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, and hundreds of GWASs have been published already (A Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/26525384). The most popular approach in GWAS is to test the association with complex traits on single nucleotide polymorphism (SNP), also known as single nucleotide variant (SNV), one by one, then select the the SNVs that meet a stringent significance level after multiple testing correction, such as the Bonferroni and false discovery rate (FDR) methods [McCarthy et al., 2008, Hirschhorn and Daly, 2005]. However, this strategy will suffer from low power when the minor allele frequency (MAF) of the SNV is low (between 1% and 5%), and as a result the signals contained within the low MAF SNVs are hard to detect [Sham and Purcell, 2014]. In addition, the usual regression coefficient estimate of SNV becomes unstable due to the small number of minor allele counts and the coefficient estimate's variance becomes very large [Sham and Purcell, 2014]. It becomes an even more severe a problem for rare variants (RVs), usually defined as SNVs with MAF below 1% [Bansal et al., 2010]. In spite of extremely low MAF, we cannot underestimate RVs' important role in conferring disease risk. Due to the constraint of purifying selection, causal and functionally deleterious variants are often RVs, which typically have larger effect size than common variants [Fu et al., 2013, Bansal et al., 2010, Sham and Purcell, 2014, McCarthy et al., 2008]. Therefore, developing new association tests tailored to SNVs with low MAF and RVs has been a very active research area in recent years. Due to the nature of low MAF, either increasing case sample size or aggregating information across multiple variants in an analysis set (e.g. gene) is expected to achieve a practically acceptable power [Capanu et al., 2011, Basu and Pan, 2011, Bansal et al., 2010, Sham and Purcell, 2014]. As increase sample size is usually expensive and demanding, e.g. more than 25,000 cases will be required, advances in RV set-based association tests, e.g. based on genes or biological pathways, are major directions researchers have been

investigating towards [Ye and Engelman, 2011, Pinto et al., 2010, Sham and Purcell, 2014]. Sets of genes can be defined by, e.g. Gene Ontology terms, protein-protein interactions, canonical gene signal pathways, gene expression networks, etc [Sham and Purcell, 2014, De la Cruz et al., 2010, Weng et al., 2011, Wang et al., 2010a, Wei et al., 2012].

1.1 Gene-based association tests

[To Dr. Wei's concern: make nice introduction of these tests by itemizing the categories]
A large number of gene-based association tests (mainly designed for RVs) have been proposed in recent years. From the earliest few methods, including the cohort allelic sums test (CAST)[Morgenthaler and Thilly, 2007] and the combined multivariate and collapsing (CMC) method [Li and Leal, 2008], to later on a full bucket of methods. These methods can be by large categorized into:

- A very famous category of these methods is the so-called "burden test" or "sum test", such as a weighted sum statistic (WSS) [Madsen and Browning, 2009], which uses MAF based weighting scheme to combine the sum 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 "burden test" category inherited and improved the WSS performance in some scenarios [Hoffmann et al., 2010, Zhang et al., 2010b, Ionita-Laza et al., 2011, Feng et al., 2011]. Such improved "burden tests" include the sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS) [Zhu et al., 2010, Feng et al., 2011], the replication-based test (RBT) [Ionita-Laza et al., 2011] built on WSS with the aim to be less sensitive to the presence of both risk and protective effects in a genetic region of interest, the yet another weighted-sum test with a "step-up" approach to choose the 'best' combination of rare variants into a single aggregated group [Hoffmann et al., 2010], the MB test with approximately optimal collapsing (AOC) method [Zhang et al., 2010b], a data-driven P-value Weighted Sum

Test (PWST) [Zhang et al., 2011] which used both significance and direction of individual variant effect from single-variant analysis to calculate a single weighted sum score, etc.

- 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. This type of tests includes the Sum of Squared U-statistics test (SSU) [Pan, 2009] which is close to a variance-component test, the C-alpha test [Neale et al., 2011] which handles RVs with mixed effect direction well but not able to adjust for covariates (such as population stratification PCs), the kernal machine regression (KMR) method [Wu et al., 2010, Kwee et al., 2008] which provides the flexibility of choosing different kernal functions $h(.)$ to measure the genomic similarity between the genotypes of subject i and j then regress response on the specified kernal functions (given linear kernal, it is equivalent to SSU test), the very famous sequence kernel association test (SKAT)[Wu et al., 2011], the SKAT-O [Lee et al., 2012b, Lee et al., 2012a] which is a weighted linear combination of a burden test and the SKAT variance component test of $\tau^2 = 0$, the adjusted-SKAT [Oualkacha et al., 2013] which allows the variant effects to have an equal correlation ρ besides the usual assumption in SKAT that the effect of variants are assumed to be independently and identically distributed with an arbitrary distribution of mean 0 and variance τ^2, the GEE-based linear kernel machine SNP set association test [Wang et al., 2013] which is very closed to the SSU test, etc.

- "Collapsing-based test" inherited the idea from CMC/CAST method, and this type of test is actually closely related to "sum test". The RARECOVER algorithm [Bhatia et al., 2010], 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) [Liu and Leal, 2010], compares the difference of weighted multi-site genotype

frequencies between cases and controls. The rare variant weighted aggregate statistic (RWAS)[Sul et al., 2011], groups rare variants and computes a weighted sum of differences between case and control mutation counts.

- Lasso and group-penalized regression based methods [Zhou et al., 2010, Kim et al., 2014]

- Functional linear model and (smoothed) functional principle component analysis based association tests [Luo et al., 2011, Luo et al., 2012b, Luo et al., 2012a, Fan et al., 2013]

- "Adaptive/Hybrid test" combines the testing advantages from at least two major categories above to make the test more data adaptive with regard to maintain a large power. The EREC method [Lin and Tang, 2011] builds a general framework for association testing which combined strength from MB test and VT test to form the most powerful test by setting the weight function ϵ proportional to the set of regression coefficients β in the limit. A data adaptive tests combines the score test, SSU and Sum tests' advantages [Pan and Shen, 2011]. An exponential combination (EC) framework for set-based association tests [Chen et al., 2012] features with the sum of exponential statistics (statistics should follow either independent normal or independent chi-square distribution) are parametric and have the adapted standardized variant statistics from previous MB test and C-alpha test. A robust and powerful test uses Fisher's method to combine linear and quadratic statistics [Derkach et al., 2013]. A unified mixed-effect model tests both group effect equal to 0 and variance component equal to 0, which 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 [Sun et al., 2013].

- Other miscellaneous tests. Some of them can be classified into more than one category mentioned above, thus we include them here as well as other miscellaneous tests. A variable-threshold (VT test) approach [Price et al., 2010] computes z-score $z(T)$ for each different MAF threshold T, defines z_{max} as the maximum z-score across values

of $z(T)$, and finally assesses the statistical significance of z_{max} by permutations on phenotypes. A data-adaptive sum test (aSum) is capable of handling both deleterious and protective direction and allowing collapsing CVs into the test [Han and Pan, 2010]. A probabilistic disease-gene finder employs an aggregative variant association test that combines both amino acid substitution and allele frequencies as implemented in VAAST [Yandell et al., 2011] and later improved in VAAST 2 [Hu et al., 2013]. The weighted score test [Cai et al., 2012] up- or down-weights the contribution from each member of the marker-set based on the Z-scores of their effects.

For a detailed comparison and discussion among some of the above mentioned tests, Basu and Pan have done a very comprehensive review and simulation-based benchmark on these tests [Basu and Pan, 2011]. Another comprehensive review on statistical analysis strategies for association studies involving rare variants was written in 2010 [Bansal et al., 2010]. Recently Pan et al also did a performance benchmark of several latest methods including PWST, EREC, aSSU, SKAT-O and their newly proposed aSPU method [Pan et al., 2014].

Due to the complexity in genetics association with phenotype, e.g. specific association effect directions and sizes, a given test favoring one scenario may or may not perform well in other scenarios [Pan, 2009, Derkach et al., 2013, Pan et al., 2014, Sun et al., 2013]. In other words, there is no single test the most powerful among all testing scenarios. Therefore, there has been a lot of efforts already made in developing adaptive/hybrid tests for RVs (e.g., [Derkach et al., 2013, Chen et al., 2012, Han and Pan, 2010, Lee et al., 2012a, Lin and Tang, 2011, Pan and Shen, 2011, Sun et al., 2013, Zhang et al., 2011]). However, due to still limited adaptability, e.g. with a fixed set or pre-determined weights on individual RVs, these tests though combined some earlier tests' advantages (e.g. MB test, burden test and SKAT), they 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 Wei Pan and Peng Wei [Pan et al., 2014]. It features as having the ability to achieve

quasi-optimal power in all data scenarios, such as varying number of SNVs within the region, varying ratio of signal SNVs, same effect alleles or a mixed effect of both protective and deleterious alleles, varying allele frequencies, varying effect size, etc. It maintains the most power as compared to other state-of-art tests when a large number of RVs within a region contains a small portion of signals, which is usually the case in association studies under exome/whole-genome sequencing scenario [Pan et al., 2014].

1.2 Longitudinal study design in GWAS and the strategy

[To all advisors' concern: explain the longitudinal study in more details, including where the larger power comes from and why result from longitudinal study may differ from cross-sectional study]

Comparison between longitudinal studies and cross-sectional studies

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

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

[WARE et al., 1990]. Now β_C still represents the 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 basic of inference about β_C is a comparison of individuals with a particular value of x to other individuals with a different value of x. In contrast, the parameter estimation of β_L is by comparing a person's responses at two times, assuming x changes with time.

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

1. Longitudinal studies allow us to estimate both the cross-sectional difference (β_C) and the rate change over time (β_L).

2. Even when $\beta_C = \beta_L$, longitudinal studies tend to be more powerful than cross-sectional studies. This is due to the fact that in longitudinal studies, each person can be thought of serving as his/her own control. 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 things 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 (smaller variance).

