Published in final edited form as:

Genet Epidemiol. 2017 November ; 41(7): 636–643. doi:10.1002/gepi.22063.

A Functional U-statistic Method for Association Analysis of Sequencing Data

Sneha Jadhav¹, Xiaoran Tong², and Qing Lu^{2,*}

¹Department of Statistics and Probability, Michigan State University

²Department of Epidemiology and Biostatistics, Michigan State University

Abstract

While sequencing studies hold great promise for uncovering novel variants predisposing to human diseases, the high-dimensionality of the sequencing data brings tremendous challenges to data analysis. Moreover, for many complex diseases (e.g., psychiatric disorders) multiple related phenotypes are collected. These phenotypes can be different measurements of an underlying disease, or measurements characterizing multiple related diseases for studying common genetic mechanism. Although jointly analyzing these phenotypes could potentially increase the power of identifying disease-associated genes, the different types of phenotypes pose challenges for association analysis. To address these challenges, we propose a non-parametric method, FU, for multivariate analysis of sequencing data. It first constructs smooth functions from individuals' sequencing data, and then tests the association of these functions with multiple phenotypes by using a U-statistic. The method provides a general framework for analyzing various types of phenotypes (e.g., binary and continuous phenotypes) with unknown distributions. Fitting the genetic variants within a gene using a smoothing function also allows us to capture complexities of gene structure (e.g., linkage disequilibrium), which could potentially increase the power of association analysis. Through simulations, we compared our method to the Multivariate Outcome Score Test (MOST), and found that our test attained better performance than MOST. In a real data application, we apply our method to the sequencing data from Minnesota Twin Study and found potential associations of several nicotine receptor subunit (*CHRN*) genes, including *CHRNA3*, associated with nicotine dependence and/or alcohol dependence.

Keywords

nonparametric method; functional data analysis; similarity measure; multivariate method

Introduction

While significant progress has been made in identifying common genetic variants associated with complex diseases, the identified common variants only explain a small proportion of heritability. Population genetics theory and recent research suggest that rare mutations could play an important role in complex diseases. Nevertheless, current findings of rare variants

*Corresponding Author: Qing Lu, 909 Fee Road, Room 601, East Lansing, Michigan 48824, Telephone: 517.353.8623 x137, qlu@msu.edu.

are still limited. With advancements in next generation sequencing technology, large-scale sequencing studies are now being conducted to comprehensively investigate the contribution of rare variants to the genetic etiology of complex diseases. Although these studies hold great promise for uncovering new disease-associated variants, the massive amount of data and complex structure of sequencing data poses great analytical challenges on association analysis. Advanced methods are in great need to address these challenges and to facilitate the discovery process of new variants predisposing to complex diseases. In addition, there is a great interest in studying genetic variants contributing to multiple related phenotypes. When multiple related phenotypes are collected in a study, there is a potential to improve the power of detecting genetic variants through the joint analysis of multiple related phenotypes. Nevertheless, these multiple phenotypes can be of different types and follow different distributions, which bring additional challenges to association analysis of sequencing data.

Because of the low frequency of rare variants, conventional single-locus tests are subject to low power for the association analysis. Various gene-based tests, including burden tests, adaptive burden tests and variance component tests, have been developed to improve the power by aggregating association signals from multiple rare variants within a gene. A comprehensive summary of literatures regarding association tests is discussed elsewhere (Lee, Abecasis, Boehnke, & Lin, 2014). Briefly, burden tests summarize an individual's genetic information into a weighted or un-weighted summary score, and then test the association of the summary score with a disease phenotype. Collapsing methods, such as the CMC method, are based on a similar idea of summarizing information. These tests are powerful when all genetic variants are causal and affect the traits in the same direction. Nonetheless, if these assumptions are not satisfied, they are subject to suboptimal performance. Adaptive burden tests can overcome this limitation, but they are computationally intensive for large-scale data (Han & Pan, 2010; Price et al., 2010). Variance component tests, such as SKAT (Michael C. Wu et al., 2011) and C-alpha test (Neale et al., 2011), can deal with variants with bi-direction effects and are generally less sensitive to the presence of noise variants. However, C-alpha test uses a permutation test to calculate p-values when genetic variants are in LD, which require more computational time. SKAT uses a random effects model and can deal with continuous or discrete phenotypes. Although SKAT has many advantages over existing methods, it has certain limitations. For example, it can analyze phenotypes with standard distributions but not those with heavy tailed distributions (e.g., t distribution). Moreover, it assumes that the link function is known. Misspecification of link function can lead power loss. Functional methods, such as functional linear models (FLM) (Fan et al., 2013) and functional analysis of variance (FANOVA) (Vsevolozhskaya, Zaykin, Greenwood, Wei, & Lu, 2014) offers an alternative way to jointly test multiple genetic variants in a gene with a phenotype, and have been shown to have a good performance. Functional methods view human genome as a stochastic function, which allow the method to consider correlations among genetic markers. These methods, however, are built for analyzing one phenotype at a time. A few methods, including several non-parametric methods, have been developed for analyzing multiple phenotypes. For instance, there is multiple-trait test, developed based on a summary score (e.g., maximum) of all the variants (Guo, Liu, Wang, & Zhang, 2013). Though such strategy makes it easy to handle the high dimensionality of data, it uses permutation test to calculate

