0.051	0.052	0.061	0.060	0.054	0.053	0.063	0.060	0.059	0.057	0.062	0.059

Table 3: Empirical Type I Error Table in the simulation setting A.

Comparison of the LaSPU Test with Other Tests

Now let we look at the empirical power benchmark among these tests:

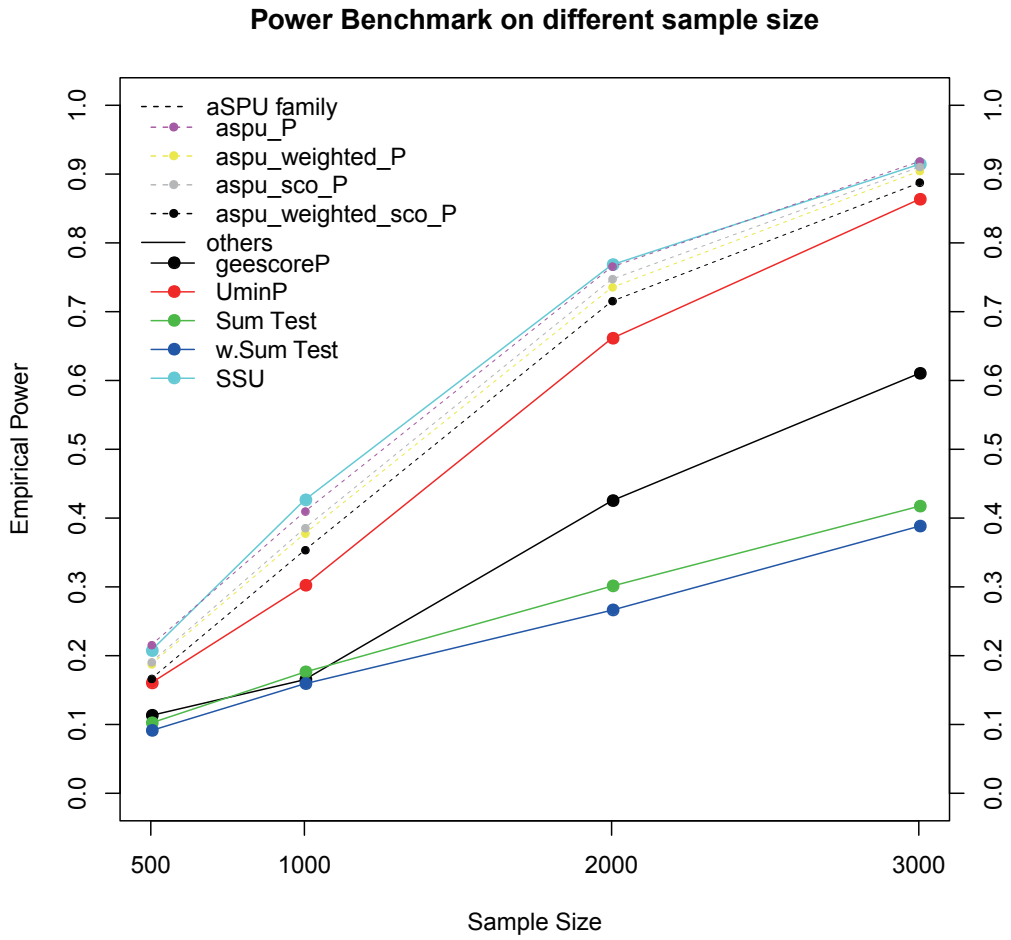

Figure 4: Empirical power benchmark under different n in the simulation setting B.

In the setting B, as shown in Figure 4, LaSPU family performed the best together with SSU test, which is close related to SPU(2) test. UminP test was after them, then followed by the GEE Score Test. Sum test and Sum weighted test were the last. The observations indicates that the SPU(2) was powerful in this testing scenario, which boosted the LaSPU family performance, while SPU(1) with its low performance was ignored by LaSPU family. A closer look within LaSPU family also tells us the original LaSPU test performed the

best while other modifications were close to it.

Tests with half number of SNPs in opposite effect direction

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

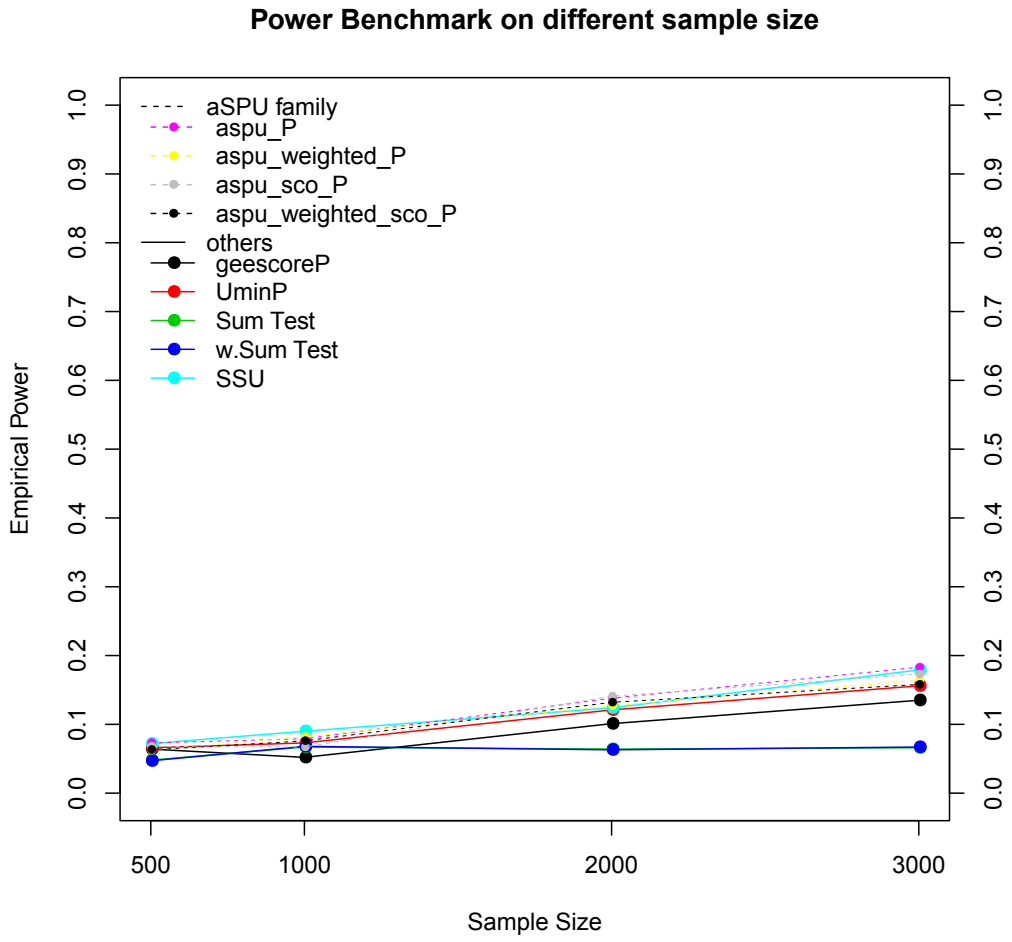

Figure 5: Empirical power benchmark under a heterogeneous SNP effects in the simulation setting C.

As shown in Figure 5, we can see all the tests lost their power a lot. LaSPU family still performed the best together with SSU test. In this scenario, we still find SPU(2) boosted

the performance of LaSPU family, while SPU(1) related test, such as Sum test, was not only of lower power, but also did not increase the power at all when n goes up!

Tests with growing number of Null SNPs

In the setting D, we gradually increased the number of null SNPs in second block from 50 to 75, 100, 200, then finally 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:

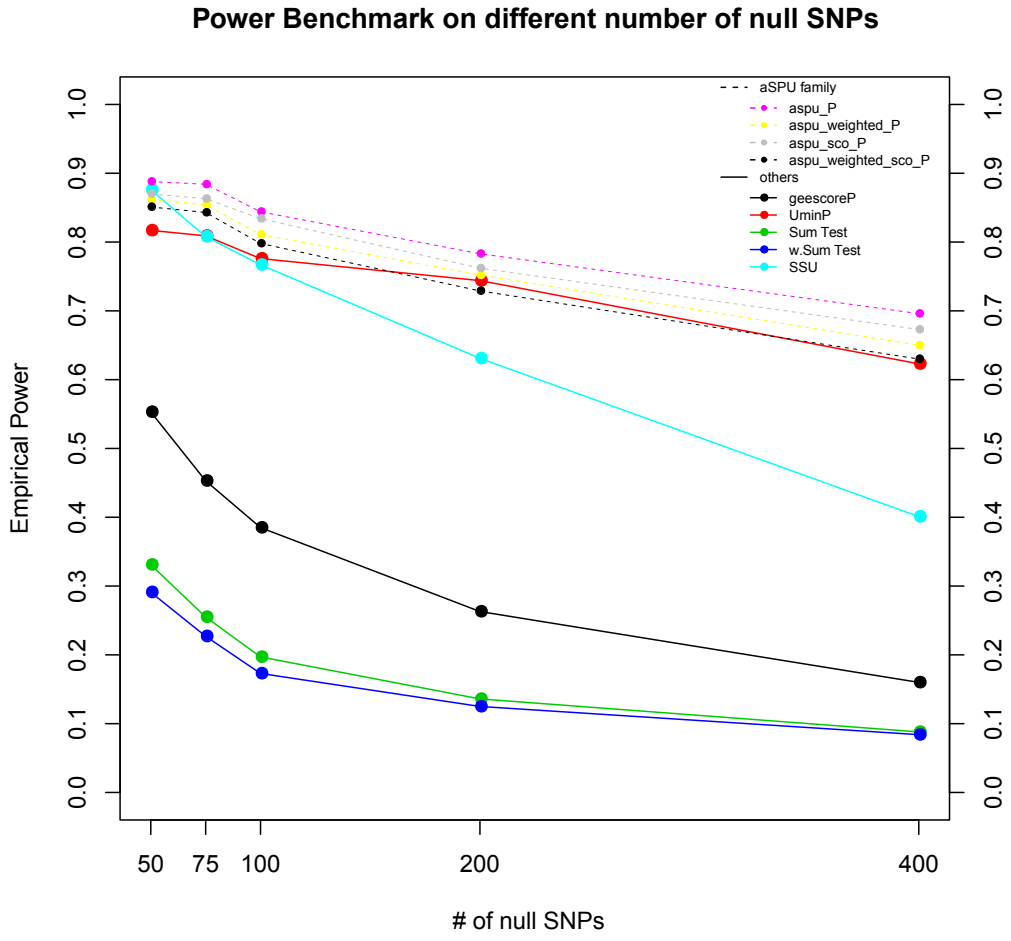

Figure 6: Empirical power benchmark under an increased number of null SNPs in the simulation setting D.

As shown in Figure 6, as the number of null SNPs increases, the LaSPU family members revealed their greatest potential to maintain a higher power than any other test. The champion was the original LaSPU test, which only decreased from a power of 0.9 to 0.7, while the worst one weighted Sum test decreased from a power of 0.3 to below 0.1. GEE Score test also lost its power quickly. The SSU test suddenly lost its power with a steep linear slope after 100 null SNPs, indicating SPU(2) was not sufficient any more in maintaining a high power for LaSPU family with a larger number of null SNPs. UminP test (as equivalent to SPU(∞)) performed quite well as it just followed the bottom of LaSPU family members, i.e., LaSPUw.Sco. This observation has at least two interpretations: first, UminP will be more and more sufficient when SNP number grows to ∞ so that number of causal SNPs will be effectively one; second, LaSPU family performance in this scenario was actually more and more boosted by SPU(γ) with γ between 3 and 8, as neither SPU(2) nor SPU(∞) performed better than any member of the LaSPU family. It is reasonable to infer when null SNP number increased from 50 to 400, a specific test from SPU(3) to SPU(8) took its advantage and thus led the LaSPU family to the highest power position.

This scenario is more the case encountered in the genetic data recorded from the next-generation sequencing (NG-seq) platforms. By using NG-seq techniques, high-density variants can be identified, and then LaSPU methods can be applied accordingly.

Comparison of the LaSPU Test with Its Other Variants

As with the same setting B, we added more tests to investigate the performance details of various LaSPU variants as compared to each other and compared to the other methods.

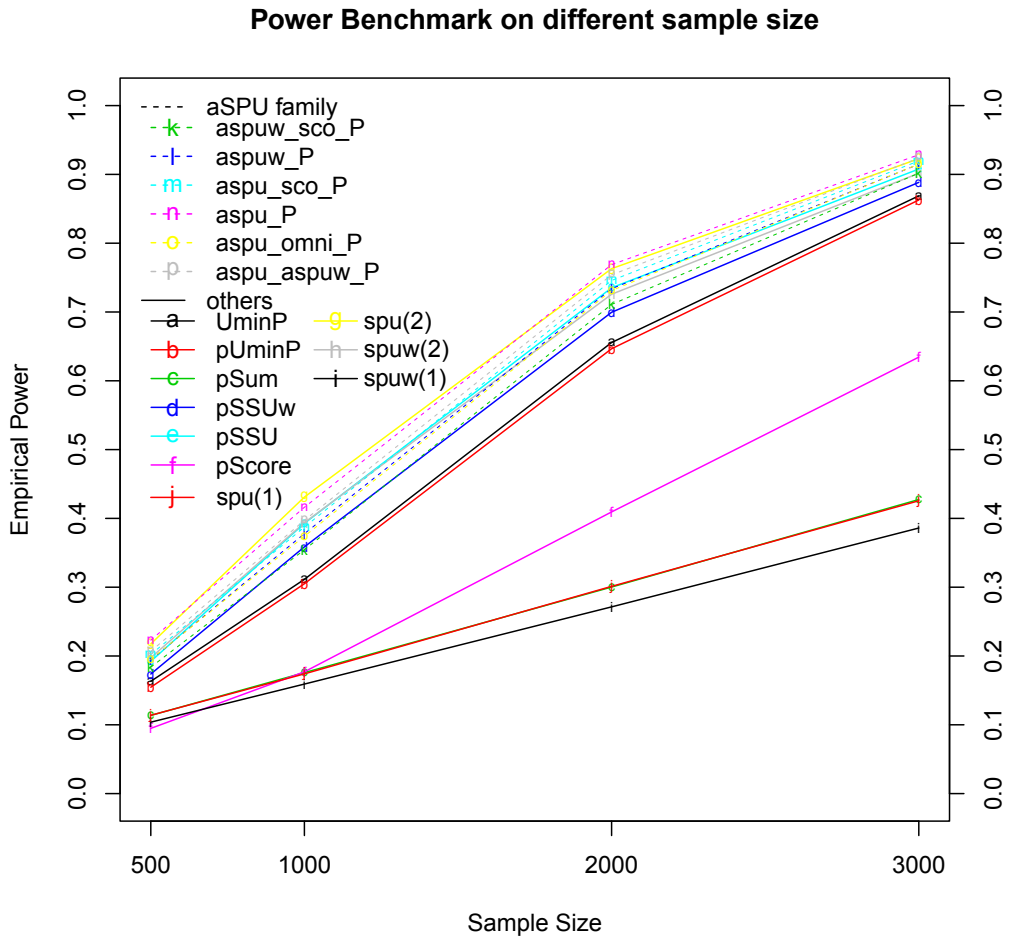

Figure 7: Empirical power benchmark with more tests in the simulation setting B.

As shown in Figure 7, LaSPU family tests dominated the other tests. LaSPU test was the best overall after sample size 2000, followed closely by SPU(2) test. SPU(2) test slightly wined over LaSPU test with the sample size of 1000. Apparently, it performed well and contributed to the LaSPU family tests' performances. The SPU(2) test's edge over its variant SPUw(2) test may indicate that, ignoring the diagonal variance of SNPs, sometimes, could help boost the power. Although the original LaSPU test was the best among the LaSPU family tests, LaSPU_LaSPUw test and LaSPU.score test were close to it, and even the LaSPU.omni test was close to them. This fact demonstrated the

robustness of LaSPU test with regard to its modification s. In some scenarios, we do not know which variant is the best in advance, but with some partial information, for example, score test may also have a good power when LD structure among SNPs matters and there are not so many SNPs in a region, then we can use the LaSPU.score test instead of LaSPU original test to achieve a better power. Similarly, when we have no information at all, we would prefer the LaSPU omnibus test, which combines the strengths of LaSPU, LaSPU weighted and Score tests in one test. It will perform sub-optimal to the unknown best LaSPU (variant) test. In real world GWAS, a more well-rounded data-adaptive association test with less risk to lose power in unknown scenarios with an negligible performance compromise may be preferred.

Power Gain from Repeated Measurements in Longitudinal Studies

In the setting E, we demonstrated the power gain from longitudinal studies by introducing different power at different number of repeated measurements under different sample sizes. We used LaSPU.omni test for illustration purpose here. As shown in Figure 9, both larger sample size and more measurements increased the power of association test. The power gain from adding more repeated measurements seemed most obvious with the sample size of 2000. This fact demonstrates the subtle relationship between the power and the sample size or the number of measurements. In longitudinal GWAS, an optimal combination of a specific sample size and a specific number of measurements could lead to the maximal power and the lowest study cost, which is of interest to researchers. Here, as an example, we can choose to recruit 2000 subjects with each one measured four times, then the study power will be equivalent to that we recruit 3000 subject with only baseline measurement. Usually recruiting more subjects, e.g., a 50% percentage more here, will cost more than following up with the subjects.