3. 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 people. 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 people is large, we might prefer to use more data for the same individual across time.

4. With longitudinal studies, we can estimate a person's current and future outcome.

The merit 2 scenario above is also the case in my dissertation. My dissertation has the underlying assumption that the SNPs in a region contribute to the outcome Y as the main effect only and the effect is the same across time ($\beta_C = \beta_L$) for a given individual. There is more to explain here about the efficiency of the longitudinal study. Let $e = Var(\hat{\beta}_L)/Var(\hat{\beta}_C)$ as the specific measure of efficiency. Apparently, the smaller the value of e the greater is

the information gained by taking additional measurements across time on each person. The value of e depends on a list of factors, which includes correlation structure R (e.g 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 visit 1. In general, increasing n_i (e.g. more measurements) and increasing δ (e.g. 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 page 24-27 in [Diggle et al., 2002].

The identified significant signal loci from a longitudinal study may be the **same** or **different** from a comparable cross-sectional study depending on a specific hypothesis test. In 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 of the main effect and interaction effect equal to 0 or individual testing of any one of the two effects equates 0 will possibly lead to different significant loci from the corresponding cross-sectional study.

The relationship between the testing power and affecting factors in longitudinal studies

In any study, investigators must provide the following quantities to determine the power P, which include Type I error rate (α), smallest meaningful difference (d) to be detected, sample size (n), variance (σ^2) in response variable. While in longitudinal studies, there are several additional factors necessary to consider, which are the number of repeated observations per person (n_i) and the correlation among the repeated observations within the same person (ρ). Let we briefly illuminate the relationship between these quantities and the power P. Increase α will increase P; increase d will increase P; increase n will increase P; reduce σ^2 will increase P; increase n_i will increase P. For ρ, the relationship with P is not fixed but depending on

what kind of hypothesis under testing. When in $\beta_C = \beta_L$ scenario, an **decreasing** ρ will lead to a larger power. In contrast, when in $\beta_C \neq \beta_L$ scenario, an **increasing** ρ will lead to a larger power while testing the $\beta_L = 0$ (the rate change over time equal to 0). This at the first glance seems counter-intuitive but indeed reasonable. In $\beta_C = \beta_L$ 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 lead to an effectively larger sample size (within-subject measurements are more distinct), which leads to a smaller variance of β_C estimate. On the contrary, in $\beta_C \neq \beta_L$ scenario and 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's whose variance is decreasing in ρ, thus an increasing ρ will lead to a larger power of testing the significant deviation from $\beta_L = 0$.

Longitudinal studies in GWAS

While many GWASs have been performed in cohorts, they collected data across multiple time points for each individual [Aulchenko et al., 2009, Ionita-Laza et al., 2007, Kamatani et al., 2010, Kathiresan et al., 2007, Sabatti et al., 2008]. However, the longitudinal information has not been fully utilized as the majority of current association tests only used either the baseline measurement or average measurement for each individual[Sabatti et al., 2008, Ionita-Laza et al., 2007, Kamatani et al., 2010, Kathiresan et al., 2007]. Compared to the total number of GWASs, very few studies involved longitudinal data analysis. One such study on smoking and nicotine dependence by [Belsky et al., 2013] have data from a 4-decade longitudinal study, and they used generalized estimating equation model to analyze the panel data account for correlation within subject. There are also several studies on Alzheimer's Disease (or more specifically ADNI-1 data collected by Alzheimer's Disease Neuroimaging Initiative) involving the analyses of longitudinal phenotypic information collected at multiple time points [Wang et al., 2012, Melville et al., 2012, Silver et al., 2012]. Increased power coming from longitudinal data has been elucidated herein before, and recently this fact has been discussed in depth by either simulation study and/or real data analysis in GWAS settings

[Xu et al., 2014, Furlotte et al., 2012]. Depending on specific parameters settings in simulation studies and case by case for real data analysis, the power gain from longitudinal data analysis as compared to baseline data analysis can range from a moderate to a significant amount. [Xu et al., 2014, Furlotte et al., 2012]. Therefore, a longitudinal study design or executing the longitudinal data analysis when data are available is always preferred in GWAS settings.

Classical longitudinal data analysis methods

Existing methods in longitudinal data analysis can be mainly categorized into three categories: 1, mixed effect models; 2, marginal models with regression coefficient estimated by generalized estimating equation (GEE); 3, transition (Markov) models. Mixed effect model was first proposed in 1982 [Laird and Ware, 1982]. Mixed effect model is a two-stage models, which treat probability distributions for the response vectors of different individuals as a single family and the random-effects parameters which hold the same for the same individual as another distribution. Parameter estimation is usually done by restricted maximum likelihood (REML) and expectation-maximization (EM) algorithm [Laird and Ware, 1982]. Another major method, the marginal models with GEE were first proposed in 1986 [Zeger and Liang, 1986, Liang and Zeger, 1986]. It is an extension to quasi-likelihood methods by Wedderburn [Wedderburn, 1974]. Rather than giving subject-specific(SS) estimates as in mixed effect models, GEE gives population-averaged (PA) estimates by only describing the marginal expectation of the outcome variable as a function of the covariates and the variance is a known function of the mean, while accounting for the correlation among the repeated observations for a given subject by specifying a "working" correlation matrix, which may not be the true underlying correlation matrix. The generalized estimating equations are thus derived without specifying the joint likelihood function of a subject's observations as SS model does need. 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 do-

ing iteratively reweighted linear regression. The last major method, transitional (Markov) models, 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 mixed models and marginal models can do. A few examples are linear link [Tsay, 1984], logit link [Cox and Snell, 1989, Zeger et al., 1985, Korn and Whittemore, 1979] and log link [Zeger and Qaqish, 1988]. Model fitting is straightforward for linear link as in Gaussian autoregressive models, the full maximum likelihood estimation is available [Tsay, 1984]. For logistic and log-linear cases, the full likelihood is unavailable and the alternative is to maximize the conditional likelihood with GEE-like iterative weighted least square algorithm to solve the conditional score function and get consistent estimates [Cox and Snell, 1989, Zeger et al., 1985, Korn and Whittemore, 1979, Zeger and Qaqish, 1988].

There is a need to discuss more on two out of the three major methods, which are mixed models and marginal models (since transitional models are not popularly used in genetics association study settings, we will omit further discussion about it), as it explains the reason why we will develop our new method within GEE framework for specific aims hereinafter. Application of GEE may be less appropriate when the time course of the response variable for each individual, e.g. BMI measurements across several time points, is of primary interest, so as to the correlation parameters within same subject [Zeger et al., 1988, Liang and Zeger, 1986]. The mixed effect model could handle such interests [Laird and Ware, 1982]. However, under the genetic association study settings, time course and/or within-subject correlation parameters are usually not of major interests (i.e. can be put as nuisance parameters). The true substantial problem is for gene or region based multiple-SNV-set association test, increased number of explanatory variables (SNVs) on the RHS of the regression-like equation will lead to large consumption of the degree of freedoms (dfs) and algorithm convergence difficulty. Large consumption of the dfs will lead to power loss and possibly inflate the type I error, e.g.

excessive inflation in Wald Test [Guo et al., 2005, Pan, 2001, Shete et al., 2004]; algorithm convergence difficulty is very often encountered in mixed model when equation RHS has a lot of covariates and for some extreme scenario, e.g. with a binary trait, the MLE of a regression coefficient of a RV does not exist if the minor alleles of this RV only appear in case or vice versa, eventually it turns out to convergence failure with an iterative algorithm to obtain MLE [Zhang et al., 2014, Pan et al., 2014]. Another caveat of the mixed model under this test setting is, mis-specification of the random-effects distribution and/or omitting part of the random-effects (e.g. keep only random intercept in the mixed model when random slope is significant) will lead to excessive type I error inflation [Litière et al., 2007, Xu et al., 2014]. Compared with mixed models, these problems are much more mild on GEE models: GEE Score test is proved to be robust to type I error inflation when equation RHS has a lot of covariates; upon usage of so-called sandwich or robust covariance matrix, GEE model estimator will keep consistent and type I error will keep at the nominal level even when the working correlation is misspecified (comparable to misspecified random effect in mixed models); GEE model fitting requires only evaluation under null hypothesis, which greatly simplifies the convergence burden and accelerates the computation; with regard to power loss in the case of increased number of covariates (SNVs) put on the equation RHS, as aforementioned, a recent work on data adaptive association test within GEE framework demonstrated convincing capability in maintaining a still high power while many other tests' power dropped dramatically [Zhang et al., 2014, Pan et al., 2014]. Though this work is for single cross-sectional trait or multiple cross-sectional traits, it can be extended to longitudinal scenario as in our aim I hereinafter.

1.3 Gene-Set/Pathway based association test

Extending the gene-based association test to sets of multiple related genes could return more biological meaningful inference, as in vivo, there are usually multiple genes work-

ing together to fulfill a biological function, analyzing "co-workers" genes together with phenotype tends to identify those signals hidden from or attenuated in single-gene based tests [for Blood Pressure Genome-Wide Association Studies et al., 2011, Hirschhorn, 2009, Zhong et al., 2010, Wang et al., 2010a]. Complex disease are known to have a combination of genetic factors in addition to environmental, lifestyle factors, and their interactions [Hirschhorn and Daly, 2005, McCarthy et al., 2008]. Thus by investigating into the sets of genes, more evidence could be extracted as risk altering factors contributing to a specific disease. The other reason for considering pathway-based association test is similar to the consideration for gene-based association test: aggregating multiple Genes/RVs against testing each Gene/RV separately will boost the statistical power. One convincing evidence is from The Cancer Genome Atlas (TCGA: http://cancergenome.nih.gov/) for doing tumor sequencing studies. While only few oncogenes (e.g. TP53, EGFR) harbor many mutations, most others harbor few mutations in a tumor-dependent manner. Single gene-based association test thus 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 disease such as cancer to investigate its association with somatic mutations, a gene-set/pathway analysis by aggregating information across genes will boost the statistical power, and is thus preferred.

