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

by

Yang Yang, M.S

APPROVED:

Dissertation Chair, PHD

Minor Advisor, PHD

Breadth Advisor, PHD

External Advisor, PHD

DEDICATION

Persistent support from my family members:

Nainan Hei

&

Tianpeng Yang and Qi Lu

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

by

Yang Yang, M.S

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
November, 2014

ACKNOWLEDGMENTS

Great thanks to my dissertation adviser Dr. Peng Wei, as he guided me ever from 2011, put countless efforts in training me to be a countable 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 by instead overcome the difficulty in a native English style of written and oral communications; 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 writings. I also want to express my special gratitude to Dr. Han Liang. As I have been a Graduate Research Assistant in MD Anderson Cancer Center under his supervision and mentoring between 2012 to 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.

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

Yang Yang, M.S
The University of Texas
School of Public Health, 2014

Dissertation Chair, Peng Wei, PhD

Minor Advisor, Alanna C. Morrison, PhD

Breadth Advisor, Yun-Xin Fu, PhD

External Advisor, Han Liang, PhD

Contents

List of Tables

List of Figures

1 Background

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 [**?**, **?**]. 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 [**?**]. 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 [**?**]. It will become an even more severe problem for rare variants (RVs) analysis. RVs are usually defined as SNVs with MAF below 1% [**?**]. 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 [**?**, **?**, **?**, **?**]. 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 [**?**, **?**, **?**, **?**]. 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 [**?**, **?**, **?**]. 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. [**?**, **?**, **?**, **?**, **?**].

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)[**?**] and the combined multivariate and collapsing (CMC) method [**?**]. 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) [**?**], 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 [**?**, **?**, **?**, **?**]. Such improved "burden tests" includes the sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS) [**?**, **?**]; the replication-based test (RBT) [**?**] 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 [**?**]; the MB test with approximately optimal collapsing (AOC) method [**?**]; a data-driven P-value Weighted Sum Test (PWST) [**?**] 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) [**?**], which is equivalent to a variance-component test; the C-alpha test [**?**], which handles RVs with mixed effect directions well but is not able to adjust for covariates (such as the principal components used to correct for population stratification confounders); the kernel machine regression (KMR) method [**?**, **?**], 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 [**?**]); the widely used sequence kernel association test (SKAT)[**?**], 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 [**?**, **?**], which is a weighted linear combination of a burden test and the SKAT variance component test; the adjusted-SKAT [**?**], 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 [**?**] 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 [**?**], which is a model-free method, collapses only a subset of the variants in a region to achieve the strongest association with a phenotype; the kernel-based adaptive cluster (KBAC) method [**?**], comparing the difference of weighted multi-site genotype frequencies between cases and controls; the rare variant weighted aggregate statistic (RWAS) method [**?**], 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 [**?**, **?**]. 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 [**?**, **?**, **?**, **?**] 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 [**?**] builds a general framework for association testing, which combines strength from MB test and variable-threshold (VT) test [**?**] 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 [**?**]. An exponential combination (EC) framework for set-based association tests [**?**] 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 [**?**]. A unified mixed-effect model [**?**] 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 [**?**] 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 [**?**]. A probabilistic disease-gene finder employs an aggregative variant association test that combines both amino acid substitution and allele frequencies as implemented in VAAST [**?**] and the later improved version VAAST 2 [**?**]. The weighted score test [**?**] 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 [**?**]. Another comprehensive review can be found in [**?**]. 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 [**?**].

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 [**?**, **?**, **?**, **?**]. 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, [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]). 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 [**?**]. 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 [**?**]. 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.

1.2 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

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

[**?**]. 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 de-

pends 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 [**?**].

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 [**?**, **?**, **?**, **?**, **?**]. 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 [**?**, **?**, **?**, **?**]. Compared with large number of GWASs, only very few studies involved longitudinal data analysis. One such study on smoking and nicotine dependence by [**?**] 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 [**?**, **?**, **?**]. 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 [**?**, **?**]. 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 [**?**, **?**]. 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 [**?**]. 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 [**?**].

Another major method, the marginal model with GEE was first proposed in 1986 [**?**, **?**]. It is an extension of the quasi-likelihood methods by Wedderburn [**?**]. 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 [**?**], logit link [**?**, **?**, **?**] and log link [**?**]. Model fitting is straightforward for linear link, as in Gaussian autoregressive models, the full maximum likelihood estimation is available [**?**]. 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 [**?**, **?**, **?**, **?**].

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 [**?**, **?**]. The mixed-effect model could handle such inference of interests [**?**]. 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 [**?**, **?**, **?**]. 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 [**?**, **?**]. 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 [**?**, **?**]. 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 [**?**, **?**]. 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.

