A Data-adaptive SNP-Set-based Association Test of Longitudinal Traits and the extension to Test on Gene Pathway

by

Yang Yang, M.S

APPROVED:

Thesis chair, PHD

Prof 2, PHD

Minor Prof, PHD

Breadth Prof, PHD

DEDICATION

Persistent support from my family members:

Nainan Hei

&

Tianpeng Yang and Qi Lu

A Data-adaptive SNP-set-based Association Test of Longitudinal Traits and the extension to Test on Gene Pathway

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 PULIC HEALTH
Houston, Texas
November, 2015

ACKNOWLEDGEMENTS

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. I also appreciate great helps from my dissertation committee members: Dr. Han Liang, Dr. Alanna C. Morrison and Dr. Yun-Xin Fu. They are talented experts in their fields and provided me with enormous valuable advice towards my research and writings.

A Data-adaptive SNP-set-based Association Test of Longitudinal Traits and the extension to Test on Gene Pathway

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

Thesis Chair, Peng Wei, PhD

Prof 2, Liang Han, PhD

Minor Prof, Alanna C. Morrison, PhD

Breadth Prof, Yun-Xin Fu, PhD

Contents

List of Tables

List of Figures

1 Background

Genome-wide association studies (GWASs) has been popular ever since 2007, and till now hundreds of GWAS have been published already [McCarthy et al., 2008]. The most popular approach in GWAS was to test the association between complex traits and single nucleotide variant (SNV) one by one, then select the the SNVs meeting a stringent significance level after multiple testing error correction, such as Bonferroni and 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 signal contained within the SNV is weak [Sham and Purcell, 2014]. Such a case becomes even more severe a problem for rare variants (RVs), which usually has MAF below 1% [Bansal et al., 2010]. Although with extremely low MAF, we cannot underestimate RVs' important effects underlying disease risk, which are usually functional and deleterious; RVs also bring over 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 test tailored to SNVs with low MAF and RVs has been 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 gene-based and sets of functionally related genes tests are major directions people 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 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., 2010].

A list of gene-based association tests (majorly designed for RVs) have been proposed in recent years ever since 2007. From the earliest few methods: 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 such as a weighted sum statistics (WSS) which assumes all the alleles to be deleterious also known as Madsen and Browning test (MB test) [Madsen and Browning, 2009]. Many other tests after it actually inherited it and improved the performance in some scenarios [Hoffmann et al., 2010, Zhang et al., 2010, Ionita-Laza et al., 2011, Feng et al., 2011]. The Sum of Squared U-statistics test (SSU) [Pan, 2009], RARECOVER algorithm [Bhatia et al., 2010], sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS) [Zhu et al., 2010, Feng et al., 2011], replication-based test (RBT) 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 [Ionita-Laza et al., 2011],the kernel-based adaptive cluster (KBAC) [Liu and Leal, 2010], a variable-threshold (VT test) approach [Price et al., 2010], a general framework for association testing which combined strength from MB test and VT test to form the most powerful test while setting the weight function ϵ proportional to the set of regression coefficients (EREC method) β in the limit [Lin and Tang, 2011], a data-adaptive sum test (aSum) capable of handling both deleterious and protective direction and allowing collapsing CVs into the test [Han and Pan, 2010],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., 2010], Lasso and group-penalized regression based method [Zhou et al., 2010], the C-alpha test which handles RVs with mixed effect direction well but not able to adjust for covariates (such as population stratification PCs)[Neale et al., 2011], the rare variant weighted aggregate statistic (RWAS)[Sul et al., 2011], the sequence kernel association test (SKAT)[Wu et al., 2011] and its later on modified versions (e.g. SKAT-O which is a weighted linear combination of a burden test and the SKAT variance component test of $\tau^2 = 0$, adjusted-SKAT 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 independtly and identically distributed with an arbitray distribtuion of mean 0 and variance τ^2)[Ionita-Laza et al., 2013, Oualkacha et al., 2013, Lee et al., 2012a, Lee et al., 2012b], a probabilistic disease-gene finder which 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 imroved in VAAST 2 [Hu et al., 2013], a data adaptive tests combing score test, SSU and Sum tests' advantages [Pan and Shen, 2011], a data-driven P-value Weighted Sum Test (PWST) which used both significance and direction of individual variant effect from single-variant analysis to calculate a single weighted sum score [Zhang et al., 2011], an exponential combination (EC) framework for set-based association tests within which 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 [Chen et al., 2012],the weighted score test [Cai et al., 2012], functional linear model and (smoothed) functional principle component analysis based association test [Luo et al., 2011, Luo et al., 2012b, Luo et al., 2012a, Fan et al., 2013], GEE-based kernal machine SNP set association test [Wang et al., 2013], a robust and powerful test using Fisher's method to combine linear and quadratic statistics [Derkach et al., 2013], a unified mixed-effect model testing both group effect equal to 0 and variance component equal to 0, which includes both burden and SKAT tests as special cases by embeding the variant functional information and allowing a variant specific random effect in the model [Sun et al., 2013],etc. 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 phenotypes, e.g. specific association effect directions and sizes, a given test favouring one scenario may or may not perform well in other scenarios [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 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 adaptivity, 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 allels 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].

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 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 [Belsky et al., 2013]. 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 seems intuitive, and recently this fact has been discussed in depth by either simulation study and/or real data analysis [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]. 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 modesl, 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 doing 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 modesls (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 devolop 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.