Among association tests on sets of functional related genes, gene pathway based association test is probably the most popular one [De la Cruz et al., 2010, Wang et al., 2010a] with others like Gene Ontology terms, protein-protein interaction, canonical gene signal pathways, gene expression networks, etc [Sham and Purcell, 2014, De la Cruz et al., 2010, Weng et al., 2011, Wang et al., 2010a] . The 'pathway' in GWAS usually means a set of co-working genes tightly related. Some commonly used public pathway databases/repositories include Kyoto Encyclopedia of Genes and Genomes (KEGG) [Ogata et al., 1999], BioCarta [Nishimura, 2001] and Gene Ontology [Ashburner et al., 2000]. KEGG and BioCarta provide manually curated pathways in different biological processes, whereas Gene Ontology mainly

contains computational annotations for human genes. Several commercialized databases are also available including Ingenuity Pathway Analysis (IPA) and MetaCore from GeneGo, whose contents combines the manually curated evidence, literature review and algorithm predicted result. There are also kinds of specialized pathway database which curate specific types of pathways, such as Science Signal Transduction Knowledge Environment and Nature Pathway Interaction Database, which both manually curated the cell signaling pathways; the MetaCyc database contains metabolic pathways. We will skip the enumeration of all such databases here.

Types of pathway based association testing methods

With regard to null hypothesis being tested, pathway based association testing methods can be categorized into two major types: self-contained approach and competitive approach [Goeman and Bühlmann, 2007, Liu et al., 2007, Nam and Kim, 2008, Wang et al., 2010a, Fridley et al., 20 Fridley and Biernacka, 2011]. self-contained (a.k.a. Constrained) approach hypothesizes there is no gene in the gene set associated with the phenotype, while competitive approach hypothesizes the same level of association of a gene set with the given phenotype as the complement of the gene set. Competitive methods are usually started with identifying SNPs/-genes that are significantly associated with a phenotype, and then evaluating whether the significantly associated SNPs tend to enrich in predefined gene-set/pathway. These methods are called 'competitive' because they compare the frequency of significantly associated SNPs in a particular set of genes/pathway with the frequency of significant associations among all genes not in the set [Fridley and Biernacka, 2011]. Representatives of competitive approach are gene set enrichment analysis (GSEA) [Subramanian et al., 2005], which uses a Kolmogorov-Smirnov test and DAVID [Dennis Jr et al., 2003], which uses a Fisher's exact test. In contrast, self-contained approach considers the null hypothesis that SNPs/genes in the gene-set of interest are not associated with the trait vs. alternative hypothesis that SNPs/genes in the gene-set are associated with the trait. Methods within Self-contained approach are more flexible, they can be assessing the deviation from the expected num-

ber of significant SNPs under null hypothesis of no association of the phenotype with the gene-set/pathway; or computing association p-values for each marker in a gene-set/pathway, followed by testing whether the deviation between observed distribution of the marker-level p-values and expected distribution under null hypothesis is significant; or modeling the effect of gene by aggregating multiple markers, followed by modeling the effect of gene-set by aggregating multiple relevant genes; or directly modeling the effect of gene-set by aggregating the markers within the gene-set, skipping gene-level statistics. There is an illustration of the hypothesis testing difference between the two types of approaches as shown in Figure 1.

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 demo** (A). Example of competitive approach; (b). Example of self-contained approach. Figure is adopted from Fridley et al 2010 [Fridley and Biernacka, 2011].

One limitation of competitive approach is they cannot be applied to studies of candidate gene-sets for which only SNPs in the candidate gene-sets have been genotyped but not complemented ones. The reason is straightforward: competitive approach requires a comparison between many different pathways. On the other hand, self-contained approach

requires only genotypes from a collection of candidate genes, which in turn brings over the benefits such as it can be used for genome-wide, specific disease gene group (such like the cardiovascular disease, metabolic traits and autoimmune disease; usually with the disease-specific genotyping platform support, e.g. ImmunoChip [Cortes and Brown, 2011], metabochip [Voight et al., 2012] and CVD35/cardiovascular-IBC-array [Cheng et al., 1999, Keating et al., 2008].), candidate gene, or pathway studies. Self-contained method has also been reported to be more powerful than competitive methods [Goeman and Bühlmann, 2007], which follows immediately from the fact that the self-contained null is more restrictive than the competitive null, as noted herein 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 has been reported such as the genomic inflation of test statistics is often observed or not adequately adjusted for and it leads to inflated type I error [Wang et al., 2007, Goeman and Bühlmann, 2007, Fridley et al., 2010]; the "single-gene-pitfall", which means when a gene-set/pathway contains only one single gene, the self-contained test will generalize the single gene testing to be equivalent to gene set testing, in the sense that the two procedures are completely equivalent for singleton gene sets, while competitive approach treats them quite differently [Goeman and Bühlmann, 2007].

Additionally, based on input data type, the tests can be broadly classified into two categories: those require raw genotypes and those 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 for associated signals. '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-marker tests which requires individual genotype data to derive gene-level test statistics (some of these methods pool all SNPs in a pathway together without calculating test statistics for

pathway gene members) and the method based on single-marker p-values but require raw genotype data for phenotype permutation based test to come up with a more unbiased pathway enrichment score. The 'raw genotype approach' is also more unbiased, e.g. gene length, the distance threshold to assign SNPs to nearby genes, the way to summarize gene-level test statistics, etc. The graphic demo of method categorization is shown in Figure 2.

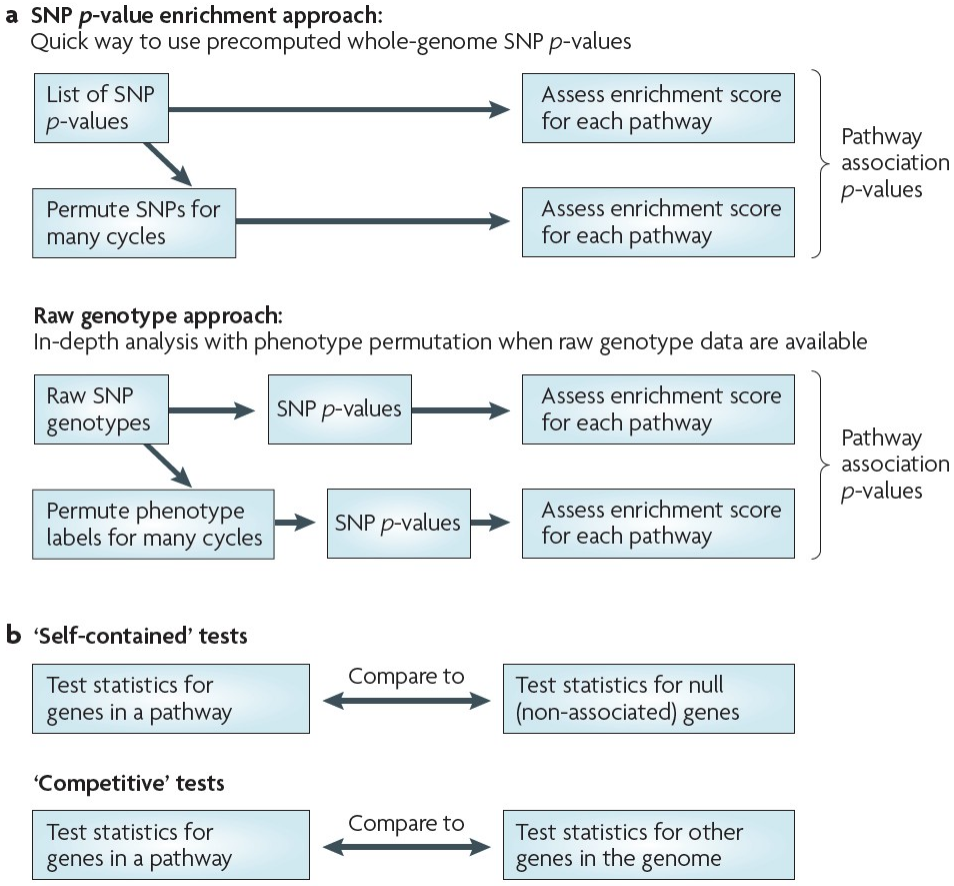

Figure 2: **Types of pathway association tests in GWAS.** (a). Categorization based on data input type; (b). Categorization based on hypothesis testing. Figure is adopted from Wang et al 2010 [Wang et al., 2010a].