1.3 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 [?, ?, ?, ?]. Complex diseases are known to be influenced by a combination of genetic factors in addition to environmental factors, lifestyle factors, and their interactions [?, ?]. 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 [?, ?]. Other types include Gene Ontology terms, protein-protein interaction, canonical genetic signaling pathways, gene expression networks as examples [?, ?, ?, ?]. 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) [?], BioCarta [?] and Gene Ontology [?]. KEGG and BioCarta provide manually curated pathways in different biological processes, whereas Gene Ontology mainly contains computational annotations for human genes. Several commercialized 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 [?] and Nature Pathway Interaction Database [?], both of which manually curated the cell signaling pathways; the MetaCyc database [?] and BioCyc database [?], 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 [?, ?, ?, ?, ?, ?]. 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 [?]. Representatives of competitive approach are gene set enrichment analysis (GSEA) [?], which is based on the Kolmogorov-Smirnov test and DAVID [?], 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 [**?**].

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 [**?**], the metabochip [**?**] and the CVD35/cardiovascular-IBC-array [**?**, **?**]. The self-contained approaches have also been reported to be more powerful than the competitive approaches [**?**]. 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 [**?**, **?**, **?**].

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.

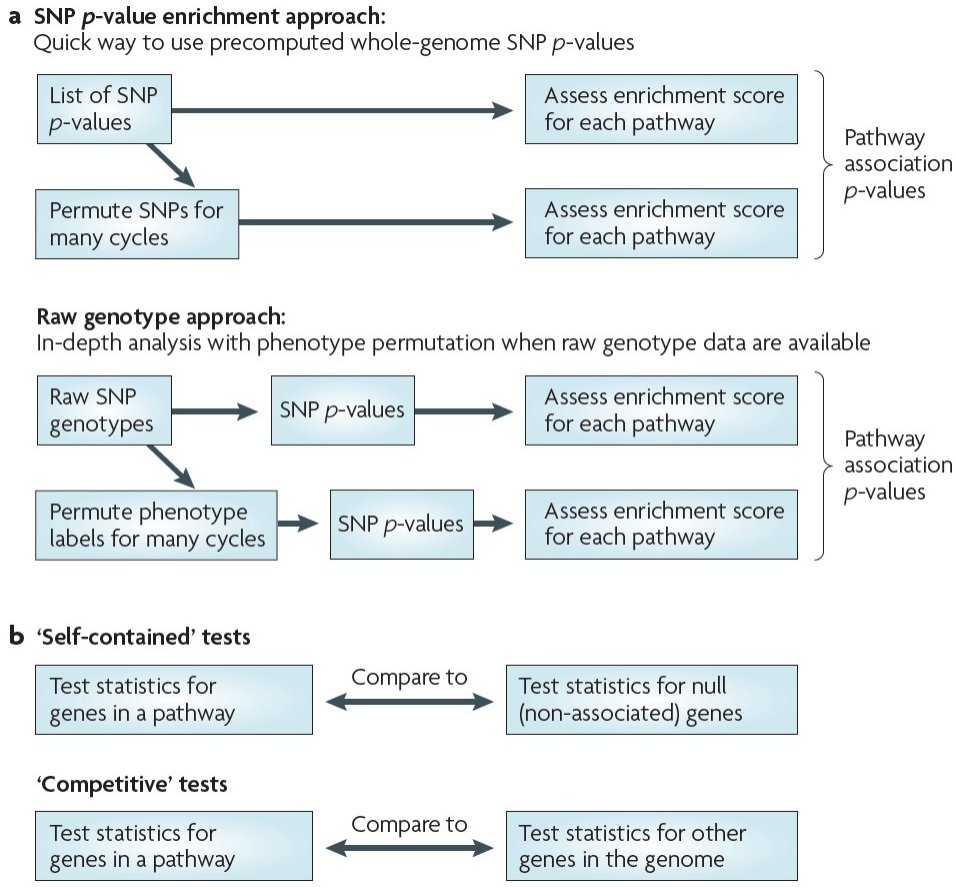

Figure 2: **Types of pathway association tests in GWAS.** (a). Categorization based on data input type; (b). Categorization based on hypothesis testing. This figure is adopted from Wang et al (2010) [**?**].

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 [**?**]. The GSEA-SNP, a modification of Wang et al's GSEA [**?**], uses the max-

test and all SNPs in a gene [**?**]. 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[**?**]. 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 [**?**]. 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 [**?**]. 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 [**?**]. 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 [**?**]. PLINK, a widely used software package in GWAS data analysis, provides an option to execute gene-set-level association analysis [**?**]. 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 [**?**]. The SNP ratio test (SRT) method, tests the ratio of significant SNPs in a pathway and computes the empirical p-value based on permutation [**?**]. 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 [**?**]. 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) [**?**]. 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 [**?**]. 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 [**?**]. 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 [**?**]. An adaptive rank truncated product (ARTP) statistic, 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 [**?**]. Detailed reviews about these and other pathway-based association tests can be found in [**?**, **?**, **?**].