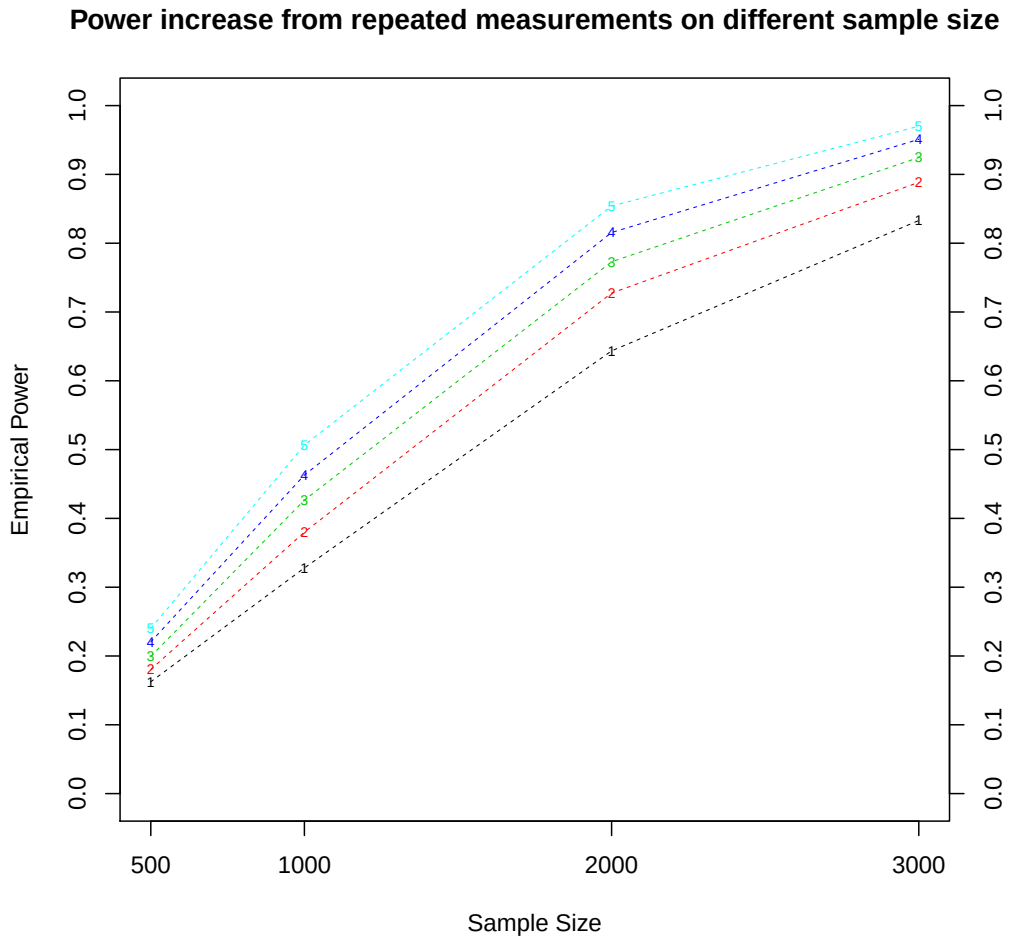

Figure 8: Power increases from repeated measurements in the simulation setting E.

3.3.2 Simulation results for RVs

With the increasing availability of sequencing data, it has become more important and urgent that the data-adaptive association test can be applied on RVs in longitudinal studies. For this purpose, we did a simulation study to assess the performance and show the potential of our proposed test for analysis of RVs in longitudinal studies. To save space, we present results only for simulation set-ups similar to the setting A and B. We call the settings A' and B' for RV simulation studies. In common for two settings, we put

15 SNPs in block one with 5 out of them are causal SNPs and causal ones were retained in the final test to mimic sequencing platform; we put 35 SNPs in block two, all of which are null RVs. We set the MAF of SNPs between 0.001 and 0.01 for RVs. In the setting A', we let $h_j^2 = 0$ to set $\beta_j = 0$ (i.e., no causal SNPs) when we were testing the type I error of those methods; in the setting B', we set causal SNPs' $h_j^2 = 0.001$ to derive their $\beta_j \neq 0$ when we were testing the empirical power of those methods.

In the setting A', as the result shown in Table 4, all the tests could maintain Type I Error rate at nominal level well.

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

Table 4: Empirical Type I Error Table in the simulation setting A'.

In the setting B', as the result shown in Figure 9, LaSPU weighted version was the overall winner with LaSPUw.score test, LaSPU_LaSPUw test and LaSPU.omni test closely followed by. Apparently, in the RV analysis, the weighted version of LaSPU, which weighs each SNP by its variance, largely determined by its MAF, had an edge over the original LaSPU test. Therefore, the LaSPUw variants, such as LaSPUw.score test, LaSPU_LaSPUw test and LaSPU.omni test, had good performance as well. Other than LaSPUw test and its variants, the SSUw and its equivalent SPUw(2) performed relatively well as they just followed the LaSPU.omni test. This fact confirmed the SNP weights are quite informative in RV analysis and thus boost the power. Original version of LaSPU test and LaSPU.score test were mixed with SSU and SPU(2) tests, after the SSUw and SPUw(2) tests. UminP test performed okay, but still it has been surpassed by the tests mentioned before with a relatively large gap in power. The classical GEE Score test followed after the UminP test. While it was the weakest test at $n = 500$, it became more powerful at bigger sample sizes gradually. The observation on Score test may infer: 1), the whole variance-covariance matrix of SNPs is not necessarily better than the variance of the SNP as extra information to boost the power. 2), while sample size is small, the Score test could not benefit from the variance-covariance matrix information or the variance-covariance matrix estimate is more likely to be biased, so that the test power was damaged here. At last, Sum, SPU(1) and SPUw(1) tests performed poorest in this simulation, which repeated the observation of the simulation setting B for CVs.

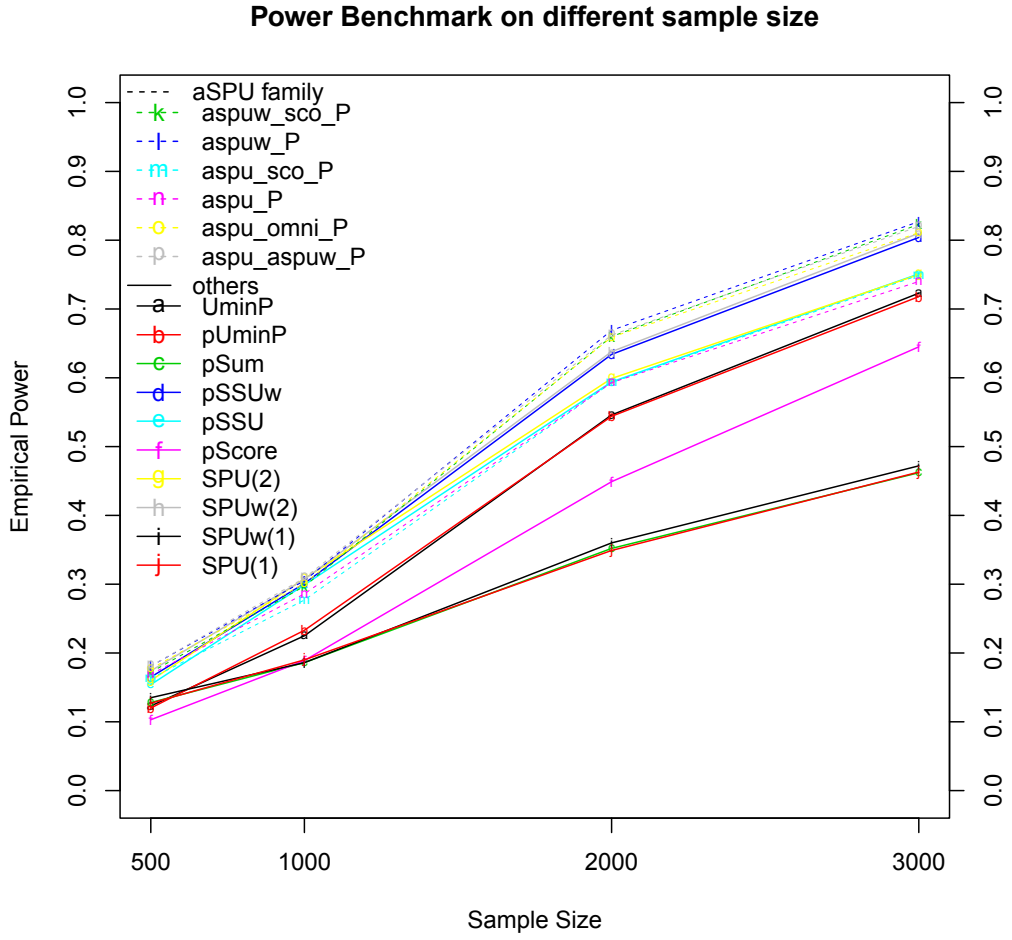

Figure 9: Empirical power benchmark under different n in the simulation setting B'.

3.3.3 Application to the ARIC study

We illustrated the proposed LaSPU method using the data from the Atherosclerosis Risk in Communities (ARIC) Study [Heiss, 1989]. ARIC study is a prospective epidemiological study designed to investigate the causes of atherosclerosis and its clinical outcomes, the variation in cardiovascular risk factors, medical care, and disease by race, gender, location, and time. We used the data from the ARIC study cohort component, which 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 examination (baseline) occurring in 1987-89, the second in 1990-92, the third in 1993-95, and the fourth exam was in 1996-98. The fifth examination was farther apart from the previous screens and was finished during 2011-2013. A detailed description of the ARIC study design and methods was published elsewhere [Investigators et al., 1989]. We exclusively used the Caucasian samples ($n = 11478$) in the ARIC cohort dataset with four available measurements. We used HDL-C as the phenotype of interest. For the genotype data part, we used rare variants defined by MAF $< 5\%$ genotyped on HumanExome BeadChip v.1.0 (Illumina).[Grove et al., 2013] We applied standard quality control (QC) procedures on individual samples, which includes checking concordance to GWAS data and excluding those individuals missing $> 5\%$ genotypes, population clustering outliers, individuals with high inbreeding coefficients or heterozygote rates, individuals with gender mismatches, and individuals with an unexpectedly high proportion of identity-by-descent sharing, with consideration for family studies, based on high-quality variants. Following the literature on association testing of rare variants, we assumed the additive genetic model and coded a variants as 0, 1, or 2, i.e., the number of the minor alleles. With respect to covariates, we included the top two principal components eigenvectors (PCs) produced by EIGENSTRAT [Price et al., 2006] in the generalized linear model to adjust for the potential population stratification within the ARIC Caucasian subjects. Additionally, we included a subject's demographic information such as age, age^2, gender and BMI. We also included the time effect as a covariate. We ran gene-based test with gene boundary defined in the analysis of Peloso et.al.[Peloso et al., 2014] We only included nonsynonymous and splice site variants. We employed a stage-wise permutation strategy for the LaSPU test: we first performed 1,000 permutations for all gene regions and then increased to 1,000,000 permutations for those gene regions with p values < 0.01 in the first stage. We set the significance threshold

at $0.05/25,000$ = 2e-06 to control the family-wise error rate (FWER) at 0.05 based on the Bonferroni correction for 25,000 human coding genes. In the real data application involving RVs, we recommend a stage-wise strategy that users could first run the simulation based LaSPU test to obtain a small number of putative significant SNP sets at a liberal threshold, e.g., $p < 0.01$, and then apply the more time-consuming permutation based LaSPU test on those putative significant SNP sets as validation (with the minimal number of permutations necessary, i.e., 500,000 here, to achieve a significance level at 2e-06.)

Figure 10 shows the manhattan plot comparison between baseline study and longitudinal study by aSPU and LaSPU test, respectively, on the association between HDL-C and rare variants in the ARIC study. Panel A shows the result by conducting baseline measurement analysis, where only one gene *LIPC* was identified to be significant; Panel B shows the result by conducting longitudinal analysis using four repeated measurements, where four genes were identified as significant including *LPL*, *LIPG*, *ANGPTL4* and *FAM65A*. After comparison with the results of Peloso et al., [Peloso et al., 2014] we found that *LIPC* and *FAM65A* have not been reported before, while *LPL*, *LIPG*, *ANGPTL4* are confirmed genes associated with HDL-C. In summary, the baseline analysis of the ARIC data alone (11478 samples) was not able to identify the signals previously found in a study with a much larger number of cohorts and samples (42,208 aggregated samples) with the ARIC study being a subset. Surprisingly, in contrast, the longitudinal data analysis with four repeated measurements and the same sample size (11478 samples) helped boost the power and was able to identify the same signals as in the Peloso et al.'s meta-analysis of baseline HDL-C. Peloso et al. also reported *APOC3* as a known gene and *ANGPTL8* (alias: *C19orf80*) and *PAFAH1B2* as novel genes associated with HDL-C. Although the LaSPU test did not identify these signals as genome-wide significant, it identified *APOC3* and *ANGPTL8* as marginally significant genes, defined by

a p value < 0.001. Specifically, LaSPU test gave a p value of 5.87e-04 for *APOC3* and 1.06e-04 for *ANGPTL8*. LaSPU test missed the novel signal *PAFAH1B2* with a p value of 0.0022. Compared to LaSPU test, the baseline measurement analysis by aSPU did not identify any previously reported signals as significant and only identified *LPL* and *APOC3* as marginally significant (p value of 0.001 for both loci). Table 5 summarized the comparison results above.

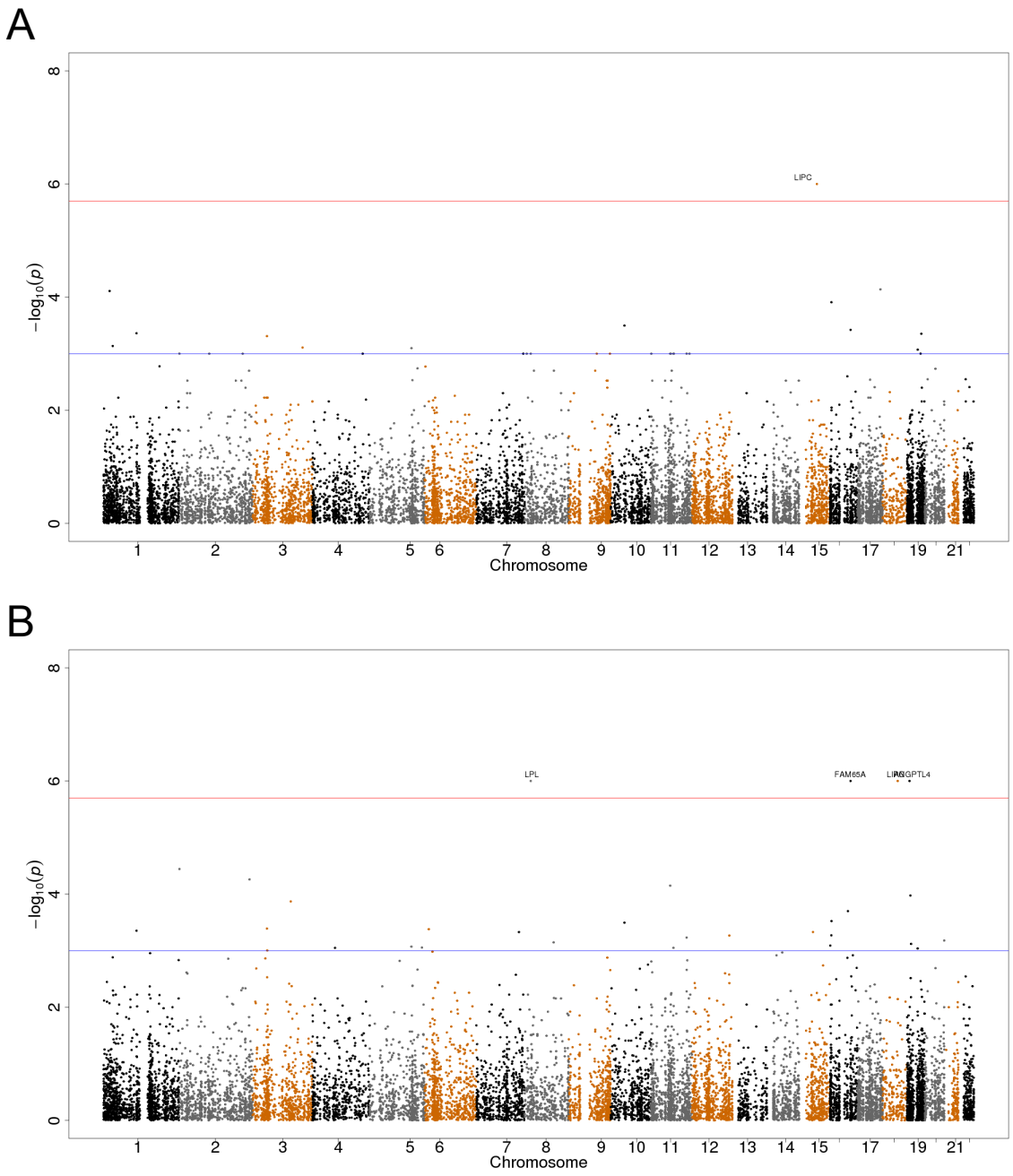

Figure 10: Manhattan Plot Comparison between baseline study and longitudinal study by LaSPU test on the association between HDL-C and Rare Variants in the ARIC study. A. baseline study; B. longitudinal study using total four measurements.

Table 5: 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)

We also compared the results of LaSPU test and Longitudinal SSU (LSSU) test. The latter is close to the SKAT-like test used in longitudinal data analysis.[He et al., 2015] Figure 11 shows that the LSSU was also able to identify the previously reported *LPL*, *LIPG* and *ANGPTL4* as significant genes. In addition, LSSU has also identified *APOC3* and *ANGPTL8* as marginally significant genes with a p values of 3.87e-04 and 4.22e-05, respectively. LSSU test also missed the novel signal *PAFAH1B2* with a p value of 0.0018. Finally, LSSU test did not identify *FAM65A* as a significant gene as did LaSPU, but reported it as marginally significant (p value of 1.25e-04). In summary, we conclude that LSSU was as powerful as LaSPU in this analysis. Of note, LSSU is not a data-adaptive test so it may not perform robustly well in other scenarios, such as those with a lot of null SNPs diluting the association signals in a SNP set.[Pan et al., 2015b] On the other hand, apparently, LaSPU mainly benefited from the SPU(2) test in this analysis. Since SPU(2)

is equivalent to SSU test, this also explained why LaSPU did not have a remarkable edge over LSSU here. It is noted that, we found the novel gene *FAM65A* identified by LaSPU test was identified as marginally significant by LSSU test, but genome-wide significant by Longitudinal Sum (LSum) test (p value of 1.99e-07). This suggests that LaSPU was able to adaptively combined the strengths of a class of tests, including LSum and LSSU.