Existing pathway based association testing methods

There are already many existed methods in pathway association testing which includes: the earliest one should be the modified gene-set enrichment analysis (GSEA) algorithm to adapt for pathway-based analysis of GWA data, which calculated a weighted Kolmogorov-

Smirnov-like running-sum statistics and used permutation-based procedure to come up with statistical significance [Wang et al., 2007]; GSEA-SNP: modification of Wang et al GSEA [Wang et al., 2007] using max-test and all SNPs in a gene [Holden et al., 2008]; i-GSEA4GWAS: SNP label permutation, assign SNPs to genes, calculate modified GSEA enrichment score[Zhang et al., 2010 the Gene Set Analysis (GSA-SNP): gene-level test based on SNP with minimum P-value (or second best), followed by gene-set-level test using either a Z-test statistic, maxmean test statistic, or GSEA [Nam et al., 2010]; the Gene Set-based Analysis of Polymorphisms (GeSBAP), calculates enrichment score using ranked gene list, assigns the best SNP p-value to a gene, uses Fisher's exact test for gene-set association [Medina et al., 2009]; modification of Fisher's method for combing SNP P-values for gene-level or gene-set-level association [De la Cruz et al., 2010]; gene set ridge regression in association studies (GRASS), which executes lasso regression (L1-norm) of eigenSNPs within each gene to achieve variable selection, while performing ridge regression (L2-norm) of eigenSNPs at the gene-set-level to achieve gene effect (e.g. disease risk odds) estimates shrinkage simultaneously [Chen et al., 2010a]; PLINK, a very famous software in GWA data analysis, provides an option to execute gene-set-level association analysis [Purcell et al., 2007]; association list go annotator (ALIGATOR), which is a 'p-value enrichment approach' requiring only pre-computed SNP p-values, uses Fisher's exact test on SNP with minimum p-value for the gene-level association. It claims it can correct for linkage disequilibrium (LD) between SNPs, various gene size, and multiple testing of nonindependent pathways [Holmans et al., 2009]; the SNP ratio test (SRT), tests the ratio of significant SNPs in a pathway and compute the empirical p-value based on permutation [O'Dushlaine et al., 2009]; supervised principal component analysis with a Gumbel extreme value mixture distribution as test statistic distribution and simulation-based standardization procedure for pathway size [Chen et al., 2010b]; Prioritizing Risk Pathways fusing SNPs and pathways (PRP): gene-level association test based on max risk statistic followed by mean risk approach to get gene-set-level risk statistic, then this statistic is weighted by specific pathway degree (i.e. total edges in a pathway) and standardized to zero

dimension and comparable for prioritization purpose [Chen et al., 2009]; Three statistics to combine a set of dependent p-values of SNPs into an overall significance level for a gene and 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, Quadratic test (QT) asymptotically following central Chi-square distribution, and decorrelateion test (DT) which combines decorrelated individual statistics by Fisher's combination test and asymptotically follows a central Chi-square distribution, under null hypothesis respectively [Luo et al., 2010]; Four combination methods of combing a list of SNP p-values or gene-level pvalues with the assumption that individual markers/-genes are independent: Fisher's, Sidak's, Simes' and the FDR method [Peng et al., 2009]; The Gene-loci Set Analysis (GLOSSI), at first uses the Cochran-Armitage trend test at single-marker level assuming an additive SNP effect, then uses Fisher's combination test to combine individual p-values of markers and corrected by Brown's approximation to better control type I error [Chai et al., 2009]; an adaptive rank truncated product (ARTP) statistic and permutation-based p-value adjustment to combine marker-level p-values to derive gene-level significance level and/or combine gene-level p-values to derive pathway-level significance level [Yu et al., 2009]. Detailed reviews about such pathway-level association tests can be found in [Tintle et al., 2011, Wang et al., 2010a, Fridley and Biernacka, 2011].

2 Public Health Significance

The majority of human diseases are complex diseases, e.g. cardiovascular disease, type 2 diabetes, Alzheimer's disease, autoimmune disease, etc, and they have high incidence rate in the US and worldwide [Craig et al., 2008, Cardon and Bell, 2001, Hirschhorn and Daly, 2005]. The development of complex diseases involves genetic factors, environmental factors, behavior factors, and the interactions among them. Identification of the casual factors as well as

the heritability of complex disease has always been hot field in Public Health research. The genetic factors, are probably the most important factors and the starting point people are looking for, then second wave analyses like gene-gene and gene-environment interactions can be followed up. The GWASs have already identified more than 1000 genetic loci associated with many human disease and traits [Hindorff et al., 2009], however, some of them are further validated to be actually false positives and a few solutions such as more powerful tests, meta analysis and wet experiments validation are expected to remedy part of this problem [Wang et al., 2005, Hirschhorn and Daly, 2005, McCarthy et al., 2008, Hindorff et al., 2009, Cantor et al., 2010]. The advent of Next-Generation Sequencing has brought human genetics to a new era [Ansorge, 2009, Metzker, 2009, Mardis, 2008, Shendure and Ji, 2008], and has the potential to explain some of the missing heritability via disease/trait-associated rare variants [Eichler et al., 2010]. Tremendous efforts have already been made by a variety of researchers world-widely 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 majorly designed for cross-sectional data analysis which utilizes less information and is thus less powerful as compared to longitudinal data analysis. Although some of these tests have the flexible framework underpinned to enable extending themselves to accommodate the longitudinal data scenario, the development is not yet done or still under process (unpublished). As the association pattern between variants and disease/trait is subtle and unpredictable, more and more novel association tests put their focus on providing user an "data-adaptive" style tests which can maintain the most power for various data sets encountered in real world. Combining all these thoughts, in my dissertation, the statistical method development will provide people a more powerful and robust data adaptive association test for either CVs or RVs under longitudinal data settings. Both gene-based and pathway-based test strategies will be implemented. Besides, an R package or independent Linux command-line based software implementing the methods will be released shortly after the methodology development, which will greatly facilitate people to use it in real world. In conclusion, my

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

3 Specific Aims and Hypotheses

Current association test methods as reviewed herein before are mainly designed for cross-sectional data analysis while many GWA data did have longitudinal measurements, e.g. the Atherosclerosis Risk in Communities (ARIC) study [Chambless et al., 1997] have multiple follow-up measurements across years. Fully utilizing the information across time points of one individual tends to achieve a greater power and has the potential to identify more disease-related loci [Furlotte et al., 2012, Xu et al., 2014]. As the common variants and rare variants are both important in identifying the disease attributing factors, a well-rounded association test should have the flexibility to work with them both. The test should also maintain a relatively high power in almost all scenarios of the real world data sets, since in real world it is hard to know which specific method has the most power so people could adopt it and avoid those methods suffering poor performance for a specific data set. To meet these urgent needs, I propose to develop a powerful data-adaptive association test adapted for longitudinal data analysis, workable on both CVs and RVs. The detailed aims are:

3.1 Aim One: Data-adaptive association tests for longitudinal data analysis within GEE framework

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

data as to my knowledge.

3.1.1 1(a): A data-adaptive association test for longitudinal data analysis within GEE framework

Develop the data-adaptive longitudinal association test within GEE framework workable for common variants, which will be done in either sliding-window based or gene-based manner for real GWAS data. Since CVs are usually coming from traditional SNP genotyping platform (e.g. selection of tagging SNPs), the SNPs are almost evenly distributed on the whole chromosome. To include, say a constant number of 40 SNPs, the regions covering the 40 SNPs on a chromosome are by large of the same physical length (e.g. 1kb), which makes the slide-window based manner reasonable to detect signals from a specific chromosome region. Gene-based manner has been extensively discussed in section 1. Since missing data scenario is usually the case in longitudinal data analysis, e.g. an individual has three out of four total measurements, the developed algorithm is expected to utilize the partial information fully instead of deleting the whole subject, and still provides consistent coefficients estimates as long as the data missing follows the Missing At Complete Random (MACR) rule [Rubin, 1976, Xu et al., 2014].

[To Dr. Morrison's concern: Add the paragraph of introducing the simulation and analysis of the ARIC data in Aim section]

We will test the proposed novel method's performance using simulated data. Specifically, we will simulate genetics data and longitudinal trait data mimicking real data scenario, e.g. genetics data will allow LD structure among SNPs and longitudinal trait will have flexible control over the between-subject variance, the within-subject variance, the number of measurements and the correlation structure among measurements. We will benchmark our new test against several existing methods, such as Sum test, UminP test and Score test, with regard to the ability to control type I error and achieve a higher empirical power under

different simulation scenarios.

3.1.2 1(b): Longitudinal aSPU family tests on Rare Variants

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

[To Dr. Morrison's concern: Add the paragraph of introducing the simulation and analysis of the ARIC data in Aim section]
We will test on simulation data (genetic and longitudinal trait) the effect of adopting permutation or parametric bootstrap strategy on testing RVs. We expect to see such procedures can provide a better contrl of type I error and provide more unbiased power estimate.

[To Dr. Morrison's concern: special ways that lipids must be treated when analyzing the ARIC data]
We will apply the novel method to ARIC cohort data as mentioned in section 4. Since the ARIC cohort data currently has **five** follow-ups till now, it's a perfect data set for my dissertation research on data-adaptive association test for longitudinal data. We propose to use four closely cardiovascular-disease-related traits measured in ARIC cohort data, which are total cholesterol (tch), High-density lipoprotein (HDL), Low-density lipoprotein (LDL) and triglycerides (trgs), as our response variables (will fit each trait separately by our proposed method). We will **take cautions before using these lipid traits** such as accounting for lipid-lowering therapy in TC and LDL traits, and natural log transformation on the trgs trait according to the procedures described in [Peloso et al., 2014]. We will use genotype data either from traditional microarray genotyping platform for CVs or ExomeChip platform for

RVs. We hope to validate known genetic loci as reported in literature [Teslovich et al., 2010, Lange et al., 2014, Peloso et al., 2014, Consortium et al., 2013, Maxwell et al., 2013] as well as identifying potential novel genetic loci associated with any of these traits. We will exclusively use Caucasian samples. For the covariates, we will include but not limited to subject's demographic information such as age, gender, BMI, etc. The application of the new methods on ARIC cohort data will on one hand help evaluate the performance of proposed methods, and on the other hand will have the potential to identify more genetic loci contributed to atherosclerosis risk.

3.2 Aim Two: Pathway-based longitudinal aSPU family tests: Path-aSPU

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

[To Dr. Morrison's concern: Add the paragraph of introducing the simulation and analysis of the ARIC data in Aim section]

We will simulate the gene pathway including multiple genes, and each gene contains multiple SNVs. We will test our Path-aSPU's performance in controlling type I error and maintaining a higher power under different simulation scenarios as compare to a few existing tests such like GRASS [Chen et al., 2010a] and ALIGATOR [Holmans et al., 2009].

With regard to data application, we will define the gene pathway by public pathway resources like KEGG [Ogata et al., 1999] and BioCarta [Nishimura, 2001]. We will apply Path-aSPU to find association between a specific pathway and longitudinal trait in ARIC data.