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 working 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., 2010]. 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. 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., 2010]. The 'pathway' in GWAS usually means a set of co-working genes tightly related. Some commonly used pathway databases/repositories include Kyoto Encyclopedia of Genes and Genomes (KEGG) [Ogata et al., 1999], BioCarta [Nishimura, 2001], . There are two major categories of testing methods: self-contained approach and competitive approach [Wang et al., 2010, Nam and Kim, 2008, Goeman and Bühlmann, 2007]. The difference between the two major tests lies in the null hypothesis each test makes: self-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. Additionally, based on input data type, the tests can be broadly classified into two categories: raw genotypes and SNP p-values. The former requires raw SNP genotyps as input while the latter requires already-calculated a list of SNP p-values. The graphic demo is shown in

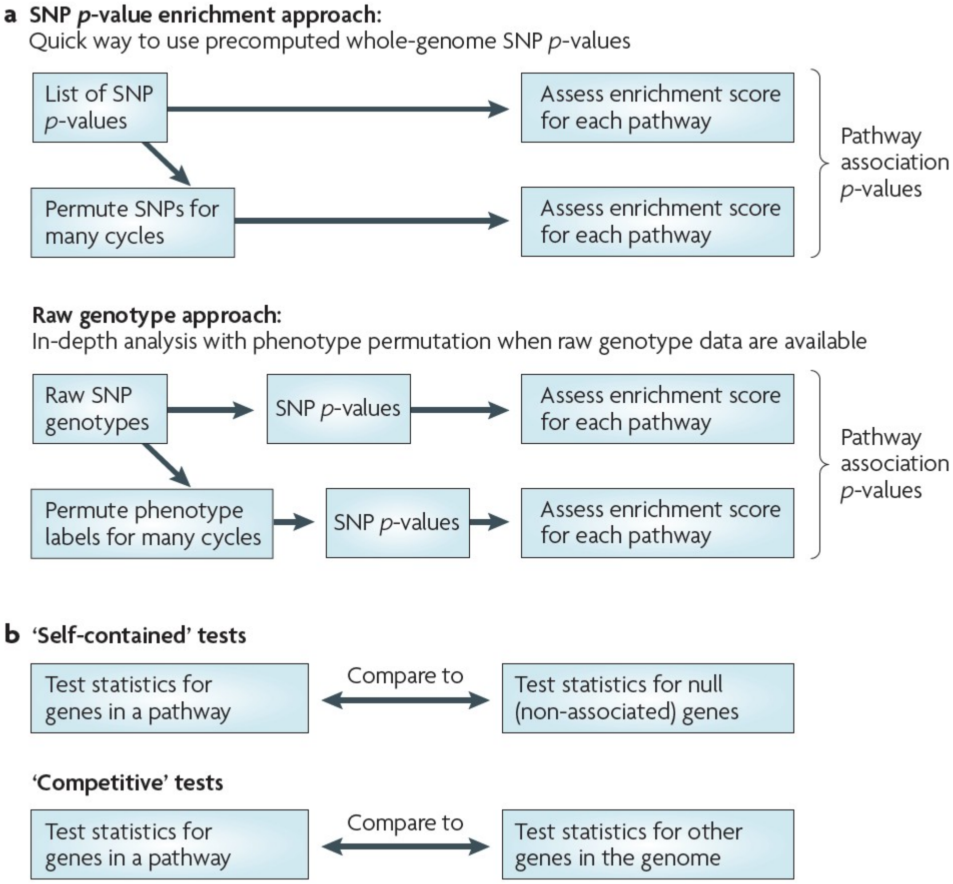
a SNP p-value enrichment approach:
Quick way to use precomputed whole-genome SNP p-values
List of SNP p-values
Assess enrichment score for each pathway
Permute SNPs for many cycles
Assess enrichment score for each pathway
Pathway association p-values
Raw genotype approach:
In-depth analysis with phenotype permutation when raw genotype data are available
Raw SNP genotypes
SNP p-values
Assess enrichment score for each pathway
Permute phenotype labels for many cycles
SNP p-values
Assess enrichment score for each pathway
Pathway association p-values
b 'Self-contained' tests
Test statistics for genes in a pathway
Compare to
Test statistics for null (non-associated) genes
'Competitive' tests
Test statistics for genes in a pathway
Compare to
Test statistics for other genes in the genome

Figure 1: **Types of pathway association tests in GWAS.** (a). Categorization based on data input type; (b). Categorization based on hypothesis testing.

There are several methods current exists including:

pathway: grass [1], gseaSnp [4], plinkSet [3] and aligator [2]. Use

2 Specific Aims and Hypotheses

1. One

2. Two

3. ...

3 Data

4 Methods

4.1 Subsection one

4.1.1 Subsubsection one

4.1.2 Subsubsection two

4.2 Subsection two

4.2.1 Subsubsection one

4.2.2 Subsubsection two

5 Plan for Simulation Studies

5.1 Subsection one

5.2 Subsection two

Table 1: Example table

		A		*B*		*C*	
par	*truth*	*est*	*95% CI*	*est.*	*95% CI*	*est.*	*95% CI*
β_1	10		(,)		(,)		(,)
β_2	1		(,)		(,)		(,)
β_3	-1		(,)		(,)		(,)

5.2.1 Subsubsection one

Figure 2: Sample figure.

References

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

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

[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 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.

[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., 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.

[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.

[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.

[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.

[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 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.

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

[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., 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.

[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.

[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.

[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.

[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.

[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.

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

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

[Wang 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.

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

[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., 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.

[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., 2010] Zhang, L., Pei, Y.-F., Li, J., Papasian, C. J., and Deng, H.-W. (2010). 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.