p-values, which can be computationally intensive. There is another score-based test that is initially built for a single genetic variant and is then modified to accommodate multiple variants by using the maximum statistic (He, Avery, & Lin, 2013). Methods considering multiple variants and multiple phenotypes (Y. Wang et al., 2015; B. Wu & Pankow, 2016) have also been proposed. However, these methods are not designed for analyzing multiple categorical phenotypes, a combination of continuous and categorical phenotypes, or phenotypes with unknown distributions.

To address the limitations of existing methods, we propose a non-parametric method, a functional U-statistic method (FU), for multivariate analysis of sequencing data. FU first fits a smooth curve on an individual's sequencing data, and then tests the association of the smooth curve with multiple phenotypes by using a U-statistic framework. The method is also a similarity-based test (Tzeng et al., 2011; C. Wei et al., 2014; Zhang, Liu, & Wang, 2010). In FU, we construct the genetic and phenotypic similarities of two individuals. If there is no association between the genotypes and phenotypes, the genetic similarity does not lead to the phenotypic similarity. However, if there is an association, then the genetic similarity should correspond to the phenotypic similarity.

The rest of the paper is organized as follows. We began with a detailed description of the proposed functional U-statistic and its asymptotic distribution, and then conducted a simulation study where the proposed method was compared with existing tests (He et al., 2013) under a variety of conditions. Finally, we conducted an empirical study by applying the new method to sequencing data from the Minnesota Twin Study.

Method

In the proposed method, we first construct a smooth function from an individual's sequencing data and then build a U-statistic to assess the association of the fitted smooth functions with multiple phenotypes.

Smoothing Function

Because of the high resolution sequencing technology, a large number of single nucleotide variants (SNVs) are often sequenced in a function unit (e.g., a gene). Prior to the association test, we construct a smooth function based on each individual's genetic data, referred to as the genetic variant function. Details of methods for constructing the smooth functions can be found elsewhere (Ramsay & Silverman, 2002; Vsevolozhskaya et al., 2014). Here, we give a brief overview of the smoothing method involving the use of basis functions. Assuming SNVs are sequenced at T ($t_1 \leq t_2 \leq \dots \leq t_T$) genomic positions, we denote the genotype profile for the i^{th} individual by $G_i = (g_{i1}, \dots, g_{iT})$, where g_{ij} are the number of minor alleles (i.e., 0, 1, or 2) for the i^{th} individual at position t_j . Let $g_i(\cdot)$ denote the genetic variant function for the i^{th} individual, we model $g_i(\cdot)$ as follows

$$g_i(t) = \sum_{k=1}^K c_{ik} \Phi_k(t),$$

where the smooth function $g_i()$ is a linear combination of K basis functions, $\Phi_k(t)$. The degree of smoothness of the function g_i is determined by the number of basis functions. A large number of basis functions (i.e. K is large) will result in a highly variable curve, which could lead to an overfitting issue. On the other hand, a small value of K could have the issue of over-smoothing. To identify the appropriate number of basis function, we start with a large value of K and then use a penalty to penalize the roughness of the function.