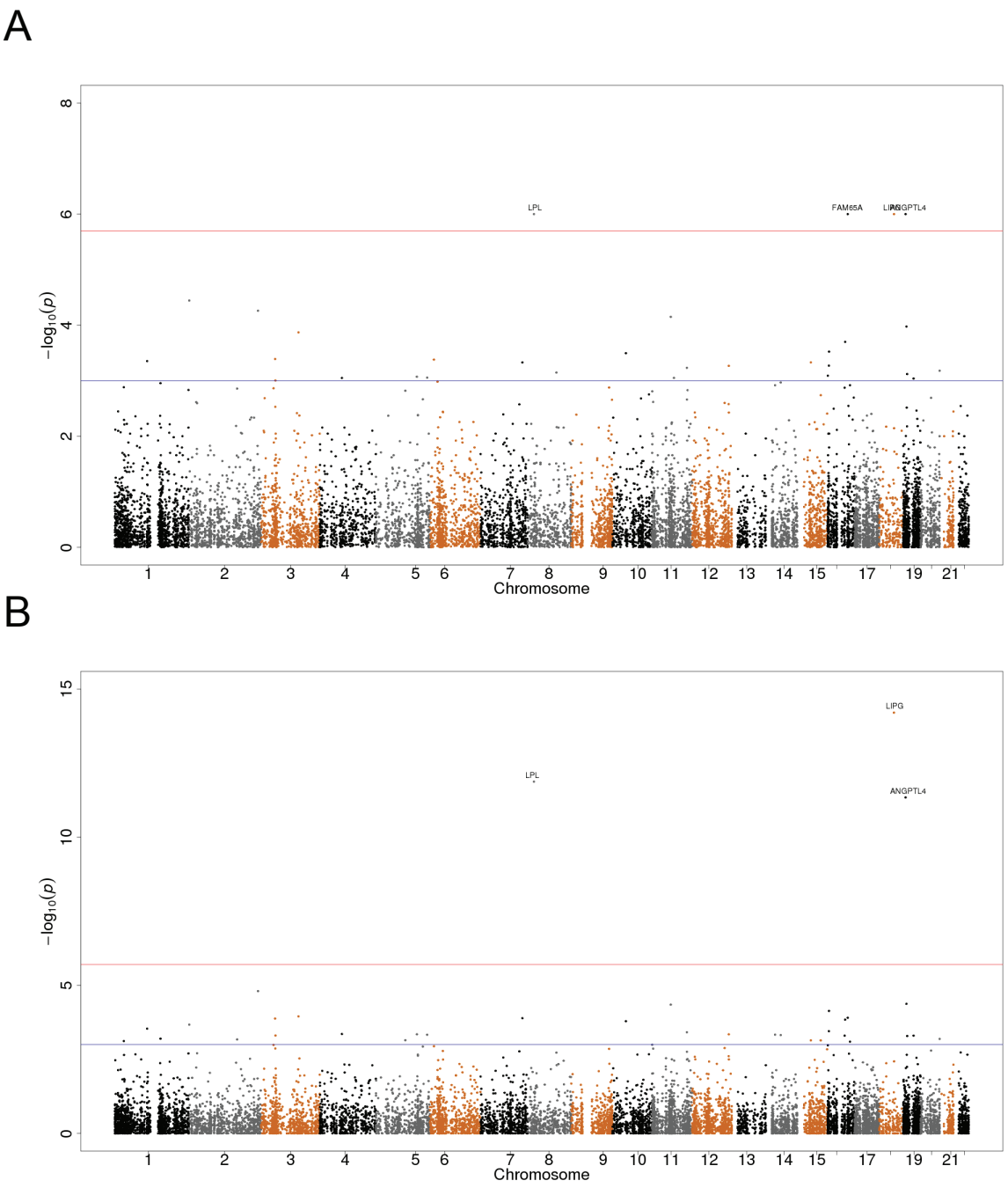

Figure 11: Manhattan Plot Comparison between LaSPU test and LSSU test on the association between HDL-C and Rare Variants in the ARIC study. A. result by LaSPU; B. result by LSSU .

3.4 Discussion

We have proposed a powerful data-adaptive set-based association test for longitudinal phenotypes that can accommodate common or rare variants. To our best knowledge, there was no similar test proposed before, except a recently proposed, longitudinal genetic random field model (LGRF) by He et al.[He et al., 2015] Their method is more alike to Longitudinal SKAT (LSKAT), and thus not data adaptive. We discussed more about this method below. On the other hand, a number of GWASs[Aulchenko et al., 2009, Ionita-Laza et al., 2007, Kamatani et al., 2010, Kathiresan et al., 2007, Sabatti et al., 2008] already included longitudinal measurement of phenotypes but researchers could not fully utilize the extra information contained in the repeated measurements. Thus, our proposed test helps fill the gap by providing a set-based test, for either CVs or RVs, which adaptively maintains a high power for various unknown genetic architecture in a longitudinal data analysis setting. Our proposed test introduces the parameter γ to achieve the data adaptivity. For example, if the SNPs in a region have similar effect magnitudes and directions, SPU(1) will be preferred; otherwise, SPU(2) will be preferred. If there are a lot of null SNPs in a region, then SPU with a higher γ or eventually SPU(∞) will be preferred. Variants of LaSPU provide us with even more data adaptivity. For example, the LaSPU.omni test combines the strengths from unweighted version of LaSPU (as preferred in our simulation study for CVs), weighted version of LaSPU (as preferred in our simulation study for RVs) and score test (as preferred in some cases where LD structure among SNPs brings more information to boost the power). The data adaptivity is missing from many existing longitudinal association tests such as LSSU, LSKAT, and LGRF tests. As supported by our numerical examples and application to the ARIC study, the proposed test is almost always powerful as the most powerful non-adaptive test with varying identity in many situations and is a competitive alternative complementary to

existing methods such as LSSU or LSKAT. By comparing with baseline measurement analysis in both simulations and the real data application, we were able to confirm that fully utilizing the information embedded in the repeated measurements could effectively boost the power.

Our LaSPU family includes several members, i.e., the original LaSPU, LaSPU weighted, LaSPU(w).Score test, and LaSPU omnibus test. The LaSPU weighted test modified the original LaSPU by weighting the contribution from each SNP by its corresponding diagonal element in the variance-covariance matrix of the genotypes. LaSPU(w).Score selects the most powerful test from a class of SPU or SPUw tests and the multivariate score test. LaSPU omnibus test select the most powerful test from SPU, SPUw and score tests. If a user prefers a single test, we recommend the original LaSPU or LaSPU omnibus test. The reason of recommending the original LaSPU test is that it is most powerful in most cases, while in rare cases, the weighted version or score test could be more powerful. For example, when the number of SNPs in a testing region is moderate and the linkage disequilibrium plays an important role in delivering the association information, the score test may be more powerful; when signal SNPs are dominated by regional SNPs with a larger variance and thus hidden from the association test, LaSPU weighted may be more powerful than the original version. We also recommend the LaSPU omnibus test. The reason is that the omnibus test will select the most powerful test from SPU, SPUw and the score tests with little attrition in power. Since the power loss is limited and it is the most comprehensive test to cover all cases, we still recommend the LaSPU omnibus test after the original LaSPU test. Furthermore, since all these tests share the pre-computed numerators (score vector), denominators (diagonal or full part of the variance-covariance matrix) and the simulated or permuted null distribution of score vectors, these tests can be executed at the same time efficiently, providing users the ability to collect all testing results after a single run of the program and need not worry about extra computation

time if they decide to change their choice of the test in advance.

Our proposed test is general and applicable to CVs or RVs. As shown for the SSU test in Basu and Pan,[Basu and Pan, 2011] via suitable weighting on SNPs, it can be easily modified to analyze both CVs and RVs. In addition, we can also introduce some weights on SNPs incorporating biological knowledge, such as the likelihood of SNPs being causal. See Wu et al.[Wu et al., 2011] for more discussion on specifying weights. With regard to longitudinal data analysis in GWAS settings, our proposed test treats the SNP effects as time averaged group effects, and is able to identify the causal effects with increased power. The increased power comes from treating each person as his/her own control. For most outcomes Y, there is considerable variability across individuals due to the influence of unmeasured characteristics such as genetic make-up, environmental exposures, personal behaviors/habits, and so forth. While these factors tend to persist over time for the same individual, their influences are canceled in the estimation of the time averaged SNP effect, and thus lead to more accurate estimate (with smaller variance) and a larger power. Our proposed test can be also easily modified to include the SNP by time interaction into the test, i.e., joint test the time averaged SNP effect and SNP-by-time interaction effect. A recent work by He et al.[He et al., 2015] proposed a longitudinal genetic random field model (LGRF), which utilized the concept of genetics similarity and within-subject similarity. LGRF views phenotypic values as a random field on a genetic space where the vector of genotype sequences determines the location in the space. The phenotypes of different subjects with closer locations in genetic space may be more similar if there is any correlation between genotype and phenotype. Similarly, for repeated measurements, one may expect that phenotypes from the same subject may be more similar and possibly the similarity positively correlates with time distance between measurements. Therefore, LGRF also regresses the phenotype on other phenotypes measured within the same subject, and uses a within-subject similarity matrix to weight the contribution from each

other phenotype within the same subject. Eventually, their test allows to optionally test the joint effect including both the main effect of genotypes and the interaction effect between genotypes and time. To achieve this purpose, LGRF again includes a third term, the contribution from all other phenotypes, weighted by the gene-time interaction similarity term, in the equation. Like our procedure within the GEE framework, they constructed estimating equations and derived generalized score statistics to obtain the p Value of the (joint) test. In our case, we only need to expand the X_i matrix as in $H_i = (Z_i; X_i)$, representing the genotype score, to include the SNP-by-time values as extra columns (thus we augument the X_i from $k \times p$ to $k \times p(1 + k)$). With all other settings the same, the LaSPU test will effectively implement the joint test by testing $H_o : \beta = (\beta_1, \beta_2, \ldots, \beta_{p(1+k)})' = 0$. These extensions are very interesting and warrant future investigation.

In the illustration of our proposed method in the ARIC study, we identified all previously reported HDL-C associated genes with rare variants except for *PAFAH1B2* with a p value of 0.0022. Thanks to the longitudinal design in ARIC, the extra information contained in the repeated measurements boosted the statistical power to achieve similar gene discoveries with a much smaller sample size (11,478 compared to 42,208). Additionally, we found a novel gene associated with HDL-C, $FAM65A$, as confirmed by multiple longitudinal tests. This gene warrants follow up and replication in future studies.

4 Journal Article 2

Title of Journal Article

Pathway-based data-adaptive association tests for longitudinal phenotypes

Name of Journal Proposed for Article Submission

American Journal of Human Genetics

Abstract

Genome-wide association studies (GWASs) have been successful in the last decade for identifying many genetic loci associated with many human complex diseases. However, only a small portion of heritability for each complex trait has been explained by genetic factors, such as SNPs. Possible reasons include genetic heterogeneity (i.e., multiple causal genetic variants) and small effect sizes of causal variants, for which pathway analysis has been proposed to relieve such problems. On the other hand, GWASs with repeated measurements collected for complex traits may bring extra power to identify associated genetic variants than their cross-sectional counterparts. Due to the complexity in the genetic architecture and the association patterns in local signals, a single association testing method may not perform well in many scenarios. Thus, a data-adaptive association testing method is preferred as it adaptively selects a test from a class of tests to maintain a high statistical power. In this article, we developed a pathway-based association test that is adaptive at both the gene and SNP levels for longitudinal traits. The pathway contains a group of functionally related genes and each gene contains multiple SNPs. We developed different strategies to correctly handle common variants or rare variants in the association testing. We showed the methods can incorporate biological knowledge on SNPs and genes to boost the statistical power. We showed the methods are robust to the misspecification of within-subject correlation, which is a desired feature for longitudinal data analysis. We also showed the methods can handle subject with missing measurements effectively. We implemented extensive simulation studies to demonstrate our methods' promising performance as compared to several existing longitudinal SNP-set-based association tests. Furthermore, we applied the proposed methods to the Exome Chip data from the Atherosclerosis Risk in Communities (ARIC) study. Through real data application, we were able to identify three significant KEGG pathways

associated with high-density lipoprotein cholesterol (HDL-C). Finally, we have built an efficient software package to implement the proposed methods and make it scalable for large-scale genome-wide association studies with longitudinal traits.

4.1 Introduction

Over the last decade, hundreds of genome-wide association studies (GWASs) for complex and common human diseases have been successful completed (see, for example, A Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/26525384). GWAS is a powerful approach to examine the genetic components for human diseases. Yet to date, the genetic variants discovered by GWAS, mainly based on the traditional univariate analyses of individual single-nucleotide polymorphisms (SNPs), only explain a small portion of the estimated heritability for each human complex trait. One possible reason is, due to the polygenic basis of disease susceptibility (i.e., multiple causal genetic variants) and small effect sizes of the genetic variants, the traditional single-SNP-based analysis is not suitable to detect many variants with smaller effects. It is likely that alternative approaches that focus on the combined effects of many loci, each making a small contribution to overall disease susceptibility, might have the power to extract extra information and reveal more insights into the missing genetic heritability. One such alternative is the gene set or pathway based analysis. These two terms are often interchangeably used while others have pointed out a difference, the gene set does not model specific relationship among genes as the pathway does.[Fridley and Biernacka, 2011] Our proposed new method below is able to work on any set of genes, thus attributing itself to the gene set based analysis. However, we will omit this difference and use the term, pathway analysis, consistently hereinafter. A genetic pathway contains a set of biologically meaningful combinations of genes, such as annotated in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database,[Ogata et al., 1999] the gene ontology (GO) database,[Ashburner et al., 2000] the Reactome database,[Joshi-Tope et al., 2005] the BioCyc database,[Caspi et al., 2008] the BioCarta database[Nishimura, 2001] and the pathway interaction database (PID).[Schaefer et al., 2009] Genetic variants located across

multiple genes may contribute to the same or similar phenotypes. By analyzing functionally related genes together with the trait, we might have improved power and are more likely to identify novel genes containing variants related to the trait, which otherwise are hidden due to insufficient power in the single-SNP-based or single-gene-based analysis.[for Blood Pressure Genome-Wide Association Studies et al., 2011, Hirschhorn, 2009, Zhong et al., 2010, Wang et al., 2010, Holmans, 2009] Another advantage of pathway-based test is its ability to aggregate variants information. One convincing evidence is from the Cancer Genome Atlas (TCGA)[Network et al., 2011] in tumor sequencing studies. While only few oncogenes (for example, *TP53* [MIM: 191170] and *EGFR* [MIM: 131550]) harbor many mutations, most others harbor few mutations in a tumor-dependent manner. For instance, a tumor might contain mutation in *PTEN* [MIM: 601728], not in *NFKB1* [MIM: 164011], whereas another tumor contains mutation in *NFKB1*, not in *PTEN*. Individually, each of the genes in a related pathway has only a low mutation frequency, whereas collectively, they have a much higher aggregated mutation frequency. Therefore, for some diseases such as cancer, a pathway analysis by aggregating information across multiple genes in a relevant pathway will boost the statistical power and is thus preferred. For example, among the 315 ovarian cancer [MIM: 167000] tumors studied by TCGA, 45% of them had gnomic alterations including both somatic mutations and DNA copy number variations in the PI3K/RAS signaling pathway. This pathway contains seven genes, including *PTEN*, *PIK3CA* [MIM: 171834], *AKT1* [MIM: 164730], *AKT2* [MIM: 164731], *NF1* [MIM: 613113], *KRAS* [MIM: 190070], and *BRAF* [MIM: 164757]. Each gene contains only low to moderate genomic alterations in 7%, 18%, 3%, 6%, 12%, 11%, and 0.5% of the tumor samples respectively; therefore, aggregating information at the pathway level should bring more power to detect genomic alternations than at the individual gene level.

Previous pathway analyses in GWASs were mostly conducted under a cross-sectional

study design. However, as it has been increasingly recognized that, many genetic studies have included follow-up measurements as well as initial baseline measurement on each subject. For genetic studies on cardiovascular disease risk factors, for example, the Atherosclerosis Risk in Communities (ARIC) study, repeated measurements of traits across study time are available for each individual [Heiss, 1989]. Association tests that fully utilize the information across time points tend to achieve a higher power and identify more disease-associated loci [Furlotte et al., 2012, Xu et al., 2014]. However, current statistical methods for testing the association in longitudinal settings, even for one single nucleotide polymorphism (SNP) at a time, are limited [Fan et al., 2012, Furlotte et al., 2012]. Investigators often take simplified approaches that collapse the repeated measurements into a single value, such as use baseline measurement and use average measurement, to handle the longitudinal data in GWAS. These approaches cannot fully harness the power of the complete information contained in the longitudinal trajectory. On the other hand, researchers can also apply the standard methods for correlated outcome model to longitudinal GWAS settings, namely, random effect models [Laird and Ware, 1982] and marginal models estimated by generalized estimating equations (GEE) [Liang and Zeger, 1986, Zeger et al., 1988]. However, these standard methods are not optimized for testing a large number of SNPs, a common scenario in pathway-based association test.

Recently, the data-adaptive association test emerged and showed their potential in dealing with the complex scenarios in association testings.[Pan et al., 2015b, Pan et al., 2014] Pathway analysis in GWAS usually contains more genetic variants than individual-gene-based or single-SNP-based tests. As higher-dimensional data are involved, more complexity are anticipated in the testing. On the other hand, as usually it is the case, not all the SNPs in any gene or any pathway are related to a trait. It is thus important and challenging to adaptively aggregate information of multiple unknown causal

SNPs while minimizing the noise from non-causal SNPs. Existing approaches of pathway analysis used in cross-sectional GWASs have some limitations in data adaptability as discussed in Pan et.al.[Pan et al., 2015b] For example, Wang et.al's method[Wang et al., 2007] used the minimum p value of the multiple SNPs in a gene to represent the gene's p value in the pathway and thus suffers from power loss when there are multiple weakly associated SNPs inside the gene. Two other methods, GATES-Simes[Gui et al., 2011] and HYST,[Li et al., 2012] combine gene-level p values from GATES,[Li et al., 2011] a gene-based test. GATES uses an extended Simes procedure to combine individual SNP summary statistics and correct for multiple testing. The pathway analysis version modified from GATES, GATES-Simes, further uses an extended Simes procedure to extract the most significant gene-level p value for a pathway. HYST, as based on GATES and the modification from GATES-Simes, uses Fisher's method instead to combine multiple genes' p values for a pathway. Pan et.al. later confirmed that GATES-Simes behaves like the minimum p value method, losing power if there are multiple weakly associated SNPs and/or multiple weakly associated genes. In contrast, HYST, as Fisher's method, loses power if an increasing number of the genes in a pathway are not associated with the trait. The essence of the problem is the non-adaptive nature of these methods at both the SNP and gene levels. Our proposal is based on a highly data-adaptive test called adaptive sum of powered score (aSPU) test originally proposed for analysis of rare variants (RVs).[Pan et al., 2014] The main idea of the aSPU test is, because we do not know in advance the association pattern, e.g., how many, how strong and whether in same direction, the SNPs in a region are associated with a trait, we first construct a class of tests overweighting a sequence of increasingly smaller set of significant SNPs, then select the test with the most significant P value with a proper adjustment for multiple testing. The aSPU test often outperforms other tests for relatively small sets of RVs.[Pan et al., 2014] More recently, aSPUpath test[Pan et al., 2015b] was proposed. It extends

the aSPU test's gene-based testing style to the pathway-based testing style for either common variants (CVs) or RVs. It includes additional adjustment for gene length (i.e., number of SNPs within a gene) and additional adaptivity at the gene level in addition to the SNP level. The adjustment for gene length is necessary because gene shall be treated equally a priori in a pathway; the additional adaptivity at the gene level complements the adaptivity at the SNP level in a pathway analysis, as we do not know a priori how many genes in a pathway and how many SNPs in a gene are associated with the given trait. The aSPUpath test outperforms other tests in many situations.[Pan et al., 2015b] In this paper, we extended the aSPUpath to a test capable of handling traits with repeated measurements, so called longitudinal aSPUpath (LaSPUpath). We showed the methods are robust to the misspecification of within-subject correlation, which is a desired feature for longitudinal data analysis. We also showed the methods can handle subject with missing measurements effectively. For RVs, we provided a special permutation strategy in addition to the simulation based strategy to obtain accurate p values. We executed extensive simulation studies to evaluate the LaSPUpath's performance with comparison to several existing SNP-set-based methods, including uniformly min-Pvalue (UminP) test, GEE-Score test, the sum of squared U-statistics (SSU) and its weighted version SSUw tests,[Pan, 2009] and the Sum test.[Pan, 2009] Although these tests were initially designed for cross-sectional data analysis, it is straightforward to extend them to longitudinal versions. Since previously discussed GATES-Simes and HYST are not applicable to longitudinal trait, we did not compare them with LaSPUpath. Instead, Pan et al. compared aSPUpath with GATES-Simes and HYST under a cross-sectional GWAS setting, and concluded that aSPUpath has advantages over the other two tests in many situations.[Pan et al., 2015b] Lastly, we demonstrated the application of LaSPUpath in the Exome Chip[Grove et al., 2013] data from the ARIC study.[Heiss, 1989] Specifically, we tested the association between high density lipoprotein cholesterol (HDL-C) with re-

peated measurements and the rare variants in gene pathways annotated by the KEGG database.[Ogata et al., 1999] We identified three significant pathways associated with HDL-C.