3.3 Aim Three: Package/software development

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

4 Real Data Introduction

[To Dr. Morrison's concern: revise real data introduction, e.g. remove unnecessary figure and add description of fifth follow-up measurement] Real data used in my dissertation will be obtained from the Atherosclerosis Risk in Communities (ARIC) Study (https://www2.cscc.unc.edu/aric/). 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, and variation in cardiovascular risk factors, medical care, and disease by race, gender, location, and date. To date, the ARIC project has published over 800 articles in peer-reviewed journals. ARIC includes two parts: the Cohort Component and the Community Surveillance Component.

The Cohort Component of the ARIC study, which will be applied to our new methods, began in 1987, and 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 screen (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 screen is farther apart from the previous screens and was finished during 2011-2013. A detailed description of the ARIC study design and methods was published elsewhere [Investigators et al., 1989].

A demographic introduction of the ARIC cohort data is put 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

Human Subjects

This dissertation study will focuses on statistical method development. I will use ARIC cohort component data for method demonstration purpose only. The ARIC data can be freely requested at (https://www2.cscc.unc.edu/aric/distribution-agreements). I will use the phenotype (several traits as above mentioned) and genotype (SNP array or ExomeChip) data, which are all de-identified. The IRB approval of the use of ARIC data in my dissertation research has been obtained by my dissertation advisor, Dr. Peng Wei, from UTHealth IRB (HSC-SPH-13-0492) for the project "Association analysis of rare variants with sequencing data".

5 Methods

5.1 Aim One (1a): A data-adaptive association test for longitudinal data analysis within GEE framework

5.1.1 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 of the SNPs and covariates on the longitudinal phenotype/trait measurements. Thus, we construct the design matrix for the SNPs and covariates as:

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

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

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

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

The consistent and asymptotically Normal estimates of β and φ can be obtained by solving the GEE [Liang and Zeger, 1986]:

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

with

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

and

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

ϕ in V_i is the dispersion parameter in GEE and is usually treated as nuisance parameter. $v(\mu_{im}) = \phi \text{Var}(y_{im}|x_i, z_i)$. $R_w(\alpha)$ is a working correlation matrix depending on some un-

known parameter α. For convenience, a working independence model with $R_w = I$ is often used. With a canonical link function and a working independence model, we have a closed form of the U vector with two parts corresponding to SNPs and covariates, and its covariance estimator:

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

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

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

Quantitative traits

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

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

if the assumption of a common covariance matrices across Y_i for i is valid, e.g. for quantitative continuous traits study [Pan, 2001], we can adopt a more efficient covariance estimator:

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

which is used by default for its better finite-sample performance [Pan, 2001].

In my dissertation, I will **focus on** the case with quantitative traits, since they are most typ-

ical traits used as response variable in longitudinal data analysis. Nevertheless, I introduce the binary traits strategy as below. In general, the only difference lies in which canonical link we will use, 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 Current 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 more importantly, the type I error tends to inflate in such case [Pan et al., 2014, Zhang et al., 2014].

- **The Score Test:** $T = U_{.2}'\Sigma_{.2}^{-1}U_{.2}^{-1}$, where $U_{.2}$ and $\Sigma_{.2}$ are discussed above; the statistic is asymptotically equivalent to the Wald test with the same null distribution $T \sim \chi_p^2$. Since we only need to fit the null model with covariates, it is computationally easier and less likely to have numerical convergence problems. More importantly, the score test controls the type I error well [Pan et al., 2014, Zhang et al., 2014].

- **The UminP Test:** $T = \underset{j}{max} \frac{U_{.2,j}^2}{\Sigma_{.2,jj}}$ for $j \in 1, 2, \ldots, p$, of jth SNP effect. The $\Sigma_{.2,jj}$ is the jth entry on the diagonal of $\Sigma_{.2}$. With max T, we can get minimal p-value accordingly. An asymptotic multivariate normal distribution numerical inte-

gration based method provided a fast way to calculate its p-value [Pan et al., 2009, Pan, 2009]; alternatively, a simulation based method relying on the asymptotic normal distribution of the score vector can be used to calculate its p-value [Pan et al., 2014, Zhang et al., 2014]. Specifically, we first simulate the score vector $U_{(b)} = (U_{(b).1}, U_{(b).2}, \ldots, U_{(b).p})'$ from its null distribution $U_{(b)} \sim N(0, \Sigma_{.2})$ for $b = 1, 2, \ldots, B$, then calculate a total number of B null statistics: $T^{(b)} = \max_{j=1,\ldots,p} \frac{U^2_{(b).j}}{\Sigma_{.2,jj}}$, and the p-value is calculated as $\sum_{b=1}^{B} \frac{I(T^{(b)} \geq T)+1}{B+1}$.

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

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

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

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

These two tests are called Sum test and SSU test [Pan, 2009]. The former is closely related to other burden tests such like those in [Morgenthaler and Thilly, 2007, Li and Leal, 2008, Madsen and Browning, 2009] If there is a common association either in direction or strength for causal SNVs with no or few non-associated SNVs, then Sum test and the likes will be most powerful; otherwise, the SSU test and its closely relatives, such as kernel machine regression (KMR or SKAT) [Lee et al., 2012a, Ionita-Laza et al., 2013, Oualkacha et al., 2013, Lee et al., 2012b, Wu et al., 2011] and C-alpha test [Neale et al., 2011], will be most powerful.

Sum test and SSU test are all based on score vector. A more general form of score-based

statistic can be generalized as:

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

where $W = (W_1, \ldots, W_p)'$ is a vector of weights for the p SNVs [Lin and Tang, 2011]. Different researchers proposed different weighting schemes to pool the information of all SNVs in a region of interest, such as those used in [Madsen and Browning, 2009, Sul et al., 2011, Pan and Shen, 2011, Han and Pan, 2010, Li and Leal, 2008, Zhang et al., 2011, Lin and Tang, 2011, Basu and Pan, 2011]. However, all of these weighting schema used fixed weights, e.g. proportional to the MAF of SNV, proportional to standard deviation of SNV, proportional to regression coefficient, proportional to single SNV p-value, etc, and there is no uniformly best weighting scheme as shown in [Pan et al., 2014, Basu and Pan, 2011, Pan and Shen, 2011].

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

By above explanation, we can see SPU(γ) test can connect to a few current test with a simplified framework though. Treating $U^{\gamma-1}$ as the weight with $\gamma \geq 1$ have at least two advantages:

First, score vector is as informative as a vector of the estimated regression coefficient while being computationally much simpler and more stable in the case of low frequency SNVs.

Specifically, since U_j contains association information about SNV j and U_j under null hypothesis follows $N(0, V)$, a larger component of $|U_j|$ corresponds to strong evidence of association between the jth SNV and the trait;

Second, it leads to a simple interpretation and a guidance: as the value of γ increases, we up-weight more and more the larger components of the score vector while gradually ignoring the remaining components. Such process smoothly combines the variable weighting and variable selection schema. Besides, an even integer of γ automatically eliminates the effect of opposite signs of U_j, thus avoid power loss due to opposite direction effects canceling out each other; an odd integer of γ might be more appropriate, as in SPU(1), Sum test or other burden tests, when the SNV effects are all in the same direction.

In our experience, SPU(γ) test with a large $\gamma > 8$ usually gave similar results as that of SPU(∞) test [Pan et al., 2014], thus we 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 SNV is large enough for the asymptotic normal distribution of score vector to hold under null hypothesis, we will use a simulation method to calculate the p-value from each $T_{SPU(\gamma)}$ [Lin, 2005, Seaman and Müller-Myhsok, 2005]. Specifically, suppose T is short notation of $T_{SPU(\gamma)}$ for a specific γ and $\hat{\Sigma}_{.2}$ is the covariance matrix of the score vector $U_{.2}$ based on original data (see 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

Although we have a list of SPU(γ) statistics and p-values, we are not sure which one is the most powerful in a specific data situation. Thus, it will be convenient to have a test which data-adaptively and automatically select/combine the best SPU(γ) test(s). We hereby propose an adaptive SPU (aSPU) test to achieve such purpose. There are a list of combining methods, such as exponential combine [Chen et al., 2012], linear combine, quadratic com-

bine and fisher's combine methods [Luo et al., 2010, Peng et al., 2009, Derkach et al., 2013], however in this dissertation work we will use minimum-p combining method exclusively with room left for trying other combining methods. As for different γ, it is difficult to characterize the power curve of an SPU test in real data situation, we will use the p-value of a SPU test to approximate its power; this idea has been prevalent in practice. Accordingly, we will have the aSPU test statistic:

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

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

Similarly as the above simulation method to get p-value of $T_{SPU(\gamma)}$, the *same strategy* can be applied to get the p-value of T_{aSPU} and actually it fully utilizes the previous simulated intermediate result, hereby saves another *unnecessary* simulation work. Specifically, at the SPU test stage we already have the $U_{.2}^{(b)}$ for $b = 1, 2, \ldots, B$. 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 "data-adaptive" aSPU test strategy that is: we first start with a smaller B, say $B = 1000$, to scan the genomes, then gradually increase B to say 10^6 for a few groups of SNVs, e.g. specific genes or windows, which pass an pre-determined significance cutoff (e.g. p-value $\leq 5/B$) in the previous step; repeat this process according to user's specific need until satisfying the significance level accuracy, e.g.

a p-value of $\leq 10^{-7}$ requires $B \geq 10^7$. In this "data-adaptive" way of implementing the simulation based p-value calculating method for aSPU test, we will be able to apply the aSPU test to GWA data.

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

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

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

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

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

  The **aSPUw** test is designed to complement the performance of aSPU test. As the standard deviations of SNVs in a region may vary a lot, there is possibility that a *non-informative* SNV has *larger* standard deviations than other associated SNVs, and the SPU test statistic will be dominated by the noise coming from the null but with larger standard deviation SNV, thus leads to concealing association signals and eventually

reduce the test power. Another advantage **aSPUw** brings about is it makes jointly analyze the effect of RVs and CVs possible by giving them an inverse-standard-deviation weight closely related to MAF.