Let $\Phi_{T \times K} = [\Phi_K(t_j)]$ be the $T \times K$ matrix of basis functions. Penalized least squares criterion can then be used to determine coefficients c_{ik} , i.e. we choose coefficients to minimize

$$\sum_{j=1}^T \left[g_{ij} - \sum_{k=1}^K c_{ik} \Phi_k(t_j) \right]^2 - \lambda \text{PEN}_2(g_i).$$

$\text{PEN}_2(g_i) = \int D^2 g_i(s) ds$, where D^2 denotes the second derivative. A large value of the tuning parameter λ leads to a smoother curve whereas a small λ results in a more variable curve. Common choices for basis functions are the Fourier, Polynomial bases, and B-splines. Because the Fourier and Polynomial basis functions are unable to accommodate local features, we use the cubic B-splines to construct genetic variant functions. The spline smoothing can be implemented by using the smooth.spline function from the fda package in R. The tuning parameter can be determined using cross validation. The smooth.spline function in SPSS has the option of determining the tuning parameter from the data. The fitted genetic variant functions can then be used for association testing.

Functional U-statistic

After fitting the smoothing functions on SNVs, we can test for their association with multiple phenotypes. For such purpose, we propose a functional U-statistic, which is able to deal with different types of phenotypes (e.g., continuous and binary phenotypes) with various underlying distributions. Let n be the number of individuals sequenced in the study.

For the i^{th} individual, we measure m phenotypes, $\mathbf{y}_{i(1 \times m)} = (y_i^1, \dots, y_i^m)$, and obtain the genetic variant function, $g_i()$, using the smoothing method described above. We assume observations $(y_i, g_i(t), t \in R)$, $i=1, \dots, n$, are independent and identically distributed, where R is the genomic region under consideration (e.g., a gene). We standardize the positions so that R has a range of $[0, 1]$. We further center both phenotypes and the

smoothing functions, and assume $E\|Y\|^2 < \infty$ and $E \int_0^1 g(t)^2 dt < \infty$, where the norm is the usual Euclidean norm. Because SNVs are coded as 0, 1 and 2, the smoothing function is bounded.

The proposed method can be seen as an extension of the method proposed by Zhang et.al (Zhang et al., 2010) in the sense that the new method can consider multiple loci and is not conditioned on phenotypes. Our method also has the connection to other similarity based methods. Similar to other similarity based methods, we quantify the phenotypic similarity

between i^{th} and j^{th} individuals and measure their genetic similarity based on the fitted genetic variant functions. For simplicity, we define the phenotypic similarity, u_{ij} , as the difference between the phenotype vectors of two individuals, i.e. $u_{ij} = \mathbf{y}_i - \mathbf{y}_j$. The genetic similarity can be defined based on the difference of two genetic variant functions,

$v_{ij} = \int_0^1 g_i(t) dt - \int_0^1 g_j(t) dt$. In addition to the above genetic similarity measure, other similarity measures, such as L^2 norm, can also be used. Given the formed phenotypic and genetic similarities, we use the following statistic to test whether there is an association of SNVs in a genetic region with at least one of the m phenotypes,

$$T = \frac{\sum_{1 \leq i < j \leq n} u_{ij} v_{ij}}{\binom{n}{2}}.$$

A major advantage of using the above statistic is that we can use an existing theorem (Hoeffding, 1948) to get the asymptotic distribution of the test statistic. By using the existing theorem, we are able to show that under H_0 ,

$$\sqrt{n}T \sim N(\mathbf{0}, V_0),$$

where $V_0 = 4 \text{cov}(\mathbf{Y}) \text{var}(Z)$ and $Z = \int_0^1 g(t) dt$. From this result, we have $nT'V_0^{-1}T \sim \chi_r^2$, where $r = \text{rank}(V_0)$. In practice, the covariance matrix V_0 is unknown, and we can estimate the covariance matrix using $\hat{V}_0 = \hat{\sigma}_z \hat{V}_y$, where $\hat{\sigma}_z$ and $\hat{V}_y$ are moment estimates of variance of Z and covariance matrix of $\mathbf{Y}$, respectively.

Simulation Studies

We performed simulations to assess the performance of the proposed method and compared it with an existing method, MOST (He et al., 2013) under different sample sizes, phenotype distributions, and the proportion of causal variants. MOST proposes 3 tests (i.e., Q test, T test and T' tests) for association analysis of a single variant with multiple phenotypes. Q test is a chi-square test based on the score function. T and T' test are designed to increase power when the genetic effects are similar for all phenotypes (e.g., the effects have same direction) and asymptotically follow a normal distribution. All three methods are then extended to deal with multiple variants by using a maximum function without the requirement of multiple testing adjustments. To mimic the real data scenario, sequencing data from 1000 Genome project was used. In particular, we used a 1Mb region of the genome (Chromosome 17: 7344328-8344327) from 1092 individuals in 1000 Genome Project. For each replicate, we randomly chose a 30 kb segment from the genome. From this segment, we randomly selected 2 regions as the disease-related region and varied the total number of disease-related SNVs according to the specified proportion of causal SNVs (i.e., 0% - 50%). For each simulation setting, we simulated 1000 replicates and evaluated the type I error and power of two methods. For each replicate, the simulated samples were randomly chosen from the