4.2 Material and Methods

4.2.1 Data and Notation

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. We have p SNPs of interest as a genotype score row vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ with x_{ij} coded as 0,1 or 2 (i.e., additively) for the count of the minor alleles of SNP $j = 1, \ldots, p$. The p SNPs could be drawn from multiple genes in a pathway (see Figure 12).

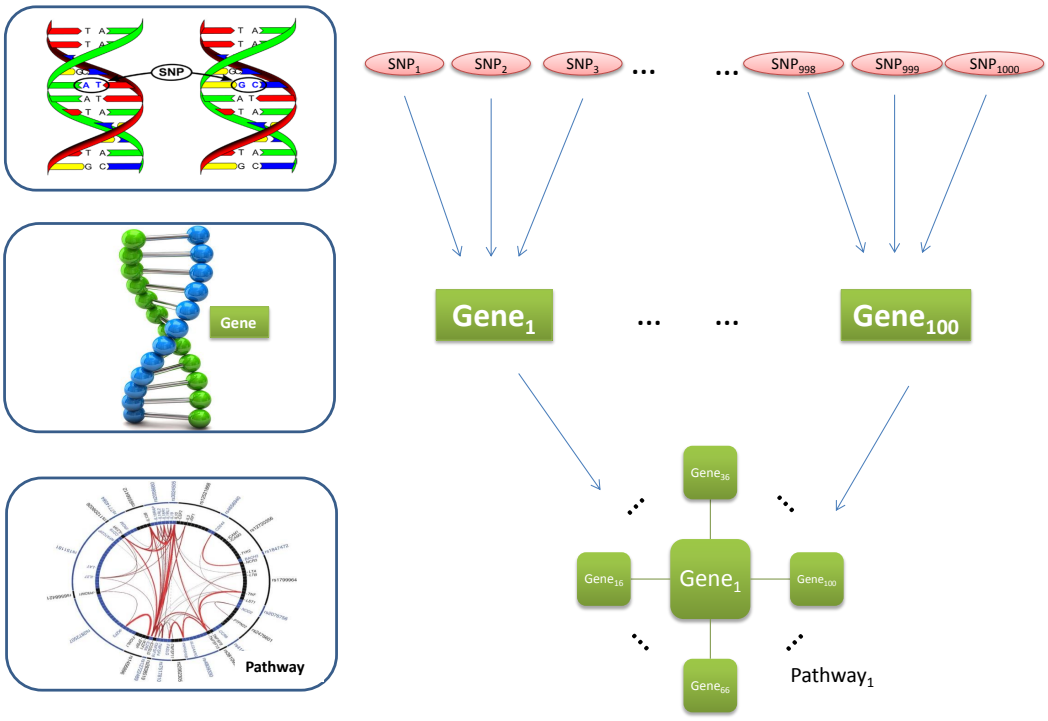

Figure 12: Aggregation of SNPs in a Pathway

We also have $z_i = (z_{i1}, z_{i2}, \ldots, z_{iq})$ is a row vector for q covariates including measurement time. We assume common effect sizes (i.e., time-averaged group effect) of the SNPs and covariates (other than measurement time) on the repeated 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 $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ and $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_{q+1})'$ for X_i and Z_i respectively. The marginal mean of each measurement, $E(y_{im}|x_i, z_i) = \mu_{im}$ where $m = 1, 2, \ldots, k$ 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):

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

with $H_i = (Z_i, X_i), \theta = (\varphi', \beta')'$ and $g(.)$ 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 [Liang and Zeger, 1986]:

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

with

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

and

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

ϕ in V_i is the dispersion parameter in GEE and is usually treated as nuisance parameter. $v(\mu_{im})=\phi\mathrm{Var}(y_{im}|x_i,z_i)$. $R_w(\alpha)$ is a working correlation matrix depending on some unknown parameter α. For convenience, a working independence model with $R_w=I$ is often used. 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:

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

$$\widetilde{\Sigma}=\widehat{\mathrm{Cov}}(U)=\sum_i(Z_i,X_i)'\widehat{\mathrm{var}(Y_i)}(Z_i,X_i)=\sum_i(Z_i,X_i)'(Y_i-\hat{\mu_i})(Y_i-\hat{\mu_i})'(Z_i,X_i)=\begin{pmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix} \tag{20}$$

where $\hat{\mu}_i$ is an estimator of μ_i, $\widetilde{\Sigma}$ is an estimate of the covariance of score (U) vector. $\widetilde{\Sigma}$ is partitioned with the dimensions according to the score vector component $U_{.1}$ and $U_{.2}$

for φ and β respectively.

As we use the identity link for longitudinal continuous trait, i.e. $g(\mu_{im}) = \mu_{im}$ and $v(\mu_{im}) = \phi \times 1 = \phi$, we will have:

$$
\begin{aligned}
U &= \sum_i (Z_i, X_i)' R_w^{-1} (Y_i - \mu_i) \\
\widetilde{\Sigma} &= \sum_i (Z_i, X_i)' R_w^{-1} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' R_w^{-1} (Z_i, X_i)
\end{aligned} \tag{21}
$$

if the assumption of a common covariance matrices across Y_i for i is valid, for example for quantitative continuous traits study [Pan, 2001], 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 [Pan, 2001].

Although repeated measurements are usually for continuous traits, our method can be easily adapted to longitudinal binary trait. In brief, the only difference between dealing with continuous trait and dichotomous trait is the canonical link function. 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

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

4.2.2 Several Existing Set-based Association Tests

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 13.

The classical score test statistic is $T = U_{.2}^{'}\Sigma_{.2}^{-1}U_{.2}$, which, however, will lose power when SNP number p increases with a fixed sample size n. As shown theoretically [Fan, 1996], as the dimension p increases, the power of the score test will diminish, tending to the type I error rate α. The popular UminP test statistic is $T = \underset{j}{max}\frac{U_{.2,j}^2}{\Sigma_{.2,jj}}$ with $\Sigma_{.2,jj}$ is the jth entry on the diagonal of $\Sigma_{.2}$. It might also has low power if we have many small $|\beta_j| \neq 0$. Two alternatives, called the Sum and SSU tests, are

$$T_{Sum} = 1'U/\sqrt{1'V1} = \sum_{j=1}^{p} U_j/\sqrt{1'V1}, \qquad T_{SSU} = U'U = \sum_{j=1}^{p} U_j^2.$$

The Sum test is closely related to other burden tests such like those in [Morgenthaler and Thilly, 2007, Li and Leal, 2008, Madsen and Browning, 2009]. If there is a common association either in direction or strength for causal SNVs with no or few non-associated SNVs, then Sum test and the likes will be most powerful; otherwise, the SSU test and its closely relatives, such as kernel machine regression (KMR or SKAT) [Lee et al., 2012a, Ionita-Laza et al., 2013, Oualkacha et al., 2013, Lee et al., 2012b, Wu et al., 2011] and C-alpha test [Neale et al., 2011], will be more powerful. Nevertheless, as shown in Pan et al.,[Pan et al., 2015a, Pan et al., 2014] a variance-component test is not adaptive and may lose power in the presence of many non-associated SNPs as anticipated in the current context of pathway analysis. Accordingly, a more powerful and adaptive test was proposed as reviewed next.

4.2.3 Review: The Data-Adaptive aSPU Test

Is is straightforward to see both the Sum test statistic and the SSU test statistic are based on score vector but using different weights. A more general form of the score-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 SNPs [Lin and Tang, 2011]. The aSPU test proposed $W_j = U_{.2,j}^{\gamma-1}$ 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 $\gamma = 1$, the SPU(1) test uses $\mathbf{1}$ as weight and sums up the information contained in all the SNPs in the region of interest, equivalent to Sum test or burden test; when $\gamma = 2$, the SPU(2) test uses U as weight to itself and is equivalent to SSU test and other variance-component test such as SKAT; when γ keeps increasing, the SPU(γ) test puts higher weights on the jth SNP with larger $|U_{.2,j}|$, while gradually decreasing the weights of other SNPs with smaller $|U_{.2,j}|$. As the large value of $|U_{.2,j}|$ indicates strong association information stored in SNP j and small value of $|U_{.2,j}|$ indicates weak or none association information stored in SNP j, a higher γ tends to put more and more weights on those informative SNPs. When $\gamma \to \infty$ as an extreme situation, where ∞ is assumed to be an even number, we have

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

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

A simulation based method [Lin, 2005, Seaman and Müller-Myhsok, 2005] could be used to calculate the p value for SPU test. Specifically, suppose T is a short notation of $T_{SPU(\gamma)}$ for a specific γ and $\hat{\Sigma}_{.2}$ is the covariance matrix of the score vector $U_{.2}$ from the original data (Equation 13). 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 $P_{SPU(\gamma)} = \sum_{b=1}^{B} \frac{I(T^{(b)} \geq T^{obs})+1}{B+1}$. We used $B = 1000$ in our simulations for a nominal significance level at 5%.

There is no uniformly most powerful test in set-based association testing; on the other hand, it has been found empirically that the Sum, SSU and UminP tests are preferred over different scenarios. For a given dataset, to adaptively choose the value of γ for the SPU tests, Pan et al. [Pan et al., 2014] proposed an adaptive SPU (aSPU) test that simply combines the results of a series of SPU tests: suppose we have some candidate value of $\gamma \in \Gamma = \{1, 2, \ldots, 8, \infty\}$ as used in our later experiments, and suppose the p vaule of the SPU(γ) test is $P_{SPU(\gamma)}$, then the aSPU test simply takes the minimum p value:

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

Of course, T_{aSPU} is no longer a genuine p value; we recourse to the previous simulation based strategy to obtain its p value. Specifically, since we have the previous simulated $U_{.2}^{(b)}$ and computed $T_{SPU(\gamma)}^{(b)}$ for $b = 1, 2, \ldots, B$ and $\gamma \in \Gamma$, we can directly calculate the p

value for $T^{(b)}_{SPU(\gamma)}$:

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

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:

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

In practice for genome wide scan purpose, we can use a stage-wise aSPU test strategy: we first start with a smaller B, for example, use $B = 1000$ to scan the genomes, then gradually increase B to, for example, 10^6 for a few selected groups of SNPs. For example, we could choose specific genes or windows which passed a pre-determined significance cutoff (for example, p-value $\leq 5/B$) in the previous stage; we then repeat this process until the pre-determined significance level is reached. For example, a p-value of $\leq 10^{-5}$ requires the $B \geq 10^5$. In this stage-wise way, aSPU test could be applied on GWAS data more efficiently.

4.2.4 A Data-Adaptive Pathway-Based longitudinal test: LaSPUpath

Let the short notation $U_{g.}$ equates $U_{.2}$ for genotype score part in Equation 20 , then $U_{g.} = (U_{g,1}, U_{g,1}, \ldots, U_{g,p_g})'$ represents the score vector for gene g with p_g RVs from the GEE fitting. Given a pathway (or a gene set) S with $|S|$ genes, we have the score vector partitioned according to the genes as $U = (U_1', U_2', \ldots, U_{|S|}')'$. Note, all U vectors have already taken into account the longitudinal trait information through GEE fitting. Then, the gene-specific SPU statistic is as follows:

$$T_{SPU(\gamma, w_g; g)} \propto ||Ug.||_\gamma = \left(\frac{\sum_{j=1}^{p_g} w_{g,j} |U_{g,j}|^\gamma}{p_g} \right)^{\frac{1}{\gamma}} \tag{22}$$

Then accordingly, the pathway-based SPU statistic is

$$T_{Path-SPU(\gamma,\gamma 2,w,w_G;S)} = \sum_{g\in S}(w_{G,g}T_{SPU(\gamma,w_g;g)})^{\gamma 2} \tag{23}$$

Now we have two scalars $\gamma > 0$ and $\gamma_2 > 0$, gene-specific weights for SNPs $w = (w_1', \ldots, w_{|S|}')'$ and $w_g = (w_{g1}, \ldots, w_{g,p_g})'$, and gene-specific weights for genes $w_G = (w_{G,1}, \ldots, w_{G,|S|})'$ are pre-specified. w_g is used to incorporate prior information on SNPs, e.g., to up-weight SNPs associated with gene expression, whereas w_G can be chosen according to gene functional annotations or gene expression data to represent prior likelihoods of their being functional and thus associated with the trait; without prior knowledge or data, or for simplicity, we can just use $w_g = 1$ and $w_G = 1$, which are to be used by default in this paper unless otherwise specified.

Note the $T_{SPU(\gamma,w_g;g)}$ is now standardized by the gene-specific number of RVs, p_g, so that large genes will not dominate the test (since the genes in a pathway are the analysis units and are thus treated equally a priori if no weighting schema is used); for a given gene g, $T_{SPU(\gamma,w_g;g)}$ is equivalent to $T_{SPU(\gamma)}$ as in aSPU method by large. The intuition of $\gamma 2$ is like that of γ: If we treat the pathway as the gene and the gene as the RVs. A larger $\gamma 2$ (γ) put more weights on heavily associated genes (RVs), when gradually ignoring the less associated genes (RVs) in a pathway (gene). An extreme case is $\gamma 2 = \infty$, it indicates the pathway-based analysis actually used only one gene - the most heavily associated gene with the trait. Since the goal of pathway-based analysis is to take advantage of multiple "co-working" genes, and aggregate more RVs, it is less meaningful to consider the use of a $\gamma 2 = \infty$. Instead, we propose to use $\gamma 2 \in \Gamma 2 = \{1, 2, 4, 8\}$. The reason is that, at the pathway level, the statistic $T_{SPU(\gamma;g)}$ is always a positive number, as compared to $U_{.2,j}$ from the GEE model fitting for SNPs which can have different signs (SNP effect directions). Thus, deliberately assigning both odd and even number of $\gamma 2$ becomes

unnecessary, and we can actually use most representative $\gamma 2$ values and expect them to have most distinct effects from each other. Fewer $\gamma 2$ candidates will also expedite the computation. To these purposes, $\gamma 2 \in \Gamma 2 = \{1, 2, 4, 8\}$ will cover Sum-like test, SSU-like test, and two more tests preferring the sparse-causal-gene situation (for example, only 2 or 3 genes are associated with the longitudinal trait in a pathway with 20 genes). For γ at the SNP level, we stick to $\gamma \in \Gamma = \{1, 2, \ldots, \infty\}$ as used in Pan et.al.[Pan et al., 2014]

For any given $(\gamma, \gamma 2)$, we recourse to the simulation-based strategy[Lin, 2005, Seaman and Müller-Myhsok, 2005], as used in aSPU test, to calculate the p-value $P_{Path-SPU(\gamma,\gamma 2,w,w_G;S)}$ from $T_{Path-SPU(\gamma,\gamma 2,w,w_G;S)}$. Its power depends on the choice of $(\gamma, \gamma 2)$. Then we will have the pathway-based aSPU test statistic defined as

$$T_{Path-aSPU(S)} = min_{\gamma,\gamma 2} P_{Path-SPU(\gamma,\gamma 2,w,w_G;S)} \tag{24}$$

aiming to select a most powerful test from multiple tests coming from different combinations of $(\gamma, \gamma 2)$. Lastly, as similar to that for the aSPU test, we propose using a single layer of the simulation based strategy to calculate the p value $P_{Path-aSPU(S)}$.

For the possible situation where multiple genes in a pathway contain quite different proportions of causal SNPs, we might use a more general pathway-based test with a gene-specific γ_g for each gene g in $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{|S|})'$. We can thus modify the tests as

$$T_{Path-SPU2(\boldsymbol{\gamma},\gamma 2,w,w_G;S)} = \sum_{g \in S} (w_{G,g} T_{SPU(\gamma_g,w_g;g)})^{\gamma 2} \tag{25}$$

$$T_{Path-aSPU2(S)} = \min_{\boldsymbol{\gamma},\gamma 2} P_{Path-SPU2(\boldsymbol{\gamma},\gamma 2,w,w_G;S)} \tag{26}$$

The corresponding aSPUpath2 test will be computationally more demanding, since $\gamma_g in \boldsymbol{\gamma}$ now can be different for each gene in the pathway and to find a best $\boldsymbol{\gamma}$ and $\gamma 2$ combo need

more grid searching of all possible combinations. Besides, a more flexible combination will also introduce more variability to the results and thus might lead to loss of power. This needs to be studied further.

4.2.5 Extension to Rare Variant Analysis

While minor allele frequencies (MAF) of RVs are usually low, for example, between 0.001 to 0.01, the property of asymptotically normal distribution of either the regression coefficient or score vector may not hold. The simulation based strategy in LaSPUpath (mainly for CVs) may not be sufficient for RV analysis. It will be nice if we can provide an alternative to estimate the p value more precisely for RV analysis in LaSPUpath method. Thereby, we propose a permutation based method to generates the empirical null distribution of $U_{.2}^{(b)}$ instead of the previous simulation based method. The permutation strategy will not break the time-dependent measurement order or the association between the longitudinal trait and possible covariates, but only permute the genotype codes across individuals. It will be also robust to missing data as this situation is usually anticipated in the longitudinal data analysis.

The permutation strategy can be implemented as follows:

1. identify the max k across all n subjects, which is the number of longitudinal measurements. For example, $k = 4$ to illustrate.