The **aSPUw** test also has disadvantages, otherwise, we will not keep mentioning **aSPU** test as our flagship test within the aSPU test family (including aSPU, aSPUw, and below aSPU.Score and aSPUw.Score tests). When **variance** of SNVs are quite **homogeneous**, put a variance-based weight (always positive) in the denominator will shrink the test statistics and thus lead to less power. In brief, there will be some scenarios, the aSPU test will dominate aSPUw test, and vice versa. Therefore, it is worth generating both test results for all real-data scenarios of which we don't know the underlying SNV variance situation (homogeneous or heterogeneous). We can compare the results afterwards. The best thing is already mentioned earlier: the two tests can be executed at the same time without extra computation burden.

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

  Although the **GEE Score test** will lose power in some scenario of gene-based GWA analysis as mentioned before, it still has the unique advantage in some scenarios when the correlation structure among SNVs really matters. GEE Score test in the form of $T = U_{.2}^{'}\Sigma_{.2}^{-1}U_{.2}^{-1}$ will keep the covariance matrix in the denominator, which preserves the information of possible linkage disequilibrium among SNVs. To combine the pros of GEE Score test and aSPU(w) test, we propose to adopt the minimum p-value combining strategy again, yielding the aSPU(w).Score test with 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. To calculate the p-value of the aSPU(w).Score test, it is just as simple as to include the Score test p-value along with the other SPU(γ) p-values, select the minimum p-value among them to form the new statistic $T_{aSPU.Score}$,

then use the same simulation algorithm as discussed earlier to get the the $P_{aSPU.Score}$.

The advantage of **aSPU(w).Score** test is we only need to sacrifice a little bit test performance in all scenarios (based on our extensive simulation studies, which is though not shown here), to exchange for a huge improved stability in maintaining a high power in all scenarios (usually when aSPU family performs not so impressive, and Score test happen to be on the edge due to its retaining of the LD information among SNVs).

5.1.2 Methods for Simulation Studies

[To Dr. Fu's concern about whether there is SNP x time interaction effect involved in the simulation, we clarify here]

We generated simulated genotypes following [Wang and Elston, 2007, Pan, 2009, Basu and Pan, 2011]. In brief, we generated two independent blocks of SNPs for each subject: the first block included causal SNPs and others (null SNPs) in linkage disequilibrium (LD); the second block contained only some null SNPs in LD as well. We used first-order auto-regression (AR(1)) correlation structure to imitate real-world LD. We simulated longitudinal response variables using AR(1) as well following [Song et al., 2013]. Then we added the SNPs main effect as a fixed effect to longitudinal response variables **without consideration of SNP $\times$ time interaction**. We did not consider covariate effects in simulation studies, though they can be simply added without any change to the algorithm. We referred to several literatures [Pan et al., 2014, Basu and Pan, 2011, Pan and Shen, 2011, Han and Pan, 2010, Pan, 2009] and the Atherosclerosis Risk in Communities (ARIC) data (https://www2.cscc.unc.edu/aric/ and see section 4) to set up the simulation parameters, e.g. ρ_y across longitudinal measurements and ρ_x across SNVs as used in AR(1) correlation structure model.

[To Dr. Fu's concern about genotype simulation method, we clarify here]

We did notice that there are other strategies in simulating the genetic data, such as the forward time simulation method to generate population genetic data, which includes coales-

cence models like two-epoch model and six parameter complex bottleneck model, and allows for simulation of purifying selection effect and scaled fitness effect as well [Boyko et al., 2008, Hernandez, 2008]. Compared to the populational genetics data simulation method, our simulation method does not take into account the population coalescence theory and assumes each sampled individual genotype is independent to the others. However, our simulation method takes the edges at the flexible control over the correlation magnitude among SNVs, the desired MAF of SNVs and the proportion of casual SNVs. Such advantages were proved and utilized in developing new association tests in a list of past researches [Wang and Elston, 2007, Pan, 2009, Han and Pan, 2010, Pan and Shen, 2011, Basu and Pan, 2011, Pan et al., 2014, Zhang et al., 2014].

Simulation of genotype data

To construct one block of SNPs for subject i, a latent vector $G_i = (G_{i1}, \ldots, G_{ip})'$ was first drawn from a **multivariate normal distribution** $N(0, R)$, where R had a AR(1) correlation structure with its (i, j)th element in terms of purely correlation $r_{ij} = \text{Corr}(G_{if}, G_{ig}) = \rho^{|f-g|}$ between any two latent components, G_{if} and G_{ig} for $f \neq g$. In our simulations we set $\rho = 0.8$. Secondly, the latent vector was dichotomized to yield a haplotype with each latent element G_{ij} dichotomized to 0 or 1 with probability $\text{Prob}(G_{ij} = 1) = \text{MAF of } j\text{th SNP}$; the MAFs were randomly drawn from a uniform distribution: for causal SNPs the MAFs were set between 0.3 and 0.4; for null SNPs the MAFs were set between 0.1 and 0.5. Thirdly, we combined two independent haplotypes to form the genotype $X_i = (X_{i1}, \ldots, X_{ip})'$ for subject i. The haplotypes for different subject were generated **independently**, i.e. the samples are simulated as perfectly **independent** subjects with no Identity by Descent (IBD).

By this strategy we placed 35 SNPs in first block with AR(1) correlation structure to imitate the real LD structure among these SNPs; out of 35 SNPs we randomly set 5 SNPs to be causal (i.e. has a non-zero coefficient in later introduced regression model); to mimic the real data situation in SNP genotyping platforms, e.g. tag SNPs are usually in LD with casual SNPs but not the casual SNPs themselves, we excluded the 5 casual SNPs in the test (thus

in first block, only null SNPs in LD with these 5 casual SNPs will enter the test). We further placed 15 null SNPs in the second block with AR(1) correlation structure as the same as we did in the first block. Note the first block and second block are independent though. All the SNPs from second block will participate in the test.

Simulation of phenotype data

For longitudinal quantitative phenotypes/traits, we adopted the strategy used in [Song et al., 2013]. Specifically, we first did an exploratory analysis (generalized least square estimation with AR(1) correlation structure) on ARIC data to get an approximate estimate of the correlation coefficient between traits across time points, that is $\rho_{data} = 0.7$ on average.

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

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

with $m = 1, \ldots, k$ indexes the longitudinal measurements within subject i as already stated in 5.1.1; $\mu_i = Z_i\varphi + X_i\beta = H_i\theta$ as in quantitative trait case (see 5.1.1); 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, so we can plugin our estimate from real data here by setting up $\rho = 0.7$. $e_{i,m}$ is the total residual, which can be divided into two parts: first part depends on $e_{i,m-1}$ and second part is an independent term. We assume the following distribution:

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

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

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

It's not hard to see the $\rho e_{i,m-1} + s_{i,m}$'s variance by algebraically summing up two parts is equal to the variance of $e_{i,m}$. 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 of the above variance-covariance matrix (6), the first one defines the inter-subject variances, and the second one 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.

Connect phenotype data with genotype data

In association test, we always expect different SNPs contribute to the phenotypes/traits in different unknown patterns. Thus, the SNP effect magnitude tuning in simulation study is not trivial. Instead of assigning a β_d coefficient to a SNP with a random numerical value, say 0.1 or 10000, there is a way to use genetic heritability to control the association magnitude from the jth SNP [Lynch et al., 1998]. Let we first introduce the below splitting of the

phenotype variance:

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

where the Hard-Weinberg equilibrium is assumed to hold. f is the MAF of the SNP; σ_{oth}^2 is the residual variance after removing the effect of 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 extended to:

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

After this point, by systematically solving the equations (7) and (9), we can easily calculate the β_j for jth SNP once we have determined the value of h_j^2, σ_b^2, σ_e^2 and f. Usually a h_j^2 for a single SNP j will not be high for complex disease and we used $h_j^2 = 0.001$ in our simulation study to control β_j.

5.1.3 Plans for Simulation Studies

[To Dr. Fu's concern: 1000 replicates may be not enough to evalute type I error accurately.] We first summarize the default parameters used in simulation studies here:

- $h_j^2 = 0.001$

- $\sigma_b^2 = 1$
- $\sigma_e^2 = 1$
- n varies between 500 and 3000
- $k = 4$
- 1000 replicates of simulated dataset (may increase to **5000** to evaluate the type I error more accurately)
- $\alpha = 0.05$
- $\rho_y = 0.7$
- $\rho_x = 0.8$
- $R = AR(1)$ as the correlation structure to simulate longitudinal trait
- $Rw = I$ as the working independence correlation structure used in GEE estimation

We then have the plans for simulation studies as below:

1. **Power comparison between longitudinal study and cross-sectional study**

   We will test 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, i.e., cross-sectional study, to k, say four as the maximum measurement number). We will also test the power gain magnitude under different levels of within-subject correlation coefficient.

   We are interested in:

   (a) the magnitude of power gain on different level of ρ, the within-subject correlation coefficient as used in 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, say 3, the power gain after it will be negligible as compared to previous power gains. That 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, say a power of 0.9 in longitudinal studies, while avoiding unnecessary cost from pursuing a even larger k.

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

   We will evaluate the type I error of aSPU test and its modifications (we will call them aSPU family hereinafter) as compared with several existing tests: Score GEE, UminP, Sum Test, weighted Sum Test and SSU test. We set up the nominal type I error at 0.05. We provide a sample table below to show the future result presentation format (dummy number 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 default simulation settings with varying sample size**

We will benchmark the empirical power among aSPU family tests and several existing tests. We will keep the type I error 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 type I error table above.