data, and the phenotypes were then generated from the pre-specified distribution based on the disease-associated SNVs. To evaluate the performance of the two methods on multiple phenotypes, we simulated different combinations of phenotypes, including binary phenotypes and continuous phenotypes with normal and t distributions. While normal-distributed and binary phenotypes are commonly observed phenotypes, the t -distributed phenotype represents a continuous phenotype with more extreme values (phenotypes from a study with a small sample size or the extreme-phenotype design). Normally-distributed phenotype was generated from a normal distribution,

$$y_i \sim N(G_i^t \beta, 1),$$

where y_i is the phenotype, G_i is the observed genotype of the disease-associated SNVs, and β are the effects of the SNVs generated from a uniform distribution of $U(0.05, 0.15)$. For non-disease-associated genotypes, the effect is set to 0. For the binary phenotype, we used the logistic regression model,

$$\text{logit}(P(y_i=1)) = G_i^t \beta.$$

The effect sizes, β , were generated from a uniform distribution of $U(0.08, 0.15)$.

For t distribution, we used the following model to simulate the phenotype,

$$y \sim t(df=2, ncp=m_i), \text{ where } m_i = G_i^t \beta.$$

The effect sizes were generated from a uniform distribution of $U(0.05, 0.15) \cdot 1000$ replicates were simulated for each scenario. In the first simulation, we set 30% SNVs as causal SNVs, and evaluated the performance of methods by varying sample sizes. In the second simulation, we fixed the sample size (i.e., $n=500$) and studied the effect of proportion of causal variants on the performance of the methods.

The results of type I error rates for FU and MOST are summarized in Table 1. The results show that type I error rates of two methods remain controlled at the level of 0.05 for normal-distributed and binary phenotypes under different study sample sizes. However, when the phenotypes have more extreme values (i.e., followed a t distribution), MOST has inflated type I error rates while FU has a well-controlled type I error.

In figure 1, we compare the performance of FU with the three MOST tests by varying sample sizes (He et al., 2013). As suggested in previous paper, we used the maximum statistic based on all SNVs, thus bypassing the issue of multiple test adjustment. We used permutation to obtain the p-value of the maximum statistic. As expected, both methods gain power with the increased sample size. Overall, the new method, FU, attains higher power than the MOST method. When all the phenotypes satisfy normal assumption (i.e., NNN), two methods have close performance. Nevertheless, when the phenotype follows a heavy-tailed distribution (e.g., TTT), FU has much higher power than the MOST method. For

binary phenotypes (i.e., BBB), we also find that our method outperforms the MOST method (He et al., 2013).

In real data scenario, it is likely that only a few variants are associated with the phenotype of interest. In the second simulation, we evaluated the impact of different proportions of causal variants on the performance of two methods. We find both methods have increased power with the increased proportions of causal variants. With the increased proportions of causal variants, our method attains higher power than MOST for binary and *t*-distributed phenotypes. The two methods have similar performance when the phenotype follows a normal distribution (figure 2).

Application to the Sequencing Data from the Minnesota Twin Study

Substantial substance use disorders (SUD) and related health conditions emerged by mid-20th century and continue to represent a remarkable 21st century global burden of disease (Mathers & Loncar, 2006). As argued by others, this burden is largely driven by the substance dependence (SD) process. Twin studies and family-based studies have suggested a substantial genetic contribution to SD (e.g., nicotine dependence). With advancements in next-generation sequencing technologies, sequencing studies has become increasingly used in SD research. These studies generate a massive amount of sequencing data and allow researchers to comprehensively investigate the role of a large number of genetic variants, including rare variants, in SD. Many of these studies assess multiple SD phenotypes (e.g., nicotine dependence and alcohol dependence). Simultaneously analyzing these related phenotypes can potentially improve the power of detecting SD-associated variants, and will ultimately help elucidate the genetic etiology of SD. While progress has been made in identifying genetic variants associated with SD through linkage, genetic association, animal, and genome-wide association studies, our understanding of SD genetic etiology is still limited. Recent studies have provided evidence that the neuronal nicotinic acetylcholine receptors (*nAChRs*) may play an important role in nicotine dependence and alcohol dependence, in the sense that alcohol controls the pharmacological binding of nicotine at *nAChRs* (John, Meyer, Rumpf, & Hapke, 2003) (Riala, Hakko, Isohanni, Jarvelin, & Rasanen, 2004).