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 $Y_i = (y_{i,1}, \text{NA}, \text{NA}, y_{i,4})'$). 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., $k \times (q+p+1)$, for covariates and SNPs. Now we should have $\begin{pmatrix} Y_i & H_i \end{pmatrix}$ as an augmented matrix of dimension $k \times (q+p+2)$ 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 $nk \times (q+p+2)$.

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

After we obtain $U_{.2}^{*(b)}$ pool from permutation strategy, we could have an empirical null distribution as we did for simulation based strategy. Thus, the left work of the LaSPU-path test becomes the same for both strategies. In the real data application involving RVs, we recommend the combined strategy that users could first run simulation based LaSPUpath test to obtain the putative significant genetic loci, and then apply the more time-consuming permutation based LaSPUpath test on those selected loci as validation.

4.2.6 Simulation Set-ups

We conducted extensive simulation studies to evaluate and compare the performance of the LaSPUpath test with several alternative methods capable of dealing with longitudinal trait, such like SSU, Score, Sum, and UminP test. Our general pathway structure set-ups were similar to those in Pan et.al.[Pan et al., 2015b] and Chen et.al.[Chen et al., 2010a] This simulation strategy allows flexible manipulation of the pathway structure, that is the SNPs inside each gene might or might not be correlated whereas the SNPs from different genes were always independent, and the causal SNPs might or might not be included in the data. Specifically, as the default scenario, we considered one pathway containing 20 genes; each gene g contained p_g RVs with p_g randomly draw from a uniform distribution $U(1, 20)$. In set-up A, there was no causal genes or causal SNPs in the pathway; in set-up B, 5 of the 20 genes will be randomly selected to be causal, with each causal gene containing $U(1, 3)$ causal RVs. We simulated genotypes within a gene following Wang and Elston.[Wang and Elston, 2007] We simulated the longitudinal trait using AR(1) correlation structure following [Song et al., 2013]. We took into account the SNP main effect and time course main effect as fixed effects on the longitudinal trait without consideration of SNP by time interaction. For simplicity, we did not include other covariate effects (such as demographic variables) in the simulation studies, though they can be simply added. We referred to previous studies [Pan et al., 2014, Basu and Pan, 2011, Pan and Shen, 2011, Han and Pan, 2010, Pan, 2009] and the ARIC data to set up the simulation parameters, for example, ρ_y across longitudinal measurements and ρ_x across SNPs as used in AR(1) correlation structure model, so that our simulation can better approximate real data analysis in the later.

Methods for simulation of genotype data

First, a latent vector $G_i = (G_{i1}, \ldots, G_{ip})'$ was drawn from a multivariate normal distri-

bution $MVN(0, R)$, 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 $f \neq g$. The default ρ was set between 0.5 and 0.9 to mimic the real data. The number of SNPs inside each gene, p_g, was randomly chosen between 1 and 20. Second, the latent vector 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 jth SNP; the MAFs were randomly drawn from a uniform distribution between 0.05 and 0.4 for CVs or between 0.001 and 0.01 for RVs. Third, we combined two independent haplotypes to obtain the genotype $X_i = (X_{i1}, \ldots, X_{ip})'$ for subject i. In set-up A, there was no causal genes, thus no causal SNPs in the pathway ($\beta_j = 0$); in set-up B, 5 of the 20 genes will be randomly chosen to be causal, with each causal gene containing $U(1, 3)$ causal RVs ($\beta_j \neq 0$). SNPs across different genes are always independent.

Methods for simulation of longitudinal trait

We first obtained the estimates of the parameters used in this simulation from a preliminary analysis of ARIC dataset. For example, we estimate $\rho = 0.7$ for repeated measurements of HDL-C trait within the same subject. Secondly, we set-up the mixed effect model to achieve the AR(1) correlation structure as:

$$y_{im} = \mu_i + b_i + \underbrace{\rho e_{i,m-1} + s_{i,m}}_{e_{i,m}}, \tag{27}$$

with $m = 1, \ldots, k.$ indexes the longitudinal measurements within subject i; $\mu_i = Z_i\varphi + X_i\beta = H_i\theta$, where time as a covariate and time effect as a parameter are included in Z_i and φ respectively; 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. $e_{i,m}$ is the total residual, which can be

divided into two parts: the first part depends on $e_{i,m-1}$ and the second part is an independent term. We assume they follow:

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

It is straightforward to see the sum $\rho e_{i,m-1} + s_{i,m}$'s variance is equal to the variance of $e_{i,m}$ by algebraically summing up two parts. Under this assumption, the variance-covariance matrix across longitudinal measurements becomes (for $k = 4$ as the number of longitudinal measurements):

$$\Sigma_{4\times 4} = Var \begin{pmatrix} b_i + e_{i1} \\ b_i + \rho e_{i1} + s_{i2} \\ b_i + \rho e_{i2} + s_{i3} \\ b_i + \rho e_{i3} + s_{i4} \end{pmatrix} = \sigma_b^2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \tag{28}$$

On the rightmost end of the equation, the first part defines the within-subject variances, and the second part allows the measurements with a k-interval lag to have a correlation coefficient of ρ^k. This is closer to reality in some cases for longitudinal data.

Methods for tuning simulated genetic effect

As noted in association tests, different SNPs contribute to the trait with different effect sizes. However, the SNP effect magnitude tuning in the simulation study is not trivial. Instead of assigning a β_d coefficient to a SNP with an arbitrary numerical value, for example, 0.1 or 10000, there is a way to use genetic heritability to control the association magnitude of the jth SNP [Lynch et al., 1998]. We first introduce the formula of the

variance of the phenotype :

$$Var(y_{im}) = Var(X_{ij})\beta_j^2 + \sigma_{oth}^2 = 2f(1-f)\beta_j^2 + \sigma_{oth}^2 \tag{29}$$

where the Hard-Weinberg equilibrium (HWE) is assumed to hold. f is the MAF of the SNP; σ_{oth}^2 is the residual variance after removing the effect of jth SNP. Obviously we can see σ_b^2 and σ_e^2 are contained in σ_{oth}^2 (see equation (27)). If other SNPs' effects 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 (29):

$$h^2 = \frac{Var(A)}{Var(P)} \tag{30}$$

This is the classical formula of narrow-sense heritability, with $Var(A)$ represents the variance due to the additive effects of the alleles, and $Var(P)$ represents the total variance in the phenotype. In our situation for jth SNP, this can be expanded to:

$$h_j^2 = \frac{Var_j(A)}{Var(P)} = \frac{Var(X_{ij})\beta_j^2}{Var(y_{im})} = \frac{Var(y_{im}) - \sigma_{oth}^2}{Var(y_{im})} \approx \frac{Var(y_{im}) - \sigma_b^2 - \sigma_e^2}{Var(y_{im})} \tag{31}$$

By systematically solving the equations (29) and (31), we can easily calculate the β_j for jth SNP once we have pre-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 be small for complex disease, and we simulated the dataset with a range of moderate to small h_j^2 values so that we obtained different β_j for further testing.

Other Settings in Simulation

We fixed the test significance level at $\alpha = 0.05$. We ran 1000 replicates for all participating tests, including LaSPUpath, GEE-Score, UminP, SSU, SSU weighted (SSUw) and Sum tests. We used $B = 1000$ for either simulation or permutation strategy whenever encountered. To investigate the type I error and the power performances under different

effect sizes in a longitudinal settings, we simulated in total five repeated measurements with a h_j^2 of either 0 or 0.001, 0.0025, 0.005, 0.0075 and 0.010. causal SNPs are excluded from the final test for CVs to mimic the genotyping array while still kept in the final test for RVs to mimic the sequencing platform. We set the MAF of SNPs between 0.05 and 0.4 for CVs and between 0.001 and 0.01 for RVs. In the set-up A, we let $h_j^2 = 0$ to set $\beta_j = 0$ (i.e., no causal SNPs) when we were testing the type I error of those methods; in the setting B, we set causal SNPs' h_j^2 to different values aforementioned to derive their $\beta_j \neq 0$ when we were testing the empirical power of those methods.

4.3 Results

4.3.1 Simulation Results for CVs

Type I Error

In the set-up A, as shown in Table 6, result shows that all the tests could maintain their Type I Error rate well at the nominal level ($\alpha = 0.05$) except for GEE-Score test, which is over conservative. This is reported before, as when number of variables (variants here) increases, the score test will have over-conservative Type I Error rate and may lose power.

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

Table 6: Empirical Type I Error Table in the simulation set-up A.

Comparison of the LaSPUpath Test with Other Tests

In the setting B, as shown in Figure 13, LaSPUpath method performed the best with a relatively large advance in statistical power as compared to the second place, the UminP method. While UminP method is simple, it performed quite well in this scenario, where we have multiple causal genes in a pathway and multiple causal SNPs in a causal gene. After UminP, SSU and SSUw performed similarly, which shows in this scenario, the

weighting by SNPs' variances does not make a difference. Then, there followed the Sum test, which seems not to be powerful at all in the pathway analysis. That is possibly due to an increased number of null SNPs. The bottom one is the GEE-Score test, which had an over-conservative type I error before and lost a lot of power as a result.

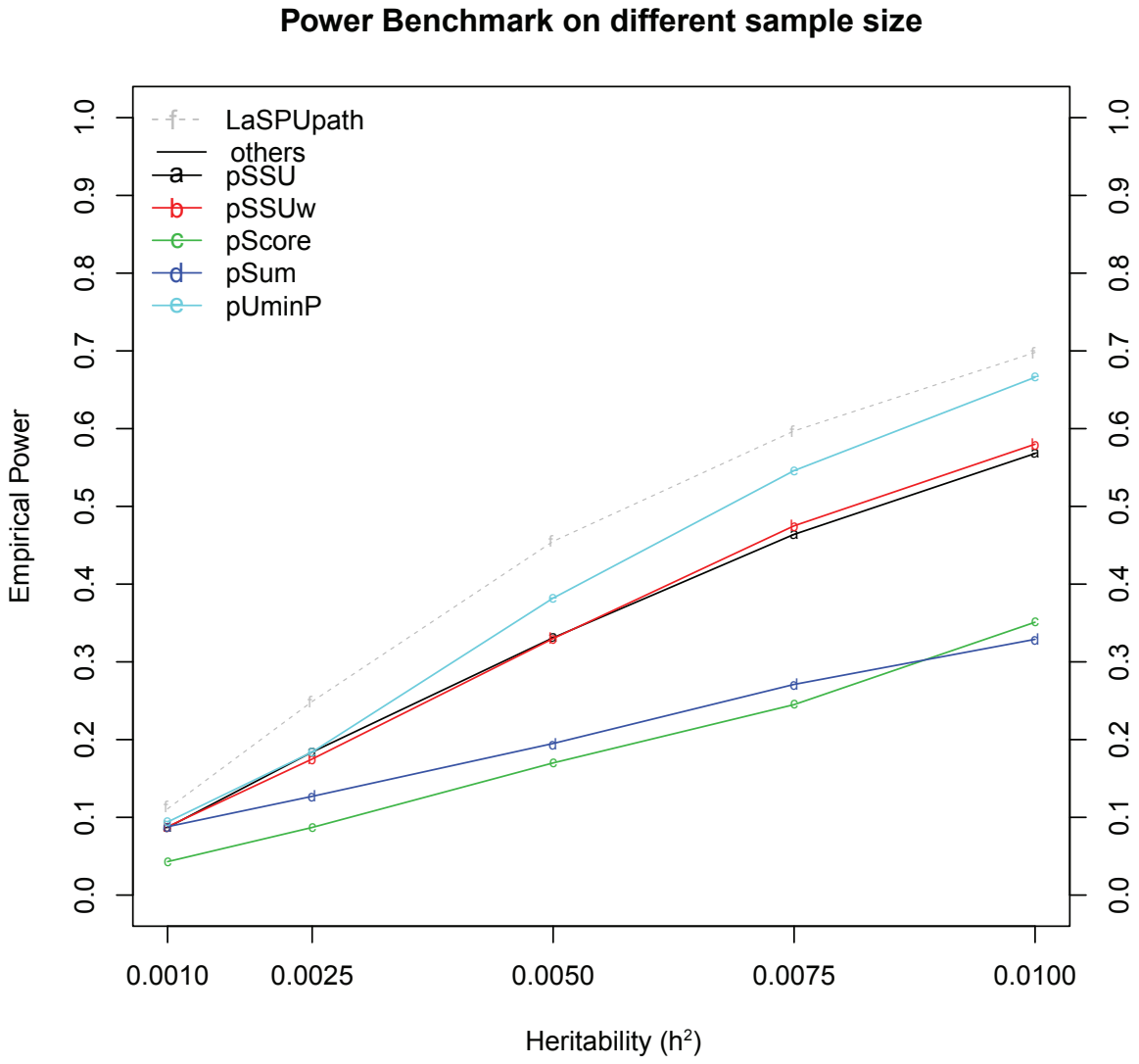

Figure 13: Empirical power benchmark under different heritability (h^2) in the simulation set-up B.

4.3.2 Application to the ARIC study

We illustrated the proposed LaSPUpath method using the data from the Atherosclerosis Risk in Communities (ARIC) Study [Heiss, 1989]. ARIC study is a prospective epidemiological study designed to investigate the causes of atherosclerosis and its clinical outcomes, the variation in cardiovascular risk factors, medical care, and disease by race, gender, location, and time. We used the data from the ARIC study cohort component, which 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 examination (baseline) occurring in 1987-89, the second in 1990-92, the third in 1993-95, and the fourth exam was in 1996-98. The fifth examination was farther apart from the previous screens and was finished during 2011-2013. A detailed description of the ARIC study design and methods was published elsewhere [Investigators et al., 1989]. We exclusively used the Caucasian samples ($n = 11478$) in the ARIC cohort dataset with four available measurements. We used HDL-C as the phenotype of interest. For the genotype data, we used rare variants defined by MAF less than 5% genotyped on HumanExome BeadChip v.1.0 (Illumina).[Grove et al., 2013] We applied standard quality control (QC) procedures on individual samples, which include checking concordance to GWAS data and excluding those individuals missing $> 5\%$ genotypes, population clustering outliers, individuals with high inbreeding coefficients or heterozygote rates, individuals with gender mismatches, and individuals with an unexpectedly high proportion of identity-by-descent sharing, with consideration for family studies, based on high-quality variants. Following the literature on association testing of rare variants, we assumed the the additive genetic model and coded a variant as 0, 1, or 2, i.e., the number of the minor alleles. With respect to covariates, we included the top two principal components eigenvectors (PCs) produced by EIGENSTRAT [Price et al.,

2006] in the generalized linear model to adjust for the potential population stratification within the ARIC Caucasian subjects. Additionally, we included a subject's demographic information such as age, age^2, gender and BMI. We also included the time effect as a covariate. We retrieved a total of 214 human biological pathways from the KEGG database.[Ogata et al., 1999] Because a too small pathway can give results not too different from a gene-based analysis, whereas, a too large pathway usually has non-specific annotated function, many authors restricted their analyses to pathways of certain sizes. For example, Chen et.al.[Chen et al., 2010a] and Wang et.al.[Wang et al., 2010] considered pathways with at least 10 genes, while Gui et.al.[Gui et al., 2011] included only pathways containing between 10 and 300 genes. Following the previous authors, to facilitate the interpretation of the results, we excluded too-small (< 10 genes) and too-big (> 500 genes) pathways, which resulted in 197 pathways left. We employed a stage-wise permutation strategy for the LaSPUpath test: we first performed 1,000 permutations for all gene regions and then increased to 1,000,000 permutations for those gene regions with p values < 0.01 in the first stage. We set the significance threshold at $0.05/197 = 0.00025$ to control the family-wise error rate (FWER) at 0.05 based on the Bonferroni correction for 197 pathways. In the real data application involving RVs, we recommend a stage-wise strategy that users could first run the simulation based LaSPUpath test to obtain a small number of putative significant pathways at a liberal threshold, e.g., $p < 0.01$, and then apply the more time-consuming permutation based LaSPUpath test on those putative pathways as validation (with the minimal number of permutations necessary, i.e., 4,000 here, to achieve a significance level at 0.00025.)

Table 7 shows three significant pathways identified by LaSPUpath. The first pathway is hsa03320 with a p value of $1e-06$ (i.e., none of the permuted test statistics exceeded the observed one based on 1,000,000 permutations). The hsa03320, also known as the PPAR signaling pathway, was reported to play a role in the clearance of circulating or

cellular lipids via the regulation of gene expression involved in lipid metabolism in liver and skeletal muscle. PPARbeta/delta is involved in lipid oxidation and cell proliferation, while PPARgamma promotes adipocyte differentiation to enhance blood glucose uptake. The pathway contains a few lipid metabolism closely related key genes including *APOA1,2 and 5*, *APOC3*, *FADS2*, *LPL*, *ANGPTL4* and *FABP1-7*. Among these genes, *APOC3*, *LPL* and *ANGPTL4* are known genetic loci associated with HDL-C through previous rare variant analysis.[Peloso et al., 2014] Hsa03320 was previously found mainly associated with lipid-related diseases like Hyperlipoproteinemia, type I; Hyperlipoproteinemia, type V; Familial partial lipodystrophy; Hypoalphalipoproteinemia and Hyperalphalipoproteinemia. The second pathway, Hsa00561 (p value: $1e-06$), also known as Glycerolipid metabolism pathway, belongs to lipid metabolism class. It contains a few lipid metabolism closely related key genes including *LIPG*, *LIPC*, *LPL*, *PNLIP* and *PNLIPRP1-3*. It was previously reported to be associated with Hyperlipoproteinemia, type I. The last pathway, hsa05010, also known as Alzheimer's disease (AD) pathway, is closely related to the autosomal dominant or familial early onset AD (FAD). We noticed that this relatively large pathway (of 152 genes) contains several lipase, phospholipase and lipoprotein related genes, including *LPL*, *APOE*, *LRP1* and *PLCB1-4*. There are a number of studies reporting a reduced level of high-density lipoprotein cholesterol is highly correlated with the severity of AD.[Fernandes et al., 1999, Merched et al., 2000, Michikawa, 2003, Puglielli et al., 2003] Furthermore, atherosclerosis, intimately related to high blood cholesterol, has been shown as well to correlate with an increased risk of AD, with higher levels of risk being associated with advanced atherosclerosis.[Hofman et al., 1997, Casserly and Topol, 2004] Thus, the identified association between hsa05010 and HDL-C has validated previous findings.

In the end, we concluded that all the three identified pathways by LaSPUpath are closely related to the regulation of HDL-C level. Our proposed method is competitive and

reliable.