4. **Empirical power benchmark under the simulation settings that half number of casual SNVs are in opposite effect direction**

   In 5 causal SNPs (simulated in the region with all other SNPs but excluded from tests), we will set 2 of them to have opposite effect direction to the left 3 SNPs by flipping the SNP main effect sign. The other settings will keep the same as the above. We will keep the type I error 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 type I error table above.

5. **Empirical power benchmark under the simulation settings that null SNPs number is growing**

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

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

   We will investigating the aSPU family tests performance under other working correlation structures than working independence, such as AR(1), Compound Symmetry, and unstructured. Please be noted, as we simulated the longitudinal trait using AR(1) correlation structure as mentioned in Section 5.1.2, fitting GEE with AR(1) working correlation matrix is actually using the true correlation matrix. We will keep all other settings the same with test scenario 3 (default simulation settings) above. We will present either a figure plotting power curve as a function of n for each participating test under a specific working correlation matrix or a power table in the similar format as type I error table above. We are interested to see the subtle effect from combining a specific working correlation matrix and a specific n for each test.

5.2 Aim One (1b): Longitudinal aSPU family tests on Rare Variants

5.2.1 Statistical Modeling

In the previous section 5.1.1 we discussed the methodology development of aSPU family tests on common variants with a longitudinal trait. In this section, we will discuss the extension of the new methods to rare variants.

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

1. identify the max k across all n subjects, which is the number of longitudinal measurements, e.g. $k = 4$ as used in simulation study in section 5.1.3.

2. detect if the data has missing values, if yes, fill the missing value with NA to complement the data dimension (for example, subject i with $Y_i = (y_{i,1},,,y_{i,4})'$ has two missing measurements at time 2 and time 3. After missing value complementing, it becomes $Y_i = (y_{i,1}, \text{NA}, \text{NA}, y_{i,4})'$). Now we should have all the subjects with each Y_i of dimension equal to $k \times 1$.

3. complement H_i to be of full dimension, i.e. $k \times (p + q + 1)$, for covariates and SNVs. Now we should have $\begin{pmatrix} Y_i & H_i \end{pmatrix}$ as an augmented matrix of dimension $k \times (p + q + 2)$ for each subject i, where $H_i = (Z_i, X_i)$. For total n subjects, we have row-wise binded matrix
$$M = \begin{pmatrix} Y_1 & H_1 \\ Y_2 & H_2 \\ \vdots & \vdots \\ Y_n & H_n \end{pmatrix}$$
of dimension $nk \times (p + q + 2)$.

4. permute the SNV chunk among different individuals, i.e. the X_i in $\begin{pmatrix} Y_i & Z_i, X_i \end{pmatrix}$ with

the X_j in $\begin{pmatrix} Y_j & Z_j, X_j \end{pmatrix}$, where $i \neq j$.

5. with permuted

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

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

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

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

5.2.2 Methods for Simulation Studies

The simulation strategy of RV data is almost the same with previous strategy for generating CV data (see section 5.1.2), 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 as we expect the whole-genome sequencing or exome sequencing platform will identify high density SNVs including the real casual ones.

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

5.2.3 Plans for Simulation Studies

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

   Similarly as we planed the simulation studies for CVs in section 5.1.3, we will test the aSPU family tests' type I error performance as compared to a few existing representative tests for RVs. We will still use the simulation-based P-value calculating method as we tested on CVs before. We compared 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. We set up the nominal type I error at 0.05. We provide a sample table below to show the future presentation format (dummy number 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 default settings with varying sample size**

   We will benchmark the empirical power among aSPU family tests and several existing tests in RV analysis using simulation-based P-value calculating method. We will keep the type I error 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 type I error table above.

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

As we already planned Type I error benchmark using the **simulation-based** P-value calculating methods, now we want to compare it to the Type I error benchmark using **permutation-based** P-value calculating method as proposed in section 5.2.1. We hereby test the aSPU family tests' type I error performance as compared to a few existing representative tests for RVs using the permutation-based P-value calculating method. We will compare 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. We set up the nominal type I error at 0.05. We will present the result use a similar table as shown in example table 2.

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

   We will benchmark the empirical power among aSPU family tests and several existing tests in RV analysis using **permutation-based** P-value calculating method, and will compare it to the one done by **simulation-based** P-value calculating method in RV analysis. We will discuss the observed difference in RV analysis between these two methods. We will keep the type I error 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 table use a similar table as shown in example table 2.

5. **Benchmark with aSPU.aSPUw.Score test**

   Since we know within aSPU, aSPUw, and the Score test, there must be at least one having a satisfactory power in a data scenario depending on association pattern and correlation structure within SNPs, we can combine the three tests like we proposed earlier in combining aSPU and Score tests to become the aSPU.Score test. We can adopt

the same minimum p-value combining strategy as we did for aSPU.Score, specifically

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

We can then test if the performance of aSPU.aSPUw.Score with regard to type I error and empirical power. We are interested in whether the proposed new test combining 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).

5.3 Aim Two: Longitudinal aSPU family tests in a pathway-based manner: Path-aSPU

In the previous sections, we have discussed the methodology development of aSPU family tests in a gene-based or region-based manner for CVs and RVs, here we will discuss the extension of the new methods to the pathway-based manner, so called **Path-aSPU**. Path-aSPU is proposed mainly for analyzing RVs, since RVs are with extreme low MAFs, they need more aggregation (by gene and then by multiple genes in a pathway) to increase the test power. As we are aggregating the RVs in genes then genes in a pathway, there must be a large number of non-associated RVs, an preferred case for the aSPU family as shown in section. This is part of the motivation we propose to develop Path-aSPU.

5.3.1 Statistical Modeling

A pathway analysis involves multiple genes (e.g. 20 as a typical number), too few or too many will yield a somewhat meaningless pathway, e.g. suppose a pathway with only two genes and another pathway with 2000 genes. As the genes within a pathway may contain different numbers of RVs, e.g. a gene has 10 RVs while another gene has 400 RVs, we need to modify the aSPU test to adjust for various gene length to avoid dominant influence from

a large (or small) gene.

Suppose we let the short notation $U_{g.}$ to represent $U_{.2}$ for the RVs X_i'part in the whole score vector (see section 5.1.1), and $U_{g.} = (U_{g,1}, U_{g,1}, \ldots, U_{g,p_g})'$ is the score vector for gene g with p_g RVs of itself, from GEE fitting as the same as in previous aSPU test. Given a pathway (or gene set) S, the gene-specific SPU statistic is as follows:

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

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 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 $(\gamma, \gamma 2)$, we recourse to same simulation or permutation strategy as used in section 5.1.1 and section 5.2.1 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 (which utilized the same simulated U in last step for calculating $P_{Path-SPU(\gamma,\gamma 2;S)}$) to calculate the final **pathway-based aSPU** p-value $P_{Path-aSPU(S)}$.

The intuition of $\gamma 2$ is like that of γ: If we think pathway as gene, the gene as 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 we already explained $\gamma = \infty$'s interpretation in section 5.1.1, it indicates the pathway-based

analysis actually used only one gene - the most heavily associated gene with traits. Since the goal of pathway-based analysis is to take advantage of multiple "co-working" genes, and aggregate more RVs, it is less meaningful to consider the use of a $\gamma 2 = \infty$. Instead, we propose to use $\gamma 2 \in \Gamma 2 = \{1, 2, 4, 8\}$. The reason is that at gene level, the statistic $T_{SPU(\gamma;g)}$ is a positive number, not like $U_{.2,j}$ from GEE model fitting which can have different signs (SNV directions). Thus, deliberately assigning both odd and even number of $\gamma 2$ becomes unnecessary, and we can actually use most representative $\gamma 2$s and expect them to have distinct effect from each other. The $1, 2, 4, 8$ will cover Sum-like test, SSU-like test, and two more tests preferring the sparse-casual-gene situation (e.g. only 2 or 3 genes are associated with traits in a pathway, say with 20 genes).

5.3.2 Methods for Simulation Studies

The simulation strategy of RVs within a gene is the the same as section 5.2.2. We will use the same simulated longitudinal phenotype data as for aim one.

5.3.3 Plan for Simulation Studies

We will simulate a pathway with 20 genes; each gene g will contain p_g RVs with p_g randomly draw from a uniform distribution $U(5, 30)$; 10 of the 20 genes will be randomly selected to be causal, with each casual gene containing 1 causal RV. The RVs within each gene will be simulated as in section 5.2.2. The phenotype data in the simulation study will be the same as before.

We will test Path-aSPU family on the simulated data to evaluate the type I error and power, with comparison to other existing tests like GRASS [Chen et al., 2010a],which executes lasso regression (L1-norm) of eigenSNPs within each gene to achieve variable selection, while performing ridge regression (L2-norm) of eigenSNPs at the gene-set-level to achieve gene effect estimates shrinkage simultaneously; ALIGATOR [Holmans et al., 2009], the association list

go annotator, which is a 'p-value enrichment approach' requiring only pre-computed SNP p-values, uses Fisher's exact test on SNP with minimum p-value for the gene-level association; Plink [Purcell et al., 2007], which is a very popular GWAS analysis tool and plinkSet module within it implements the set-based associate test; the famous GSEA test in association study settings by [Wang et al., 2007].

We can further consider more extensive simulation tests, such as changing the casual RV number within a casual gene or changing the number of casual genes; use independent RVs within a gene instead of correlated RVs in AR(1); test Path-aSPU family with different working correlation matrix in GEE modeling.