In this application, we studied the role of 15 *nAChRs* subunit genes on SD by using the sequencing data from the Minnesota Twin Study (MTS) data (Vrieze et al., 2014). MTS is a longitudinal study of twins born in the state of Minnesota between 1972 and 1984. A majority of the samples in the MTS have self-reported European ancestry, which is confirmed by genome-wide principal components. Individuals in the MTS were assessed with various substance-related phenotypes, such as alcohol, nicotine and other illicit drug dependence. The phenotypic measures were based on protocols of the Substance Abuse Module of the Composite International Diagnostic Interview. The detailed description of the phenotypes and diagnostic criteria can be found elsewhere (Hicks, Schalet, Malone, Iacono, & McGue, 2011). MTS whole-genome sequencing is based on Illumina HiSeq technology with 100 or 150 base pair paired-end reads. The variant-calling pipeline, Genomes on the Cloud, was used for genotype calls, which discovered nearly 27M SNVs. Prior to the association analysis, we first conducted quality assessment of the data (e.g., Mendelian error

check). After a quality control process, we annotated the VCF file and then extracted 15 *nAChRs* subunit genes. We applied both methods to 15 genes to evaluate their contribution to ND and/or AD by adjusting for sex and age. To consider covariates (i.e., sex and age), we first fit a generalized linear model for each phenotype with covariates and used the residuals for association analysis. Because both methods are developed for unrelated samples, we included 662 unrelated samples from the 662 families in this analysis. MOST is able to directly incorporate the covariates. In the following table, we report the p-values associated with the 15 *nAChRs* subunit genes. For MOST based tests, we report permutation based p-values.

The associations of 15 *nAChRs* subunit genes with ND and/or AD are summarized in Table 2. From the association analysis, FU identifies two genes, *CHRNA6* and *CHRNA3*, contributing to ND and/or AD with p-values of 0.00329 and 0.00253, respectively. The p-values of two associations remain significant after Bonferroni correction. The Most tests also detected association of *CHRNA6* with ND and/or AD. However, the evidence of association is inconsistent among three tests. Moreover, the associations do not reach significance after multiple testing adjustment.

Discussion

These diagnostic assessments of complex diseases can often be of different types (e.g., binary or continuous measurements) with unknown underlying distributions. For example, one of the most commonly used nicotine dependence measurements, number of cigarettes smoked per day, does not follow a standard distribution (e.g., Poisson distribution). This issue is however paid less attention by the existing analytic methods. Most of the existing methods are parametric-based or semi-parametric-based, which often rely on certain assumptions (e.g., a normal or Poisson distribution assumption). When the assumptions are violated, they are subject to power loss and/or inflated type I error results (C. Wei et al., 2014). It is also a great interest in biology to investigate genetic variants contributing to multiple phenotypes. Although simultaneously analyzing multiple phenotypes can potentially improve the power of detecting disease-associated variants, the different types of phenotypes (e.g., binary and continuous phenotypes) pose statistical challenges for data analysis. Non-parametric methods, such as U-statistic, have shown their advantage of analyzing different types of phenotypes, especially when the underlying phenotypic distributions are unknown. There are several other methods that use U-statistic for association analysis. Most tests can deal with only a single variant or a single phenotype (Li, 2012); (Schaid, McDonnell, Hebbeling, Cunningham, & Thibodeau, 2005); (Z. Wei, Li, Rebbeck, & Li, 2008). The MOST test (He et al., 2013) originally considers a single locus at a time, and is then extended to multiple variants. Guo et.al (Guo et al., 2013) proposed a test based on Kendalls Tau that can deal with multiple phenotypes and multiple variants. It can be used to test for associations of both qualitative and quantitative phenotypes. However, their method is computationally intensive as permutation is required to calculate p-values. We derive the asymptotic distribution of FU, which makes it computationally feasible for analyzing large-scale sequencing data.