Table 7: 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

4.4 Discussion

We have proposed a powerful data-adaptive test for pathway analysis of longitudinal trait and genetic SNP data. There was no such kind of a test before ours, and meanwhile, a lot of GWAS [Aulchenko et al., 2009, Ionita-Laza et al., 2007, Kamatani et al., 2010, Kathiresan et al., 2007, Sabatti et al., 2008] already included longitudinal measurements of traits but researchers could not fully utilize the extra information contained in the repeated measurements. Thus, our proposed test fills the gap by providing a pathway-based test, workable on CVs or RVs, and adaptively accommodate itself to maintain a high power for various unknown data scenarios in a longitudinal data analysis setting. Because any pathway analysis involves multiple genes, each containing multiple SNPs, it is desirable to apply a test that can maintain high power with a large number of non-associated SNPs (or genes) and multiple only weakly associated SNPs (or genes), an ideal case for our proposed test. On the other hand, because the genes in a pathway can contain different numbers of SNPs, to avoid undue influence from a large (or small) gene, we modify the tests to take account of varying gene lengths. Our proposed test introduces two parameters γ and $\gamma 2$ to achieve the data adaptivity. For example, if there are only few genes, each containing many associated SNPs (e.g., due to LD), a large

value of γ and a small value of $\gamma 2$ would yield a more powerful test; because the truth is unknown, we use data to adaptively estimate their optimal values. The adaptivity of the proposed test at the gene level and/or at the SNP level is missing from many existing tests for pathway or SNP set analysis, such as the SSU and SKAT tests. As supported by our numerical examples and illustration in the ARIC study, the proposed test is almost always powerful in many situations as a tool complementary to existing SNP-set based methods like LSSU or LSKAT.

Our proposed test is general and applicable to CVs or RVs. As shown for the SSU test in Basu and Pan,[Basu and Pan, 2011] via suitable weighting on SNPs, it can be easily modified to analyze both CVs and RVs. In addition, we can also introduce some weights on SNPs and/or genes incorporating biological knowledge, such as the likelihood of SNPs or genes being causal. See Wu et al.[Wu et al., 2011] for more discussion on specifying weights. With regard to longitudinal data analysis in GWAS settings, our proposed test treats the SNP effects as time averaged group effects, and is able to identify the causal effects with increased power. The increased power comes from treating each person as his/her own control. For most outcomes Y, there is considerable variability across individuals due to the influence of unmeasured characteristics such as genetic make-up, environmental exposures, personal behaviors/habits, and so forth. While these factors tend to persist over time for the same individual, their influences are canceled in the estimation of the time averaged SNP effect, and thus lead to more accurate estimate (with smaller variance) and a larger power. Our proposed test can be also easily modified to include the SNP by time interaction into the test, i.e., joint test the time averaged SNP effect and SNP-by-time interaction effect. We only need to expand the X_i matrix as in $H_i = (Z_i; X_i)$, representing the genotype score, to include the SNP-by-time values as extra columns (thus we augument the X_i from $k \times p$ to $k \times p(1 + k)$). With all other settings the same, the LaSPUpath test will effectively implement the joint

test by testing $H_o : \beta = (\beta_1, \beta_2, \ldots, \beta_{p(1+k)})' = 0$. Furthermore, we have focused on testing on a single pathway; an alternative is to take account of possible overlapping or hierarchical structures of some pathways as discussed in Schaid et.al.[Schaid et al., 2012] Finally, we note that our proposed approach is in the category of "self-contained tests," not "competitive tests."[Goeman and Bühlmann, 2007, Liu et al., 2007, Nam and Kim, 2008, Wang et al., 2010, Fridley et al., 2010, Fridley and Biernacka, 2011]. Self-contained (a.k.a. Constrained) tests hypothesize there is no gene in the gene set associated with the phenotype, while competitive tests hypothesize the same level of association of a gene set with the given phenotype as the complement of the gene set. Furthermore, as argued by Goeman and Buhlmann,[Goeman and Bühlmann, 2007] the former type of test is necessarily more powerful than the latter type of test. Following Zhou et.al.,[Zhou et al., 2013] we can extend our LaSPUpath to competitive testing. Our goal also differs from that of Newton et.al.,[Newton et al., 2012] which goes beyond only identifying significant pathways, but also aims to uncover the common theme shared among the identified significant pathways. These extensions are very interesting and warrant future investigation.

In data application of LaSPUpath, we are not limited to use genetic pathway as the sole definition of functional related genes in a gene set. For example, we can define a gene set with similar annotations of functional variation in genomes, using RegulomeDB[Boyle et al., 2012] and/or ENCODE (https://www.encodeproject.org/). We could also define a gene set with similar epigenome activities, such as histone markers, DNA methylations and chromosome structural alternations. The Human Epigenome Atlas (http://www.genboree.org/epigenomeatlas/index.rhtml) and Roadmap Epigenomic Data at NCBI/GEO (http://www.roadmapepigenomics.org/) provide us such opportunities to look into epigenome annotation data. We could also define a gene set with genes functionally related inferred from their products - Protein. By combining the Protein-

Protein-Interaction (PPI) information, such as from BioGRID (http://thebiogrid.org/) and STRING (http://string-db.org/), LaSPUpath could work on a gene set with their corresponding gene products associated in a protein interaction network. Likewise, the LaSPUpath could work on a gene set with more extensive definitions of functionally relatedness in the future, provide researchers a powerful tool to investigate many different hypotheses depending on the annotation type.

5 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

Name of Journal Proposed for Article Submission

Bioinformatics (application note)

Abstract

Summary: Genome-wide association studies (GWASs) with repeated measurements in traits have become available in recent years. However, there is no sophisticated tool to analyze the longitudinal traits in GWAS with appropriate adjustment of the within subject correlation and missing data in repeated measurements. In addition, due to the complex and varying association signal patterns across genome, a data adaptive test is preferred as it can accommodate itself to maintain a higher power compared with those conventional tests. Furthermore, since multiple SNPs in a set (e.g., a gene) or a biological pathway may influence the same trait, a set-based or pathway-based testing may gain additional power by aggregating moderate to weak association signals compared with single variant based analysis. Last but not the least, a flexible parallel computation framework will largely facilitate the intensive computations entailed by the whole-genome association analysis. In this article, the computational tool *LaSPU* is presented.

Availability and implementation: LaSPU is a command line based program written in R and Bash shell script and can be executed on all Unix-compatible platforms. LaSPU is licenced under the CC BY-NC 4.0 licence and freely available at `https://github.com/xyy2006/LaSPU/`. **Contact**: yang.yang@uth.tmc.edu;peng.wei@uth.tmc.edu

5.1 Introduction

Genome-wide association studies (GWAS) have been popular since 2007. Some of these studies have repeated measurements of traits across study time available for each individual, e.g., the Atherosclerosis Risk in Communities (ARIC) study[Heiss, 1989] and the multi-ethnic study of atherosclerosis.[Bild et al., 2002] Association tests that fully utilize the information across time points tend to achieve a higher power and identify more disease-associated loci [Furlotte et al., 2012, Xu et al., 2014]. Recent studies showed the advantages of set-based association tests over the single SNP based analyses. Among kinds of set-based association tests, the data-adaptive tests accommodate themselves for different association patterns encountered in the genome-wide scan, e.g., the ratio of causal SNPs in a region and their effect directions, and thus could maintain a higher power as compared to others.[Han and Pan, 2010, Sun et al., 2013, Pan et al., 2014, Derkach et al., 2013] Therefore, it is desirable to have a set-based and data-adaptive association test for longitudinal trait with correct adjustment of within subject correlation and missing data in repeated measurements. Below we describe the LaSPU tool that implements such a method. LaSPU is a command line based program written in R and Bash shell script with parallel computing support for whole genome association scan task.

5.2 Methods

5.2.1 Features

The LaSPU tool implements the longitudinal adaptive sum of powered test (LaSPU) method. The LaSPU method extends the original aSPU method[Pan et al., 2014] within the Generalized Estimating Equations (GEE) framework to analyze longitudinal trait and was modified to account for those special issues occurred only under the longitudi-

nal data analysis environment, such as working correlation structure matrix of varying dimension accounting for missing data in repeated measurements and covariate-adjusted permutation strategy accounting for time order and missing data in repeated measurements. LaSPU tool is operated under the Unix-like systems in a command line based manner, where users can specify kinds of program arguments, such as those in appointing the input/output file and setting up the serial/parallel computing environment.

5.2.2 Workflow

LaSPU will automatically install several dependent R packages upon first time running. These prerequisite R packages include doMC/doSNOW for parallel computing environment setup, stringr/plyr/dplyr/data.table for more efficient data manipulation, geepack/mvtnorm for extended statistical functions, pipeR for simple program syntax and getopt for command line prompt function. All these R packages are freely available from R CRAN (https://cran.r-project.org/).

LaSPU is straightforward to operate even by first-time users. Users first need to appoint the input file names for genotype file, genotype annotation file (annotation file should at least contain the set-grouping information for SNPs, e.g., into genes) and phenotype file (prepared in long format) respectively, which can be matrix-like data previously saved in either the tab-delimited text format or pre-stored as an R "data.frame" object in the binary ".Rdata" file (with the latter more convenient for experienced R users). Second, users need to specify the response variable name of interest as in phenotype file (i.e., the column name in the matrix) and optionally appoint the filename for covariate file if there are any covariates to adjust for in the association test. Third, users need to specify a list of parameters, including the seed number for random number generation (so that result can be reproduced), the simulation/permutation times (e.g., 1000 for initial scan), the

permutation/simulation switch flag, and the optional filename prefix to all output files. Last but not least, users can setup the computation environment by choosing one from the three schema: serial (i.e., one core on one node), single node with multiple cores, and multiple node each with multiple cores (in this case, users need to provide the multiple nodes profile file in text format).

LaSPU tool by default will execute five methods: LSSU (and LSSU weighted),[Pan, 2009] GEE-Score, L-UminP, L-Sum[Pan, 2009] and LaSPU method (manuscript in preparation), on the longitudinal trait and genotype data sets. Since these five methods share more or less the same intermediate results from the computation, it is straightforward and efficient to execute them simultaneously within the LaSPU tool. The tool, after reading in the necessary data and integrating them based on the subject level, will first run some rudimentary quality contrl (QC) procedures for set-based analysis, such as excluding SNPs with a larger missing rate, monomorphic SNP sites and those sets/genes with 0 or 1 SNPs inside. We assume users will achieve a more comprehensive QC on the data sets beforehand. After QC, LaSPU tool will first fit the null model (which includes the covariates only) via GEE, to derive the score vector of the SNPs in a set and the corresponding variance-covariance matri. Note, under the null hypothesis, there is no SNP associated with the trait. Afterwards, for LaSPU method only, optionally, a permutation procedure could be adopted to obtain a more accurate empirical null distribution of the score vector. Finally, the five aforementioned methods will be executed based on the input of previously calculated score vector and variance-covariance matrix to calculate the significance level of a set of SNPs. In the end, users will have the result saved to disk as ".Rdata" binary file, with report of p values of those genes under different tests. Users can pick one test a priori as their targeted test. We recommend using LaSPU method for achieving the highest power, while the other tests are much faster and suitable for preliminary analysis of whole-genome data. The workflow is charted in Figure

14. The example output of p values for a specific method can be plotted as Manhattan plot (Figure 15)

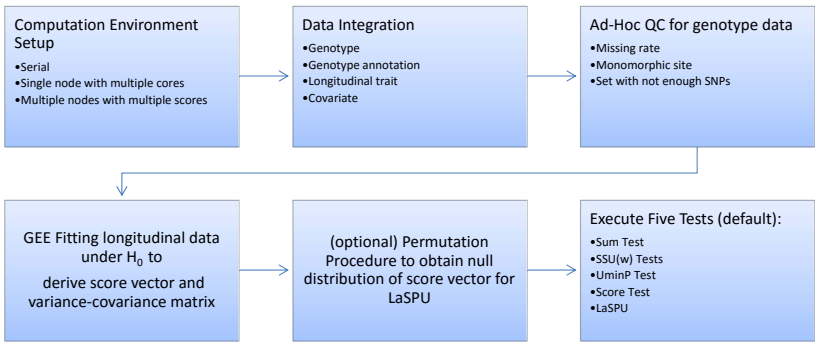

Figure 14: LaSPU Workflow Chart.

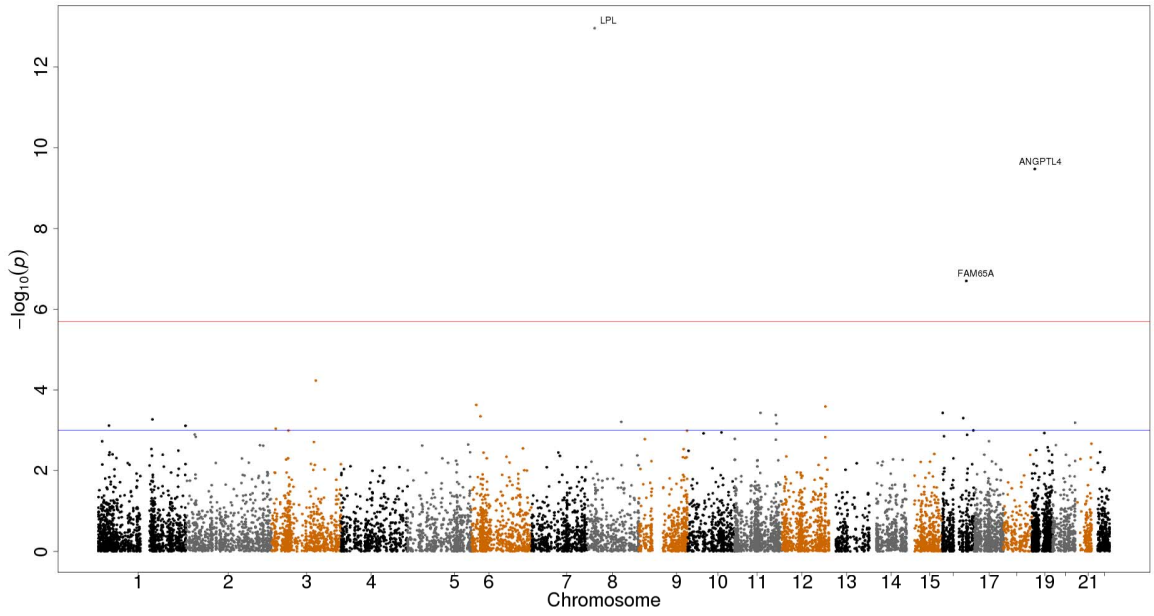

Figure 15: Example Output by LaSPU in Manhattan Plot.

5.3 Conclusion

LaSPU is an important and simple-to-use tool to handle association studies with longitudinal traits. It provides a set-based and data-adaptive association test for longitudinal traits with appropriate adjustment of within-subject correlation and missing data in repeated measurements. The tool enables users to have flexible choices of implementation of the analysis, such as the choice of simulation or permutation based strategy, number of simulations/permutations and serial or parallel computing.

5.4 Appendix for Article 3 - LaSPU Manual

The assoc-LaSPU 1.0 is a Linux command line operated program compatible with most Unix-like system. The main executable is 'assoc-aSPU.r'. There are several folders under the root path. Folder 'AnnotationData' stores the example genotype annotation file; 'CovariateNames' stores the example file appointing covariate names; 'GenotypeData' stores the example genotype file; 'PhenotypeCovariateData' stores the example file including both phenotype and covariate data, stored as a long format matrix; 'NodesProfile' stores the configuration file for multiple nodes, only applicable when users enable the corresponding parallel computing schema; 'R' stores the kernel R functions; 'OutputData' stores the output from the program; 'Demo' stores the explanation file for those examples files, and the bash script file with example command lines to operate on the example data (including example annotation file, example covariate file, example genotype file, etc.); When 'assoc-aSPU.r' is executed with proper inputs and options, all necessary functions will be called and implemented automatically. Results from the program will be saved to the folder "OutputData". If any error occurs during the program run, the error message together with normal output messages of the program will be logged in a '*.log' file under program root path. In the same time, an '*.rda' file containing R runtime environment will be saved to the 'OutputData' folder for further error debugging purpose.

Steps of operating the program:

Note: "$" starts a Linux shell command line.

1. Install all the packages as in "./R/head.r" manually (only for the first time).

2. The first line of 'assoc-aSPU.r', i.e., the shebang of Linux, should automatically call your local 'Rscript' executable; otherwise, modify it to your current customized path of 'Rscript':

#!/your-path-of-Rscript –slave –vanilla

3. "$ chmod u+x assoc-aSPU.r" to add executable privilege to current user if not existed yet.

4. "$ ./assoc-aSPU.r -h" to see help documents.

5. Check 'Demo' folder:
'ExampleData_demo.txt': shows the format of all kinds of input data you need to prepare.
'commandLine_demo.sh': shows the example command line operating on the example data.

6. You are ready to prepare your own data into preferred format and run the program.

Some Usage Notes:

1. For fastest speed, use '–include_aSPU FALSE' to only include five asymptotic tests (which are fast).

2. When including LaSPU family tests, you are recommended to use parallel computing schema as specified in '–parallel_scheme' argument.

3. Check the '–parallel_scheme' and '–parallel_over_gene' arguments to figure out a parallel computing scheme best for your dataset.

4. With '–usePermutedU TRUE', LaSPU family tests will use permutation method instead of simulation method to calculate the p value, which will cost significantly more computation time. It is advised that in the real data application involving rare variants, users could first run simulation based LaSPU test to obtain the global significant genetic loci, and then apply the more time-consuming permutation based LaSPU test on those selected loci as validation.

5. If you have multiple nodes available, check '–jobExecuteOnMultipleNodes' and '–MultipleNodes_profile' arguments to learn how to enable parallel run on multiple nodes

with multiple cores. The example node profile configuration is deposited under the folder "NodesProfile", which can be easily adapted to work under your cluster of nodes. Using multiple nodes is usually only necessary when you want to run LaSPU family tests with '–usePermutedU TRUE'.

6 Conclusion and Future Directions

This dissertation work focused on the development of a novel class of method called LaSPU or LaSPUpath, to implement the data-adaptive association test on longitudinal trait, in a set-based or pathway-based manner of aggregating information from multiple SNPs. Correctly utilizing the extra information from time trajectory as endowed in the repeated measurements could increase the analysis power of identifying associated variants. Set-based or pathway-based association test tends to have a larger power by integrating the useful information across multiple SNPs. Data-adaptive association test relieves the problem that regional signal patterns across the whole genome are complicated and unpredictable so that a single type of association test will usually lose power, by providing a class of tests rewarding an increasingly smaller set of informative SNPs in a region and adaptively selecting the most powerful one with adjustment for multiple testing. I demonstrated that LaSPU and LaSPUpath have superior performance in the numerical examples, and illustrated their application in the ARIC study. The tool LaSPU implements these two methods effectively with ease and flexible control provided to users.