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 [**?**, **?**, **?**, **?**, **?**]. The development of complex diseases involves genetic factors, environmental factors, behavior factors, and the interactions among them. In public health research, identification of the casual 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 [**?**]. These genetic loci have been validated by some validation procedures, such as replicate studies, meta analysis and wet lab experiments [**?**, **?**, **?**, **?**, **?**].

The advent of the Next-Generation Sequencing (NGS) technique has brought human genetics to a new era [**?**, **?**, **?**, **?**], and has the potential to explain some of the missing heritability via disease/trait-associated rare variants [**?**]. 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 my dissertation, I will develop statistical methods that will provide researchers with more powerful and robust data-adaptive association tests for either common variants (CVs) or rare variants (RVs) in the longitudinal data settings. Both gene-based and pathway-based test strategies will be implemented. Furthermore, an R package or independent Linux command-line based software implementing the methods will be released as part of the methodology development, which will greatly facilitate the research community to use the new methods in 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.

3 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).

4 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 [**?**] 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 [**?**, **?**]. 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.

5 Methods

5.1 Overall Study Design

5.1.1 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 [**?**, **?**, **?**, **?**, **?**] 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.

5.1.2 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 atherosclerosis 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 [**?**].

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 [**?**] 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 5.6.

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

5.2 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

5.2.1 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 [**?**]:

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

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:

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

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

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 $U_{\cdot 1}$ and $U_{\cdot 2}$ 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 [**?**], 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 [**?**].

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 3.

- **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 [**?**, **?**].

- **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 [**?**, **?**].

- **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 [**?**, **?**]; alternatively, a simulation based method relying on the asymptotic normal distribution of the score vector can be used to calculate its p-value [**?**, **?**]. 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 statis-

tics: $T^{(b)} = \max_{j=1,\ldots,p} \frac{U^2_{(b).j}}{\Sigma_{.2,jj}}$, 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 [**?**] respectively. The former is closely related to other burden tests such like those in [**?**, **?**, **?**]. 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) [**?**, **?**, **?**, **?**, **?**] and C-alpha test [**?**], 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 [**?**]. Different researchers proposed different weighting schemes to pool the information of all SNVs in a region of interest, such as those used in [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. 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 [**?**, **?**, **?**].

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 [**?**]. 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 [**?**], 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 $T_{SPU(\gamma)}$ [**?**, **?**]. 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 3). 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 [**?**], linear combine, quadratic combine and fisher's combine methods [**?**, **?**, **?**]. 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

$T = U'_{.2}\Sigma^{-1}_{.2}U^{-1}_{.2}$ will keep the covariance matrix in the denominator, which preserves the information of possible linkage disequilibrium among SNPs. To combine the strengths 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.

5.2.2 Methods for Simulation Settings

I will generate the simulated genotypes following [**?**, **?**, **?**]. 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 [**?**]. 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 [**?**, **?**, **?**, **?**, **?**] 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 coalescence 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 [**?**, **?**]. 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 casual SNPs. Such advantages were proved and utilized in developing new association tests in a number of past researches [**?**, **?**, **?**, **?**, **?**, **?**, **?**].

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 $\text{Prob}(G_{ij} = 1) =$ MAF of jth 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 casual SNPs from the later test procedure. Therefore, in the first block, only null SNPs in LD with these 5 casual 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 [**?**]. 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 5.2.1; $\mu_i = Z_i\varphi + X_i\beta = H_i\theta$ as in quantitative trait case (see section 5.2.1); 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:

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

It'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 (6), 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 [**?**]. 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 (5)), 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 (7):

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

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 (7) and (9), 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.

5.2.3 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 casual 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, casual 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 symmetry, and unstructured. Note that, as I simulated the longitudinal trait using the AR(1) correlation structure as described in Section 5.2.2, 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.

5.3 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

5.3.1 Statistical Modeling

In the previous section 5.2.1, 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.

5.3.2 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 5.2.2), except that:

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

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

5.3.3 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.

5.4 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.

5.4.1 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 5.2.1, 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-casual-gene situation (for example, only 2 or 3 genes are associated with the longitudinal trait in a pathway with 20 genes).

5.4.2 Methods for Simulation Settings

The simulation strategy of RVs within a gene is the the same as section 5.3.2. 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 casual gene containing $U(1, 3)$ causal RVs.

5.4.3 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 5.2.3).

Optionally, I may consider more extensive simulation studies, such as tuning the casual RV number within a casual gene or tuning the number of casual 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.

5.5 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, "Executable.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.

5.6 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 [**?**].

For genotype data part, in Aim 1(a), I will use the common variants genotyped by the

ExomeChip [?] 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 [?, ?]. In Aim 2, I will define the gene pathway by public pathway resources like KEGG [?] and BioCarta [?]. 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 [?] 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 [?] 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 by ANNOVAR [?], a software package providing functional annotation of genetic variants from high-throughput sequencing data or array data. 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.