By fitting SNVs in a gene with a smoothing function, the new method is able to capture additional genetic information (e.g., positional and linkage disequilibrium information), which could potentially improve the performance of the method. In our method, we use the cubic B-splines to construct the smoothing curves and penalized splines to determine the appropriate number of knots. Nevertheless, other choices of basis functions and different number of knots can be used for the smoothing procedure. We measure the genetic similarity based on the difference between the areas under the smoothing curves, which can take the position information into account. Nevertheless, such measurement could be subject to low power when the effects are bidirectional as the difference between the areas curves cancels out at locations with bidirectional effects. To catch the bidirectional effect, we can measure the genetic similarity by using functions that are symmetric. For example, we can use L^2 norm to quantify the difference between smoothing functions. The resulting statistic will still be a U-statistic but with a more complicated mean and covariance structure.

While both FU and MOST can handle multiple traits, the main difference between two methods lies in the analysis of genetic variants. MOST tests use a maximum statistic of single-locus tests to deal with multiple variants. Though this strategy bypasses the issue of multiple testing, it does not consider linkage disequilibrium (LD) between the variants. The smoothing method used in FU considers LD and the variant position information. With additional information, FU can potentially gain more power than MOST.

Functional methods, such as FU, are also related to haplotype-based methods in the sense that they both consider linkage disequilibrium (LD). By taking the LD structure of genetic data into account, both methods have great potential to increase the test's power. While both methods consider LD, two methods use different strategies to model LD. Haplotype methods first estimate haplotypes from unphased genotypes and then assess their association with phenotypes. Haplotype methods are biologically meaningful, and have an advantage over other methods when specific haplotypes are associated with phenotypes. However, the haplotype inference process could be computationally intensive for large-scale sequencing data. With the increasing level of heterozygosity, the presence of rare mutations, and the increasing number of samples, the number of possible haplotypes could substantially increase, which brings computational burden to haplotype inference. With the increasing number of estimated haplotypes, the association test could also be subject to large degrees of freedom, which might lead to low power of the test. The functional method does not infer haplotypes. It first fits a smoothing curve (i.e., a genetic variant function) on each individual's genetic data, and then evaluates the association of smoothing curves with the phenotypes. Therefore, the computation and power performance of the method is less affected by the level of heterozygosity and the presence of rare mutations. We also expect the computation time of functional methods only linearly increases with the increasing number of individuals. Nevertheless, functional methods do not estimate haplotypes, and thus they are less efficient to capture specific haplotype effects.

From the application of FU to the sequencing data from the MTS study, we found the evidence of the association of *CHRNA6* and *CHRNA3* with ND and/or AD. This result is consistent with previous findings, suggesting the potential role of *CHRNA6* and *CHRNA3* in ND and/or AD (Culverhouse et al., 2014; Haller et al., 2014; S. Wang et al., 2014). While

there is an association evidence of two genes associated with SD, further studies are necessary to further investigate their biological role in SD.

Acknowledgments

The Project was supported by the National Institute on Drug Abuse under Award Number K01DA033346 and R01DA043501. We thank Drs. Scott Vrieze, Matt Mc Gue, and S. Alexandra Burt help us access the whole-genome sequencing data from the Minnesota Twin Study.