Up to now, many prospective cohort studies and electronic health record (EHR)-based cohorts have collected longitudinal phenotype information. For example, The Atherosclerosis Risk in Communities (ARIC) study, as the data used in our application examples, collected around 11,478 samples from European ancestry and 4,314 samples from Afican ancestry, with follow-up measurements on many traits associated with atherosclerosis risk. The Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort includes over 100,000 adults in California, US, with their longitudinal phenotype stored in EHR. The UK Biobank[Sudlow et al., 2015] (http://www.ukbiobank.ac.uk/) is a very large and detailed prospective study with over 500,000 participants recruited between 2006 and 2010, where longitudinal follow-up for a wide range of heath-related outcomes are available. The Precision Medicine Initiative (PMI)[Collins and Varmus, 2015]

(http://www.nih.gov/research-training/precision-medicine-initiative) planned by NIH is an emerging approach for disease treatment and preventions that takes into account individual variability in genes, environment, and lifestyle for each person. PMI will collect over 1,000,000 participants in US and include longitudinal follow-up for multiple health-related outcomes. The list goes on. As observed, with more and more availability in studies with a larger sample size and longitudinal information of subjects, our proposed LaSPU and LaSPUpath methods will have more room to play the roles in identifying the genetic components accounted for disease onset and progression.

There are a few other interesting topics I identified during the development of the dissertation. I did discuss about these possible extension works in each article separately. Here I will give a summary. First, I focused on testing the time-averaged group effect of SNP genotype in the longitudinal settings, however, a joint test of both time-averaged group effect and the interaction effect between time and SNP genotype could be extended within the LaSPU framework. Second, proper weighting of a SNP in a gene or a gene in a pathway incorporating the biological information may increase the test power. Future research could try different weighting strategy, such as the weight follows a beta distribution, to incorporate various biological information, such as functional annotation and conservation score. Third, proper weighting of SNPs in a region, for example the weight is set to be proportional to the inverse variance of the SNP, can make the test capable of testing CV and RV at the same time, and thus lead to an aggregation of more comprehensive information in a region and then may have a larger power. Fourth, in LaSPUpath, I have focused on testing on a single pathway; an alternative is to take account of possible overlapping or hierarchical structures of some pathways as discussed in Schaid et.al.[Schaid et al., 2012] Fifth, in data application of LaSPUpath, it is not limited to use only genetic pathway as the sole definition of functional related genes in a gene set. For example, I can define a gene set with similar annotations of functional variation in genomes, using

RegulomeDB[Boyle et al., 2012] and/or ENCODE (https://www.encodeproject.org/). I could also define a gene set with similar epigenome activities, such as histone markers, DNA methylations and chromosome structural alternations. The Human Epigenome Atlas (http://www.genboree.org/epigenomeatlas/index.rhtml) and Roadmap Epigenomic Data at NCBI/GEO (http://www.roadmapepigenomics.org/) provide us such opportunities to look into epigenome annotation data. I could also define a gene set with genes functionally related inferred from their products - Protein. By combining the Protein-Protein-Interaction (PPI) information, such as from BioGRID (http://thebiogrid.org/) and STRING (http://string-db.org/), LaSPUpath could work on a gene set with their corresponding gene products associated in a protein interaction network. Likewise, the LaSPUpath could work on a gene set with more extensive definitions of functionally relatedness in the future, provide researchers a powerful tool to investigate many different hypotheses depending on the annotation type. Finally, it is noted that our proposed approach is in the category of "self-contained tests," not "competitive tests."[Goeman and Bühlmann, 2007, Liu et al., 2007, Nam and Kim, 2008, Wang et al., 2010, Fridley et al., 2010, Fridley and Biernacka, 2011]. Although as argued by Goeman and Buhlmann,[Goeman and Bühlmann, 2007] the former type of test is necessarily more powerful than the latter type of test, I can extend our LaSPUpath to competitive testing following Zhou et.al.,[Zhou et al., 2013].

In conclusion, this dissertation work breaks a new path in testing association between genotypes and longitudinal trait, which will become more and more prevalent as nowadays fashionable follow-up technology enables, such as mobile health monitoring equipment. The tool implementing the new methods accommodates itself to the trend of big data era by incorporating state-of-the-art technologies, such as parallel computing in cloud. With lots of extension possibilities in the future, the work is tremendously significant to the public health community.

References

[Ansorge, 2009] Ansorge, W. J. (2009). Next-generation dna sequencing techniques. *New biotechnology*, 25(4):195–203.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

[Association et al., 2014] Association, A. et al. (2014). 2014 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 10(2):e47–e92.

[Aulchenko et al., 2009] Aulchenko, Y. S., Ripatti, S., Lindqvist, I., Boomsma, D., Heid, I. M., Pramstaller, P. P., Penninx, B. W. J. H., Janssens, A. C. J. W., Wilson, J. F., Spector, T., Martin, N. G., Pedersen, N. L., Kyvik, K. O., Kaprio, J., Hofman, A., Freimer, N. B., Jarvelin, M.-R., Gyllensten, U., Campbell, H., Rudan, I., Johansson, A., Marroni, F., Hayward, C., Vitart, V., Jonasson, I., Pattaro, C., Wright, A., Hastie, N., Pichler, I., Hicks, A. A., Falchi, M., Willemsen, G., Hottenga, J.-J., de Geus, E. J. C., Montgomery, G. W., Whitfield, J., Magnusson, P., Saharinen, J., Perola, M., Silander, K., Isaacs, A., Sijbrands, E. J. G., Uitterlinden, A. G., Witteman, J. C. M., Oostra, B. A., Elliott, P., Ruokonen, A., Sabatti, C., Gieger, C., Meitinger, T., Kronenberg, F., Döring, A., Wichmann, H.-E., Smit, J. H., McCarthy, M. I., van Duijn, C. M., Peltonen, L., and , E. N. G. A. G. E. C. (2009). Loci influencing lipid levels and coronary heart disease risk in 16 european population cohorts. *Nat Genet*, 41(1):47–55.

[Bansal et al., 2010] Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet*, 11(11):773–785.

[Basu and Pan, 2011] Basu, S. and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol*, 35(7):606–619.

[Belsky et al., 2013] Belsky, D. W., Moffitt, T. E., Baker, T. B., Biddle, A. K., Evans, J. P., Harrington, H., Houts, R., Meier, M., Sugden, K., Williams, B., et al. (2013). Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study. *JAMA psychiatry*, 70(5):534–542.

[Benros et al., 2014] Benros, M. E., Nielsen, P. R., Nordentoft, M., Eaton, W. W., Dalton, S. O., and Mortensen, P. B. (2014). Autoimmune diseases and severe infections as risk factors for schizophrenia: a 30-year population-based register study.

[Bhatia et al., 2010] Bhatia, G., Bansal, V., Harismendy, O., Schork, N. J., Topol, E. J., Frazer, K., and Bafna, V. (2010). A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS computational biology*, 6(10):e1000954.

[Bild et al., 2002] Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Roux, A. V. D., Folsom, A. R., Greenland, P., JacobsJr, D. R., Kronmal, R., Liu, K., et al. (2002). Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology*, 156(9):871–881.

[Boyko et al., 2008] Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics*, 4(5):e1000083.

[Boyle et al., 2012] Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., et al. (2012). Annota-

tion of functional variation in personal genomes using regulomedb. *Genome research*, 22(9):1790–1797.

[Cai et al., 2012] Cai, T., Lin, X., and Carroll, R. J. (2012). Identifying genetic marker sets associated with phenotypes via an efficient adaptive score test. *Biostatistics*, 13(4):776–790.

[Cantor et al., 2010] Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing gwas results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22.

[Capanu et al., 2011] Capanu, M., Concannon, P., Haile, R. W., Bernstein, L., Malone, K. E., Lynch, C. F., Liang, X., Teraoka, S. N., Diep, A. T., Thomas, D. C., Bernstein, J. L., , W. E. C. A. R. E. S. C. G., and Begg, C. B. (2011). Assessment of rare brca1 and brca2 variants of unknown significance using hierarchical modeling. *Genet Epidemiol*, 35(5):389–397.

[Caspi et al., 2008] Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., et al. (2008). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 36(suppl 1):D623–D631.

[Casserly and Topol, 2004] Casserly, I. and Topol, E. J. (2004). Convergence of atherosclerosis and alzheimer's disease: inflammation, cholesterol, and misfolded proteins. *The Lancet*, 363(9415):1139–1146.

[Chai et al., 2009] Chai, H.-S., Sicotte, H., Bailey, K. R., Turner, S. T., Asmann, Y. W., and Kocher, J.-P. A. (2009). Glossi: a method to assess the association of genetic loci-sets with complex diseases. *BMC bioinformatics*, 10(1):102.

[Chen et al., 2009] Chen, L., Zhang, L., Zhao, Y., Xu, L., Shang, Y., Wang, Q., Li, W., Wang, H., and Li, X. (2009). Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing snps and pathways. *Bioinformatics*, 25(2):237–242.

[Chen et al., 2012] Chen, L. S., Hsu, L., Gamazon, E. R., Cox, N. J., and Nicolae, D. L. (2012). An exponential combination procedure for set-based association tests in sequencing studies. *The American Journal of Human Genetics*, 91(6):977–986.

[Chen et al., 2010a] Chen, L. S., Hutter, C. M., Potter, J. D., Liu, Y., Prentice, R. L., Peters, U., and Hsu, L. (2010a). Insights into colon cancer etiology via a regularized approach to gene set analysis of gwas data. *The American Journal of Human Genetics*, 86(6):860–871.

[Chen et al., 2010b] Chen, X., Wang, L., Hu, B., Guo, M., Barnard, J., and Zhu, X. (2010b). Pathway-based analysis for genome-wide association studies using supervised principal components. *Genetic epidemiology*, 34(7):716–724.

[Cheng et al., 1999] Cheng, S., Grow, M. A., Pallaud, C., Klitz, W., Erlich, H. A., Visvikis, S., Chen, J. J., Pullinger, C. R., Malloy, M. J., Siest, G., et al. (1999). A multilocus genotyping assay for candidate markers of cardiovascular disease risk. *Genome research*, 9(10):936–949.

[Collins and Varmus, 2015] Collins, F. S. and Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795.

[Cortes and Brown, 2011] Cortes, A. and Brown, M. A. (2011). Promise and pitfalls of the immunochip. *Arthritis Res Ther*, 13(1):101.

[Cox and Snell, 1989] Cox, D. R. and Snell, E. J. (1989). *Analysis of binary data*, volume 32. CRC Press.

[Craig et al., 2008] Craig, J. et al. (2008). Complex diseases: Research and applications. *Nature Education*, 1(1):184.

[De la Cruz et al., 2010] De la Cruz, O., Wen, X., Ke, B., Song, M., and Nicolae, D. L. (2010). Gene, region and pathway level analyses in whole-genome studies. *Genet Epidemiol*, 34(3):222–231.

[Dennis Jr et al., 2003] Dennis Jr, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., Lempicki, R. A., et al. (2003). David: database for annotation, visualization, and integrated discovery. *Genome biol*, 4(5):P3.

[Derkach et al., 2013] Derkach, A., Lawless, J. F., and Sun, L. (2013). 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(1):110–121.

[Diggle et al., 2002] Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of longitudinal data*. Oxford University Press.

[Eichler et al., 2010] Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450.

[Fan, 1996] Fan, J. (1996). Test of significance based on wavelet thresholding and neyman's truncation. *Journal of the American Statistical Association*, 91(434):674–688.

[Fan et al., 2013] Fan, R., Wang, Y., Mills, J. L., Wilson, A. F., Bailey-Wilson, J. E., and Xiong, M. (2013). Functional linear models for association analysis of quantitative traits. *Genetic epidemiology*, 37(7):726–742.

[Fan et al., 2012] Fan, R., Zhang, Y., Albert, P. S., Liu, A., Wang, Y., and Xiong, M. (2012). Longitudinal association analysis of quantitative traits. *Genetic epidemiology*, 36(8):856–869.

[Feng et al., 2011] Feng, T., Elston, R. C., and Zhu, X. (2011). Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (spwss, orwss). *Genetic epidemiology*, 35(5):398–409.

[Fernandes et al., 1999] Fernandes, M., Proenca, M., Nogueira, A., Oliveira, L., Santiago, B., Santana, I., and Oliveira, C. (1999). Effects of apolipoprotein e genotype on blood lipid composition and membrane platelet fluidity in alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1454(1):89–96.

[for Blood Pressure Genome-Wide Association Studies et al., 2011] for Blood Pressure Genome-Wide Association Studies, I. C. et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–109.

[Fridley and Biernacka, 2011] Fridley, B. L. and Biernacka, J. M. (2011). Gene set analysis of snp data: benefits, challenges, and future directions. *European Journal of Human Genetics*, 19(8):837–843.

[Fridley et al., 2010] Fridley, B. L., Jenkins, G. D., and Biernacka, J. M. (2010). Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One*, 5(9).

[Fu et al., 2013] Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., , N. H. L. B. I. E. S. P., and Akey, J. M. (2013). Analysis of

6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220.

[Furlotte et al., 2012] Furlotte, N. A., Eskin, E., and Eyheramendy, S. (2012). Genome-wide association mapping with longitudinal data. *Genetic epidemiology*, 36(5):463–471.

[Go et al., 2013] Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Borden, W. B., Bravata, D. M., SHIFAN, D., Ford, E. S., Fox, C. S., et al. (2013). Executive summary: Heart disease and stroke statistics: 2013 update: A report from the american heart association. *Circulation*, 127(1):143–146.

[Goeman and Bühlmann, 2007] Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

[Gough, 2002] Gough, N. R. (2002). Science's signal transduction knowledge environment. *Annals of the New York Academy of Sciences*, 971(1):585–587.

[Grove et al., 2013] Grove, M. L., Yu, B., Cochran, B. J., Haritunians, T., Bis, J. C., Taylor, K. D., Hansen, M., Borecki, I. B., Cupples, L. A., Fornage, M., et al. (2013). Best practices and joint calling of the humanexome beadchip: the charge consortium. *PLoS One*, 8(7):e68095.

[Gui et al., 2011] Gui, H., Li, M., Sham, P. C., and Cherny, S. S. (2011). Comparisons of seven algorithms for pathway analysis using the wtccc crohn's disease dataset. *BMC research notes*, 4(1):386.

[Guo et al., 2005] Guo, X., Pan, W., Connett, J. E., Hannan, P. J., and French, S. A. (2005). Small-sample performance of the robust score test and its modifications in generalized estimating equations. *Statistics in medicine*, 24(22):3479–3495.

[Han and Pan, 2010] Han, F. and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human heredity*, 70(1):42–54.

[He et al., 2015] He, Z., Zhang, M., Lee, S., Smith, J. A., Guo, X., Palmas, W., Kardia, S. L., Roux, A. V. D., and Mukherjee, B. (2015). Set-based tests for genetic association in longitudinal studies. *Biometrics*.

[Heiss, 1989] Heiss, G. (1989). Atherosclerosis risk in communities: a follow-up study of early arterial lesions in the general population. *Pathology of Human Atherosclerotic Plaque. New York, NY: Springer-Verlag*, 877:888.

[Hernandez, 2008] Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 24(23):2786–2787.

[Hindorff et al., 2009] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367.

[Hirschhorn, 2009] Hirschhorn, J. N. (2009). Genomewide association studies–illuminating biologic pathways. *New England Journal of Medicine*, 360(17):1699.

[Hirschhorn and Daly, 2005] Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108.

[Hoffmann et al., 2010] Hoffmann, T. J., Marini, N. J., and Witte, J. S. (2010). Comprehensive approach to analyzing rare genetic variants. *PLoS One*, 5(11):e13584.

[Hofman et al., 1997] Hofman, A., Ott, A., Breteler, M. M., Bots, M. L., Slooter, A. J., van Harskamp, F., van Duijn, C. N., Van Broeckhoven, C., and Grobbee, D. E. (1997).

Atherosclerosis, apolipoprotein e, and prevalence of dementia and alzheimer's disease in the rotterdam study. *The Lancet*, 349(9046):151–154.

[Holden et al., 2008] Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). Gsea-snp: applying gene set enrichment analysis to snp data from genome-wide association studies. *Bioinformatics*, 24(23):2784–2785.

[Holmans, 2009] Holmans, P. (2009). Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Advances in genetics*, 72:141–179.

[Holmans et al., 2009] Holmans, P., Green, E. K., Pahwa, J. S., Ferreira, M. A., Purcell, S. M., Sklar, P., Owen, M. J., O'Donovan, M. C., and Craddock, N. (2009). Gene ontology analysis of gwa study data sets provides insights into the biology of bipolar disorder. *The American Journal of Human Genetics*, 85(1):13–24.

[Hu et al., 2013] Hu, H., Huff, C. D., Moore, B., Flygare, S., Reese, M. G., and Yandell, M. (2013). Vaast 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic epidemiology*, 37(6):622–634.

[Investigators et al., 1989] Investigators, A. et al. (1989). The atherosclerosis risk in communit (aric) stui) y: Design and objectwes. *American journal of epidemiology*, 129(4):687–702.

[Inzucchi et al., 2012] Inzucchi, S. E., Bergenstal, R. M., Buse, J. B., Diamant, M., Ferrannini, E., Nauck, M., Peters, A. L., Tsapas, A., Wender, R., and Matthews, D. R. (2012). Management of hyperglycemia in type 2 diabetes: a patient-centered approach position statement of the american diabetes association (ada) and the european association for the study of diabetes (easd). *Diabetes care*, 35(6):1364–1379.

[Ionita-Laza et al., 2011] Ionita-Laza, I., Buxbaum, J. D., Laird, N. M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS genetics*, 7(2):e1001289.

[Ionita-Laza et al., 2013] Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet*, 92(6):841–853.

[Ionita-Laza et al., 2007] Ionita-Laza, I., McQueen, M. B., Laird, N. M., and Lange, C. (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. *The American Journal of Human Genetics*, 81(3):607–614.

[Joshi-Tope et al., 2005] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., et al. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428–D432.

[Kamatani et al., 2010] Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y., and Kamatani, N. (2010). Genome-wide association study of hematological and biochemical traits in a japanese population. *Nat Genet*, 42(3):210–215.

[Karp et al., 2002] Karp, P. D., Riley, M., Paley, S. M., and Pellegrini-Toole, A. (2002). The metacyc database. *Nucleic acids research*, 30(1):59–61.

[Kathiresan et al., 2007] Kathiresan, S., Manning, A. K., Demissie, S., D'Agostino, R. B., Surti, A., Guiducci, C., Gianniny, L., Burtt, N. P., Melander, O., Orho-Melander, M., Arnett, D. K., Peloso, G. M., Ordovas, J. M., and Cupples, L. A. (2007).

A genome-wide association study for blood lipid phenotypes in the framingham heart study. *BMC Med Genet*, 8 Suppl 1:S17.