5.4 Aim Three: Package/software development

We will develop the Package/software to have below properties:

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

6 Real Data Application of the methods

[To Dr. Morrison's concern: special ways that lipids must be treated when analyzing the ARIC data]

We will apply the novel method on ARIC data as detailed in section 4. Specifically, we will use the four closely cardiovascular-disease-related traits measured in ARIC cohort data, which are total cholesterol (tch), High-density lipoprotein (HDL), Low-density lipoprotein (LDL) and triglycerides (trgs), as our longitudinal response variables (will fit each one separately by our proposed model). We will **take cautions before using these lipid traits** such as accounting for lipid-lowering therapy in TC and LDL traits, and natural log transformation on the trgs trait according to the procedures described in [Peloso et al., 2014]. We hope to validate known genetic loci as reported in literatures [Teslovich et al., 2010, Lange et al., 2014, Peloso et al., 2014, Consortium et al., 2013, Maxwell et al., 2013] as well as identifying potential novel genetic loci associated with any of these traits. We will exclusively use Caucasian samples ($n = 11478$ as from 3). For the covariates, we will include but not limited to subject's demographic information such as age, gender, BMI, etc.

6.1 Data application for Aim One

For Aim (1a), we will use the CVs from traditional SNP genotyping platform used in ARIC study as our genotype data. We will use classical Quality Control (QC) criteria as for GWAS, e.g. MAF, missing rate, HWE, etc. We will include several top principal components eigenvectors (PCs) in the regression to adjust for the potential population structure within Caucasian subjects. We will run gene-based test with gene boundary defined by ANNOVAR [Wang et al., 2010b], a software providing functional annotation of genetic variants from high-throughput sequencing data or array data. We will also run sliding-window based test with 40 consecutive SNPs in a window, while neighboring windows share 10 SNPs (so we will not omit the SNPs signals in the gap between two windows).

For Aim (1b), we will use the RVs from ExomeChip used in ARIC study as our genotype data. We will run gene-based test with gene boundary defined by ANNOVAR [Wang et al., 2010b]. We will use similar QC criteria as for CVs, except we will use aggregate MAF cutoffs (based

on gene level RV counts, e.g. 20 or 40 minor alleles within a gene) as previously done in [Lange et al., 2014, Peloso et al., 2014].

6.2 Data application for Aim Two

For Aim 2, we will focus on RVs in genes and genes in a well-defined pathway. The data application procedures for RVs in genes are the same as in section 6.1. The genes in pathway will be defined by any one from the KEGG [Ogata et al., 1999], BioCarta [Nishimura, 2001] or Gene Ontology [Ashburner et al., 2000]. We will consider the medium size pathways in selected database, e.g. the pathway with 10-30 genes.

References

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

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

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

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

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

[Belsky et al., 2013] Belsky, D. W., Moffitt, T. E., Baker, T. B., Biddle, A. K., Evans, J. P., Harrington, H., Houts, R., Meier, M., Sugden, K., Williams, B., et al. (2013).

Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study. *JAMA psychiatry*, 70(5):534–542.

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

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

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

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

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

[Cardon and Bell, 2001] Cardon, L. R. and Bell, J. I. (2001). Association study designs for complex diseases. *Nature Reviews Genetics*, 2(2):91–99.

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

[Chambless et al., 1997] Chambless, L. E., Heiss, G., Folsom, A. R., Rosamond, W., Szklo, M., Sharrett, A. R., and Clegg, L. X. (1997). Association of coronary heart disease incidence with carotid arterial wall thickness and major risk factors: the atherosclerosis risk in communities (aric) study, 1987–1993. *American journal of epidemiology*, 146(6):483–494.

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

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

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

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

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

[Consortium et al., 2013] Consortium, G. L. G. et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nature genetics*.

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

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

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

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

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

[Derkach et al., 2013] Derkach, A., Lawless, J. F., and Sun, L. (2013). Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests. *Genetic epidemiology*, 37(1):110–121.

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

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

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

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

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

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

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

[Fu et al., 2013] Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., , N. H. L. B. I. E. S. P., and Akey, J. M. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220.

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

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

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

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

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

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

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

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

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

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

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

[Hu et al., 2013] Hu, H., Huff, C. D., Moore, B., Flygare, S., Reese, M. G., and Yandell, M. (2013). Vaast 2.0: Improved variant classification and disease-gene identification using a

conservation-controlled amino acid substitution matrix. *Genetic epidemiology*, 37(6):622–634.

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

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

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

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

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

[Kathiresan et al., 2007] Kathiresan, S., Manning, A. K., Demissie, S., D'Agostino, R. B., Surti, A., Guiducci, C., Gianniny, L., Burtt, N. P., Melander, O., Orho-Melander, M., Arnett, D. K., Peloso, G. M., Ordovas, J. M., and Cupples, L. A. (2007). A genome-wide association study for blood lipid phenotypes in the framingham heart study. *BMC Med Genet*, 8 Suppl 1:S17.

[Keating et al., 2008] Keating, B. J., Tischfield, S., Murray, S. S., Bhangale, T., Price, T. S., Glessner, J. T., Galver, L., Barrett, J. C., Grant, S. F., Farlow, D. N., et al. (2008).

Concept, design and implementation of a cardiovascular gene-centric 50 k snp array for large-scale genomic association studies. *PloS one*, 3(10):e3583.

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

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

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

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

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

[Lee et al., 2012a] Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., , N. H. L. B. I. G. O. E. S. P.-E. S. P. L. P. T., Christiani, D. C., Wurfel, M. M., and Lin, X. (2012a). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*, 91(2):224–237.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[Maxwell et al., 2013] Maxwell, T. J., Ballantyne, C. M., Cheverud, J. M., Guild, C. S., Ndumele, C. E., and Boerwinkle, E. (2013). Apoe modulates the correlation between triglycerides, cholesterol, and chd through pleiotropy, and gene-by-gene interactions. *Genetics*, 195(4):1397–1405.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[Peloso et al., 2014] Peloso, G. M., Auer, P. L., Bis, J. C., Voorman, A., Morrison, A. C., Stitziel, N. O., Brody, J. A., Khetarpal, S. A., Crosby, J. R., Fornage, M., et al. (2014). Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *The American Journal of Human Genetics*, 94(2):223–232.

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

[Pinto et al., 2010] Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T. R., Correia, C., Abrahams, B. S., Almeida, J., Bacchelli, E., Bader, G. D., Bailey, A. J., Baird, G., Battaglia, A., Berney, T., Bolshakova, N., Bölte, S., Bolton, P. F., Bourgeron, T., Brennan, S., Brian, J., Bryson, S. E., Carson, A. R., Casallo, G., Casey, J., Chung, B. H. Y., Cochrane, L., Corsello, C., Crawford, E. L., Crossett, A., Cytrynbaum, C., Dawson, G., de Jonge, M., Delorme, R., Drmic, I., Duketis, E., Duque, F., Estes, A., Farrar, P., Fernandez, B. A., Folstein, S. E., Fombonne, E., Freitag, C. M., Gilbert, J., Gillberg, C., Glessner, J. T., Goldberg, J., Green, A., Green, J., Guter, S. J., Hakonarson, H., Heron, E. A., Hill, M., Holt, R., Howe, J. L., Hughes, G., Hus, V., Igliozzi, R., Kim, C., Klauck, S. M., Kolevzon, A., Korvatska, O., Kustanovich, V., Lajonchere, C. M., Lamb, J. A., Laskawiec, M., Leboyer, M., Le Couteur, A., Leventhal, B. L., Lionel, A. C., Liu, X.-Q., Lord, C., Lotspeich, L., Lund, S. C., Maestrini, E., Mahoney, W., Mantoulan, C., Marshall, C. R., McConachie, H., McDougle, C. J., McGrath, J., McMahon, W. M., Merikangas, A., Migita, O., Minshew, N. J., Mirza, G. K., Munson, J., Nelson, S. F., Noakes, C., Noor, A., Nygren, G., Oliveira, G., Papanikolaou, K., Parr, J. R., Parrini, B., Paton, T., Pickles, A., Pilorge, M., Piven, J., Ponting, C. P., Posey, D. J., Poustka, A., Poustka, F., Prasad, A., Ragoussis, J., Renshaw, K., Rickaby, J., Roberts, W., Roeder, K., Roge, B., Rutter, M. L., Bierut, L. J., Rice, J. P., Salt, J., Sansom, K., Sato, D., Segurado, R., Sequeira, A. F., Senman, L., Shah, N., Sheffield, V. C., Soorya, L., Sousa, I., Stein, O., Sykes, N., Stoppioni, V., Strawbridge, C., Tancredi, R., Tansey, K., Thiruvahindrapduram, B., Thompson, A. P., Thomson, S., Tryfon, A., Tsiantis, J., Van Engeland, H., Vincent, J. B., Volkmar, F., Wallace, S., Wang, K., Wang, Z., Wassink, T. H., Webber, C., Weksberg, R., Wing, K., Wittemeyer, K., Wood, S., Wu, J., Yaspan, B. L., Zurawiecki, D., Zwaigenbaum, L., Buxbaum, J. D., Cantor, R. M.,

Cook, E. H., Coon, H., Cuccaro, M. L., Devlin, B., Ennis, S., Gallagher, L., Geschwind, D. H., Gill, M., Haines, J. L., Hallmayer, J., Miller, J., Monaco, A. P., Nurnberger, Jr, J. I., Paterson, A. D., Pericak-Vance, M. A., Schellenberg, G. D., Szatmari, P., Vicente, A. M., Vieland, V. J., Wijsman, E. M., Scherer, S. W., Sutcliffe, J. S., and Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368–372.

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

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

[Rubin, 1976] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

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

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

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

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

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

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

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

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

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

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

[Teslovich et al., 2010] Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713.

[Tintle et al., 2011] Tintle, N., Aschard, H., Hu, I., Nock, N., Wang, H., and Pugh, E. (2011). Inflated type i error rates when using aggregation methods to analyze rare variants in the

1000 genomes project exon sequencing data in unrelated individuals: summary results from group 7 at genetic analysis workshop 17. *Genetic epidemiology*, 35(S1):S56–S60.

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

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

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

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

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

[Wang et al., 2010b] Wang, K., Li, M., and Hakonarson, H. (2010b). Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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