References

- Culverhouse RC, Johnson EO, Breslau N, Hatsukami DK, Sadler B, Brooks AI, Bierut LJ. Multiple distinct CHRN3-CHRNA6 variants are genetic risk factors for nicotine dependence in African Americans and European Americans. *Addiction*. 2014; 109(5):814–822. DOI: 10.1111/add.12478 [PubMed: 24401102]
- Fan R, Wang Y, Mills JL, Wilson AF, Bailey-Wilson JE, Xiong M. Functional linear models for association analysis of quantitative traits. *Genet Epidemiol*. 2013; 37(7):726–742. DOI: 10.1002/gepi.21757 [PubMed: 24130119]
- Guo X, Liu Z, Wang X, Zhang H. Genetic association test for multiple traits at gene level. *Genet Epidemiol*. 2013; 37(1):122–129. DOI: 10.1002/gepi.21688 [PubMed: 23032486]
- Haller G, Kapoor M, Budde J, Xuei X, Edenberg H, Nurnberger J, Goate A. Rare missense variants in CHRN3 and CHRNA3 are associated with risk of alcohol and cocaine dependence. *Hum Mol Genet*. 2014; 23(3):810–819. DOI: 10.1093/hmg/ddt463 [PubMed: 24057674]
- Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered*. 2010; 70(1):42–54. DOI: 10.1159/000288704 [PubMed: 20413981]
- He Q, Avery CL, Lin DY. A general framework for association tests with multivariate traits in large-scale genomics studies. *Genet Epidemiol*. 2013; 37(8):759–767. DOI: 10.1002/gepi.21759 [PubMed: 24227293]
- Hicks BM, Schalet BD, Malone SM, Iacono WG, McGue M. Psychometric and genetic architecture of substance use disorder and behavioral disinhibition measures for gene association studies. *Behav Genet*. 2011; 41(4):459–475. DOI: 10.1007/s10519-010-9417-2 [PubMed: 21153693]
- Hoeffding W. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*. 1948; 19(3):293–325.
- John U, Meyer C, Rumpf HJ, Hapke U. Probabilities of alcohol high-risk drinking, abuse or dependence estimated on grounds of tobacco smoking and nicotine dependence. *Addiction*. 2003; 98(6):805–814. [PubMed: 12780369]
- Lee S, Abecasis Gonçalo R, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American Journal of Human Genetics*. 2014; 95(1):5–23. DOI: 10.1016/j.ajhg.2014.06.009 [PubMed: 24995866]
- Li H. U-statistics in Genetic Association Studies. *Human genetics*. 2012; 131(9):1395–1401. DOI: 10.1007/s00439-012-1178-y [PubMed: 22610525]
- Mathers CD, Loncar D. Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLoS Med*. 2006; 3(11):e442.doi: 10.1371/journal.pmed.0030442 [PubMed: 17132052]
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melandar M, Daly MJ. Testing for an Unusual Distribution of Rare Variants. *PLoS Genet*. 2011; 7(3):e1001322.doi: 10.1371/journal.pgen.1001322 [PubMed: 21408211]
- Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *The American Journal of Human Genetics*. 2010; 86(6):832–838. DOI: 10.1016/j.ajhg.2010.04.005 [PubMed: 20471002]
- Ramsay, JO., Silverman, BW. *Applied functional data analysis: methods and case studies*. Vol. 77. Citeseer; 2002.
- Riala K, Hakko H, Isohanni M, Jarvelin MR, Rasanen P. Teenage smoking and substance use as predictors of severe alcohol problems in late adolescence and in young adulthood. *J Adolesc Health*. 2004; 35(3):245–254. DOI: 10.1016/j.jadohealth.2003.08.016 [PubMed: 15313508]

- Schaid DJ, McDonnell SK, Hebring SJ, Cunningham JM, Thibodeau SN. Nonparametric Tests of Association of Multiple Genes with Human Disease. *American Journal of Human Genetics*. 2005; 76(5):780–793. [PubMed: 15786018]
- Schlaepfer IR, Hoft NR, Ehringer MA. The genetic components of alcohol and nicotine co-addiction: from genes to behavior. *Curr Drug Abuse Rev*. 2008; 1(2):124–134. [PubMed: 19492010]
- Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy Mark I, Sale Michèle M, Sullivan Patrick F. Studying Gene and Gene-Environment Effects of Uncommon and Common Variants on Continuous Traits: A Marker-Set Approach Using Gene-Trait Similarity Regression. *American Journal of Human Genetics*. 2011; 89(2):277–288. DOI: 10.1016/j.ajhg.2011.07.007 [PubMed: 21835306]
- Vrieze SI, Malone SM, Vaidyanathan U, Kwong A, Kang HM, Zhan X, Iacono WG. In search of rare variants: preliminary results from whole genome sequencing of 1,325 individuals with psychophysiological endophenotypes. *Psychophysiology*. 2014; 51(12):1309–1320. DOI: 10.1111/psyp.12350 [PubMed: 25387710]
- Vsevolozhskaya OA, Zaykin DV, Greenwood MC, Wei C, Lu Q. Functional analysis of variance for association studies. *PLoS One*. 2014; 9(9):e105074.doi: 10.1371/journal.pone.0105074 [PubMed: 25244256]
- Wang S, van der Vaart A, Xu Q, Seneviratne C, Pomerleau OF, Pomerleau CS, Li MD. Significant Associations of CHRNA2 and CHRNA6 with Nicotine Dependence in European American and African American Populations. *Hum Genet*. 2014; 133(5):575–586. DOI: 10.1007/s00439-013-1398-9 [PubMed: 24253422]
- Wang Y, Liu A, Mills JL, Boehnke M, Wilson AF, Bailey-Wilson JE, Fan R. Pleiotropy Analysis of Quantitative Traits at Gene Level by Multivariate Functional Linear Models. *Genet Epidemiol*. 2015; 39(4):259–275. DOI: 10.1002/gepi.21895 [PubMed: 25809955]
- Wei C, Li M, He Z, Vsevolozhskaya O, Schaid DJ, Lu Q. A weighted U-statistic for genetic association analyses of sequencing data. *Genet Epidemiol*. 2014; 38(8):699–708. DOI: 10.1002/gepi.21864 [PubMed: 25331574]
- Wei Z, Li M, Rebbeck T, Li H. U-Statistics-based Tests for Multiple Genes in Genetic Association Studies. *Annals of human genetics*. 2008; 72(Pt 6):821–833. DOI: 10.1111/j.1469-1809.2008.00473.x [PubMed: 18691161]
- Wu B, Pankow JS. Sequence Kernel Association Test of Multiple Continuous Phenotypes. *Genet Epidemiol*. 2016; 40(2):91–100. DOI: 10.1002/gepi.21945 [PubMed: 26782911]
- Wu, Michael C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics*. 2011; 89(1):82–93. DOI: 10.1016/j.ajhg.2011.05.029 [PubMed: 21737059]
- Zhang H, Liu CT, Wang X. An Association Test for Multiple Traits Based on the Generalized Kendall's Tau. *J Am Stat Assoc*. 2010; 105(490):473–481. DOI: 10.1198/jasa.2009.ap08387 [PubMed: 20711441]