[Keating et al., 2008] Keating, B. J., Tischfield, S., Murray, S. S., Bhangale, T., Price, T. S., Glessner, J. T., Galver, L., Barrett, J. C., Grant, S. F., Farlow, D. N., et al. (2008). Concept, design and implementation of a cardiovascular gene-centric 50 k snp array for large-scale genomic association studies. *PloS one*, 3(10):e3583.

[Kim et al., 2014] Kim, S., Pan, W., and Shen, X. (2014). Penalized regression approaches to testing for quantitative trait-rare variant association. *Frontiers in genetics*, 5.

[Korn and Whittemore, 1979] Korn, E. L. and Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics*, pages 795–802.

[Kwee et al., 2008] Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397.

[Laird and Ware, 1982] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

[Lange et al., 2014] Lange, L. A., Hu, Y., Zhang, H., Xue, C., Schmidt, E. M., Tang, Z.-Z., Bizon, C., Lange, E. M., Smith, J. D., Turner, E. H., et al. (2014). Whole-exome sequencing identifies rare and low-frequency coding variants associated with ldl cholesterol. *The American Journal of Human Genetics*, 94(2):233–245.

[Lee et al., 2012a] Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., , N. H. L. B. I. G. O. E. S. P.-E. S. P. L. P. T., Christiani, D. C.,

Wurfel, M. M., and Lin, X. (2012a). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*, 91(2):224–237.

[Lee et al., 2012b] Lee, S., Wu, M. C., and Lin, X. (2012b). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.

[Li and Leal, 2008] Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*, 83(3):311–321.

[Li et al., 2011] Li, M.-X., Gui, H.-S., Kwan, J. S., and Sham, P. C. (2011). Gates: a rapid and powerful gene-based association test using extended simes procedure. *The American Journal of Human Genetics*, 88(3):283–293.

[Li et al., 2012] Li, M.-X., Kwan, J. S., and Sham, P. C. (2012). Hyst: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. *The American Journal of Human Genetics*, 91(3):478–488.

[Liang and Zeger, 1986] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

[Lin, 2005] Lin, D. (2005). An efficient monte carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21(6):781–787.

[Lin and Tang, 2011] Lin, D.-Y. and Tang, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, 89(3):354–367.

[Litière et al., 2007] Litière, S., Alonso, A., and Molenberghs, G. (2007). Type i and type ii error under random-effects misspecification in generalized linear mixed models. *Biometrics*, 63(4):1038–1044.

[Liu and Leal, 2010] Liu, D. J. and Leal, S. M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet*, 6(10):e1001156.

[Liu et al., 2007] Liu, Q., Dinu, I., Adewale, A. J., Potter, J. D., and Yasui, Y. (2007). Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, 8:431.

[Luo et al., 2011] Luo, L., Boerwinkle, E., and Xiong, M. (2011). Association studies for next-generation sequencing. *Genome research*, 21(7):1099–1108.

[Luo et al., 2010] Luo, L., Peng, G., Zhu, Y., Dong, H., Amos, C. I., and Xiong, M. (2010). Genome-wide gene and pathway analysis. *European Journal of Human Genetics*, 18(9):1045–1053.

[Luo et al., 2012a] Luo, L., Zhu, Y., and Xiong, M. (2012a). Quantitative trait locus analysis for next-generation sequencing with the functional linear models. *Journal of medical genetics*, 49(8):513–524.

[Luo et al., 2012b] Luo, L., Zhu, Y., and Xiong, M. (2012b). Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation. *European Journal of Human Genetics*, 21(2):217–224.

[Lynch et al., 1998] Lynch, M., Walsh, B., et al. (1998). Genetics and analysis of quantitative traits.

[Madsen and Browning, 2009] Madsen, B. E. and Browning, S. R. (2009). A group-wise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384.

[Mardis, 2008] Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402.

[McCarthy et al., 2008] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369.

[Medina et al., 2009] Medina, I., Montaner, D., Bonifaci, N., Pujana, M. A., Carbonell, J., Tarraga, J., Al-Shahrour, F., and Dopazo, J. (2009). Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic acids research*, 37(suppl 2):W340–W344.

[Melville et al., 2012] Melville, S. A., Buros, J., Parrado, A. R., Vardarajan, B., Logue, M. W., Shen, L., Risacher, S. L., Kim, S., Jun, G., DeCarli, C., et al. (2012). Multiple loci influencing hippocampal degeneration identified by genome scan. *Annals of neurology*, 72(1):65–75.

[Merched et al., 2000] Merched, A., Xia, Y., Visvikis, S., Serot, J., and Siest, G. (2000). Decreased high-density lipoprotein cholesterol and serum apolipoprotein ai concentrations are highly correlated with the severity of alzheimer's disease. *Neurobiology of aging*, 21(1):27–30.

[Metzker, 2009] Metzker, M. L. (2009). Sequencing technologies the next generation. *Nature Reviews Genetics*, 11(1):31–46.

[Michikawa, 2003] Michikawa, M. (2003). Cholesterol paradox: is high total or low hdl cholesterol level a risk for alzheimer's disease? *Journal of neuroscience research*, 72(2):141–146.

[Morgenthaler and Thilly, 2007] Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutat Res*, 615(1-2):28–56.

[Morris and Zeggini, 2010] Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic epidemiology*, 34(2):188.

[Nam et al., 2010] Nam, D., Kim, J., Kim, S.-Y., and Kim, S. (2010). Gsa-snp: a general approach for gene set analysis of polymorphisms. *Nucleic acids research*, page gkq428.

[Nam and Kim, 2008] Nam, D. and Kim, S.-Y. (2008). Gene-set approach for expression pattern analysis. *Brief Bioinform*, 9(3):189–197.

[Neale et al., 2011] Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS genetics*, 7(3):e1001322.

[Network et al., 2011] Network, C. G. A. R. et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615.

[Newton et al., 2012] Newton, M. A., He, Q., and Kendziorski, C. (2012). A model-based analysis to infer the functional content of a gene list. *Statistical applications in genetics and molecular biology*, 11(2).

[Neyman, 1937] Neyman, J. (1937). Smooth test for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4):149–199.

[Nishimura, 2001] Nishimura, D. (2001). Biocarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient*, 2(3):117–120.

[O'Dushlaine et al., 2009] O'Dushlaine, C., Kenny, E., Heron, E. A., Segurado, R., Gill, M., Morris, D. W., and Corvin, A. (2009). The snp ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics*, 25(20):2762–2763.

[Ogata et al., 1999] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34.

[Oualkacha et al., 2013] Oualkacha, K., Dastani, Z., Li, R., Cingolani, P. E., Spector, T. D., Hammond, C. J., Richards, J. B., Ciampi, A., and Greenwood, C. M. T. (2013). Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genet Epidemiol*, 37(4):366–376.

[Pan, 2001] Pan, W. (2001). On the robust variance estimator in generalised estimating equations. *Biometrika*, 88(3):901–906.

[Pan, 2009] Pan, W. (2009). Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic epidemiology*, 33(6):497–507.

[Pan et al., 2015a] Pan, W., Chen, Y.-M., and Wei, P. (2015a). Testing for polygenic effects in genome-wide association studies. *Genetic epidemiology*, 39(4):306–316.

[Pan et al., 2009] Pan, W., Han, F., and Shen, X. (2009). Test selection with application to detecting disease association with multiple snps. *Human heredity*, 69(2):120–130.

[Pan et al., 2014] Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics*, pages genetics–114.

[Pan et al., 2015b] Pan, W., Kwak, I.-Y., and Wei, P. (2015b). A powerful pathway-based adaptive test for genetic association with common or rare variants. *The American Journal of Human Genetics*, 97(1):86–98.

[Pan and Shen, 2011] Pan, W. and Shen, X. (2011). Adaptive tests for association analysis of rare variants. *Genet Epidemiol*, 35(5):381–388.

[Peloso et al., 2014] Peloso, G. M., Auer, P. L., Bis, J. C., Voorman, A., Morrison, A. C., Stitziel, N. O., Brody, J. A., Khetarpal, S. A., Crosby, J. R., Fornage, M., et al. (2014). 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(2):223–232.

[Peng et al., 2009] Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J. D., Jin, L., et al. (2009). Gene and pathway-based second-wave analysis of genome-wide association studies. *European Journal of Human Genetics*, 18(1):111–117.

[Pinto et al., 2010] Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T. R., Correia, C., Abrahams, B. S., Almeida, J., Bacchelli, E., Bader, G. D., Bailey, A. J., Baird, G., Battaglia, A., Berney, T., Bolshakova, N., Bölte, S., Bolton, P. F., Bourgeron, T., Brennan, S., Brian, J., Bryson, S. E., Carson, A. R., Casallo, G., Casey, J., Chung, B. H. Y., Cochrane, L., Corsello, C., Crawford, E. L., Crossett, A., Cytrynbaum, C., Dawson, G., de Jonge, M., Delorme, R., Drmic, I., Duketis, E., Duque, F., Estes, A., Farrar, P., Fernandez, B. A., Folstein, S. E., Fombonne, E., Freitag, C. M., Gilbert, J., Gillberg, C., Glessner, J. T., Goldberg, J., Green, A., Green, J., Guter, S. J., Hakonarson, H., Heron, E. A., Hill, M., Holt, R., Howe, J. L., Hughes, G., Hus, V., Igliozzi, R., Kim, C., Klauck, S. M., Kolevzon, A., Korvatska, O., Kustanovich, V., Lajonchere, C. M., Lamb, J. A., Laskawiec, M.,

Leboyer, M., Le Couteur, A., Leventhal, B. L., Lionel, A. C., Liu, X.-Q., Lord, C., Lotspeich, L., Lund, S. C., Maestrini, E., Mahoney, W., Mantoulan, C., Marshall, C. R., McConachie, H., McDougle, C. J., McGrath, J., McMahon, W. M., Merikangas, A., Migita, O., Minshew, N. J., Mirza, G. K., Munson, J., Nelson, S. F., Noakes, C., Noor, A., Nygren, G., Oliveira, G., Papanikolaou, K., Parr, J. R., Parrini, B., Paton, T., Pickles, A., Pilorge, M., Piven, J., Ponting, C. P., Posey, D. J., Poustka, A., Poustka, F., Prasad, A., Ragoussis, J., Renshaw, K., Rickaby, J., Roberts, W., Roeder, K., Roge, B., Rutter, M. L., Bierut, L. J., Rice, J. P., Salt, J., Sansom, K., Sato, D., Segurado, R., Sequeira, A. F., Senman, L., Shah, N., Sheffield, V. C., Soorya, L., Sousa, I., Stein, O., Sykes, N., Stoppioni, V., Strawbridge, C., Tancredi, R., Tansey, K., Thiruvahindrapduram, B., Thompson, A. P., Thomson, S., Tryfon, A., Tsiantis, J., Van Engeland, H., Vincent, J. B., Volkmar, F., Wallace, S., Wang, K., Wang, Z., Wassink, T. H., Webber, C., Weksberg, R., Wing, K., Wittemeyer, K., Wood, S., Wu, J., Yaspan, B. L., Zurawiecki, D., Zwaigenbaum, L., Buxbaum, J. D., Cantor, R. M., Cook, E. H., Coon, H., Cuccaro, M. L., Devlin, B., Ennis, S., Gallagher, L., Geschwind, D. H., Gill, M., Haines, J. L., Hallmayer, J., Miller, J., Monaco, A. P., Nurnberger, Jr, J. I., Paterson, A. D., Pericak-Vance, M. A., Schellenberg, G. D., Szatmari, P., Vicente, A. M., Vieland, V. J., Wijsman, E. M., Scherer, S. W., Sutcliffe, J. S., and Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368–372.

[Price et al., 2010] Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L.-J., and Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet*, 86(6):832–838.

[Price et al., 2006] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for

stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909.

[Puglielli et al., 2003] Puglielli, L., Tanzi, R. E., and Kovacs, D. M. (2003). Alzheimer's disease: the cholesterol connection. *Nature neuroscience*, 6(4):345–351.

[Purcell et al., 2007] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

[Sabatti et al., 2008] Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., et al. (2008). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35–46.

[Schaefer et al., 2009] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl 1):D674–D679.

[Schaid et al., 2012] Schaid, D. J., Sinnwell, J. P., Jenkins, G. D., McDonnell, S. K., Ingle, J. N., Kubo, M., Goss, P. E., Costantino, J. P., Wickerham, D. L., and Weinshilboum, R. M. (2012). Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genetic epidemiology*, 36(1):3–16.

[Seaman and Müller-Myhsok, 2005] Seaman, S. and Müller-Myhsok, B. (2005). Rapid simulation of p values for product methods and multiple-testing adjustment in association studies. *The American Journal of Human Genetics*, 76(3):399–408.

[Sham and Purcell, 2014] Sham, P. C. and Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet*, 15(5):335–346.

[Shendure and Ji, 2008] Shendure, J. and Ji, H. (2008). Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145.

[Shete et al., 2004] Shete, S., Beasley, T. M., Etzel, C. J., Fernández, J. R., Chen, J., Allison, D. B., and Amos, C. I. (2004). Effect of winsorization on power and type 1 error of variance components and related methods of qtl detection. *Behavior genetics*, 34(2):153–159.

[Silver et al., 2012] Silver, M., Janousova, E., Hua, X., Thompson, P. M., and Montana, G. (2012). Identification of gene pathways implicated in alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *NeuroImage*, 63(3):1681–1694.

[Song et al., 2013] Song, P., Xue, J., and Li, Z. (2013). Simulation of longitudinal exposure data with variance-covariance structures based on mixed models. *Risk Anal*, 33(3):469–479.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550.

[Sudlow et al., 2015] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):1–10.

[Sul et al., 2011] Sul, J. H., Han, B., He, D., and Eskin, E. (2011). An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genetics*, 188(1):181–188.

[Sun et al., 2013] Sun, J., Zheng, Y., and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic epidemiology*, 37(4):334–344.

[Tintle et al., 2011] Tintle, N., Aschard, H., Hu, I., Nock, N., Wang, H., and Pugh, E. (2011). Inflated type i error rates when using aggregation methods to analyze rare variants in the 1000 genomes project exon sequencing data in unrelated individuals: summary results from group 7 at genetic analysis workshop 17. *Genetic epidemiology*, 35(S1):S56–S60.

[Tsay, 1984] Tsay, R. S. (1984). Regression models with time series errors. *Journal of the American Statistical Association*, 79(385):118–124.

[Voight et al., 2012] Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS genetics*, 8(8):e1002793.

[Wang et al., 2012] Wang, H., Nie, F., Huang, H., Yan, J., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J., Shen, L., et al. (2012). From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps. *Bioinformatics*, 28(18):i619–i625.

[Wang et al., 2007] Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6):1278–1283.

[Wang et al., 2010] Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*, 11(12):843–854.

[Wang and Elston, 2007] Wang, T. and Elston, R. C. (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. *The american journal of human genetics*, 80(2):353–360.

[Wang et al., 2005] Wang, W. Y., Barratt, B. J., Clayton, D. G., and Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6(2):109–118.

[Wang et al., 2013] Wang, X., Lee, S., Zhu, X., Redline, S., and Lin, X. (2013). Gee-based snp set association test for continuous and discrete traits in family-based association studies. *Genetic epidemiology*, 37(8):778–786.

[WARE et al., 1990] WARE, J. H., DOCKERY, D. W., LOUIS, T. A., XU, X., FERRIS, B. G., and SPEIZER, F. E. (1990). Longitudinal and cross-sectional estimates of pulmonary function decline in never-smoking adults. *American journal of epidemiology*, 132(4):685–700.

[Wedderburn, 1974] Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447.

[Wei et al., 2012] Wei, P., Tang, H., and Li, D. (2012). Insights into pancreatic cancer etiology from pathway analysis of genome-wide association study data. *PloS one*, 7(10):e46887.

[Weng et al., 2011] Weng, L., Macciardi, F., Subramanian, A., Guffanti, G., Potkin, S. G., Yu, Z., and Xie, X. (2011). Snp-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*, 12:99.

[Wu et al., 2010] Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *Am J Hum Genet*, 86(6):929–942.

[Wu et al., 2011] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, 89(1):82–93.

[Xu et al., 2014] Xu, Z., Shen, X., Pan, W., Initiative, A. D. N., et al. (2014). Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PloS one*, 9(8):e102312.

[Yandell et al., 2011] Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L. B., and Reese, M. G. (2011). A probabilistic disease-gene finder for personal genomes. *Genome research*, 21(9):1529–1542.

[Ye and Engelman, 2011] Ye, K. Q. and Engelman, C. D. (2011). Detecting multiple causal rare variants in exome sequence data. *Genet Epidemiol*, 35 Suppl 1:S18–S21.

[Yu et al., 2009] Yu, K., Li, Q., Bergen, A. W., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N., Kraft, P., and Chatterjee, N. (2009). Pathway analysis by adaptive combination of p-values. *Genetic epidemiology*, 33(8):700–709.

[Zeger and Liang, 1986] Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130.

[Zeger et al., 1988] Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060.

[Zeger et al., 1985] Zeger, S. L., Liang, K.-Y., and Self, S. G. (1985). The analysis of binary longitudinal data with time independent covariates. *Biometrika*, 72(1):31–38.

[Zeger and Qaqish, 1988] Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, pages 1019–1031.

[Zhang et al., 2010a] Zhang, K., Cui, S., Chang, S., Zhang, L., and Wang, J. (2010a). i-gsea4gwas: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic acids research*, 38(suppl 2):W90–W95.

[Zhang et al., 2010b] Zhang, L., Pei, Y.-F., Li, J., Papasian, C. J., and Deng, H.-W. (2010b). Efficient utilization of rare variants for detection of disease-related genomic regions. *PloS one*, 5(12):e14288.

[Zhang et al., 2011] Zhang, Q., Irvin, M. R., Arnett, D. K., Province, M. A., and Borecki, I. (2011). A data-driven method for identifying rare variants with heterogeneous trait effects. *Genetic epidemiology*, 35(7):679–685.

[Zhang et al., 2014] Zhang, Y., Xu, Z., Shen, X., and Pan, W. (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage*, 96:309–325.

[Zhong et al., 2010] Zhong, H., Yang, X., Kaplan, L. M., Molony, C., and Schadt, E. E. (2010). Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *The American Journal of Human Genetics*, 86(4):581–591.

[Zhou et al., 2010] Zhou, H., Sehl, M. E., Sinsheimer, J. S., and Lange, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):2375.

[Zhou et al., 2013] Zhou, Y.-H., Barry, W. T., and Wright, F. A. (2013). Empirical pathway analysis, without permutation. *Biostatistics*, 14(3):573–585.

[Zhu et al., 2010] Zhu, X., Feng, T., Li, Y., Lu, Q., and Elston, R. C. (2010). Detecting rare variants for complex traits using family and unrelated data. *Genetic epidemiology*, 34(2):171–187.