Figure 1. The power comparison of two methods by varying study sample sizes
 B, N, and T represent binary-distributed, Gaussian-distributed, t -distributed phenotypes, respectively; FU, Q, T and T' represent the new test and three tests from MOST, respectively.

Figure 2. The power comparison of two methods for various proportions of causal variants B, N, and T represent binary-distributed, Gaussian-distributed, t -distributed phenotypes, respectively; FU, Q, T and T' represent the new test and three tests from MOST, respectively.

Table I

The type I error rates of two methods for different combinations of phenotypes

Phenotypes	Sample Size	FU ^a	MOST		
			Q	T	T'
NNN ^b	50	0.048	0.040	0.061	0.048
	250	0.042	0.042	0.032	0.031
	500	0.042	0.042	0.033	0.031
BBB	50	0.044	0.042	0.040	0.042
	250	0.043	0.050	0.060	0.060
	500	0.039	0.050	0.050	0.050
TTT	50	0.044	0.041	0.047	0.058
	250	0.049	0.049	0.037	0.038
	500	0.034	0.068	0.072	0.080

^aFU, Q, T and T' represent the proposed test and three tests from MOST, respectively;

^bN, B, and T represent Gaussian-distributed, binary-distributed, τ -distributed phenotypes, respectively.

Table II

Association of 15 *nACHRs* subunit genes with ND and/or AD

Gene	Chromosome	FU ^a	MOST		
			Q	T	T'
<i>CHRNA1</i>	2	0.700	0.231 ^c	0.060	0.527
<i>CHRNA2</i>	8	0.893	0.338	0.284	0.405
<i>CHRNA3</i>	15	0.832	0.950	0.595	0.828
<i>CHRNA4</i>	20	0.326	0.353	0.568	0.174
<i>CHRNA5</i>	15	0.767	0.389	0.467	0.219
<i>CHRNA6</i>	8	0.003 ^b	0.104	0.452	0.107
<i>CHRNA7</i>	15	0.388	0.818	0.842	0.512
<i>CHRNA9</i>	4	0.590	0.902	0.454	0.725
<i>CHRNA1</i>	17	0.619	0.792	0.941	0.640
<i>CHRNA2</i>	1	0.265	0.145	0.254	0.681
<i>CHRNA3</i>	8	0.003 ^b	0.191	0.327	0.591
<i>CHRNA4</i>	15	0.293	0.299	0.399	0.223
<i>CHRNA</i>	2	0.253	0.024	0.146	0.010
<i>CHRNA</i>	17	0.366	0.062	0.816	0.077
<i>CHRNA</i>	2	0.716	0.761	0.831	0.600

^aFU, Q, T and T' represent the proposed test and three tests from MOST, respectively;
^bP-values remain significant after multiple testing adjustment;
^cThe smallest unadjusted p-values among all SNVs are reported for the MOST tests.