Wang et al. BMC Proceedings 2014, 8(Suppl 1):S91
http://www.biomedcentral.com/1753-6561/8/S1/S91

BMC Proceedings

PROCEEDINGS Open Access

Detecting association of rare and common variants by testing an optimally weighted combination of variants with longitudinal data

Shuaicheng Wang, Shurong Fang, Qiuying Sha, Shuanglin Zhang*

From Genetic Analysis Workshop 18
Stevenson, WA, USA. 13-17 October 2012

Abstract

Increasing evidence shows that complex diseases are caused by both common and rare variants. Recently, several statistical methods for detecting associations of rare variants have been developed, including the test for testing the effect of an optimally weighted combination of variants (TOW) developed by our group in 2012. These methodologies consider phenotype measurement at only one time point. Because many sequence data have been developed on population cohorts that contain phenotype measurements at multiple time points, such as the data set provided in the Genetic Analysis Workshop 18 (GAW18), we extend TOW from phenotype measurement at one time point to phenotype measurements at multiple time points. We then apply the newly proposed method to the GAW18 data set and compare the power of the new method with TOW using only one phenotype measurement. The application results show that the newly proposed method jointly modeling phenotype measurements at all time points has increased power over TOW.

Background

There is increasing interest in detecting associations between rare variants and complex traits. Although statistical methods to detect common variant associations have been well developed, these variant-by-variant methods may not be optimal for detecting associations of rare variants as a result of allelic heterogeneity as well as the extreme rarity of individual variants [1]. Recently, several statistical methods for detecting associations of rare variants have been developed, including the cohort allelic sums test [2], the combined multivariate and collapsing method [1], the weighted sum statistic [3], and the variable minor allele frequency threshold method [4], among others. These methods are essentially testing the effect of a weighted combination of variants. Thus, choosing appropriate weights is critical to the performance of these methods. In Sha et al [5], we proposed a novel test for testing the effect of an optimally weighted combination of variants (TOW). The optimal weights are analytically derived. Based on the optimal weights, TOW tests the effect of a weighted combination of variants. Simulation studies showed that TOW performed better than the existing methods across a wide range of scenarios. Aforementioned methods are for phenotypes at a single time point and cannot be applied to longitudinal phenotypes directly.

Meanwhile, quite a few statistical methods on the analysis of longitudinal data in the context of genetic mapping and association studies have been developed for common variants [6-10]. A typical method is functional mapping, which uses mathematical models to connect the actions of genes and the development of a trait. Several mathematical functions have been established to describe the development of a phenotype, including parametric functions [6], semiparametric functions [8], and nonparametric functions [9]. From a statistical standpoint, any modeling using longitudinal phenotypes is more informative than that using phenotypes at a single time point and thus can increase power to test association [7,10]. Functional mapping capitalizes on the full information provided by growth and development of phenotypes over time, increasing the power of gene

* Correspondence: shuzhang@mtu.edu
Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA

identification. However, no statistical methods on the analysis of longitudinal data are available for rare variants.

To analyze the sequencing data with phenotype measurements at multiple time points provided by Genetic Analysis Workshop 18 (GAW18) [11], in this article, we propose a novel method to test rare-variant association with longitudinal phenotypes by extending our previously proposed method, TOW. Applying the proposed method to the GAW18 data set, we compare the power of the proposed method with TOW using only one phenotype measurement.

Methods

Consider a random sample of n individuals. Each individual has been genotyped at M variants in a genomic region (a gene or a pathway). Denote $(x_{i1}, ..., x_{iM})$ as the genotypic score of the ith individual, where $x_{im} \in \{0, 1, 2\}$ is the number of minor alleles. Let $x_i = \sum_{m=1}^{M} w_m x_{im}$ denote the weighted combination of genotypic scores at the M variants, where $w_1, \ldots, w_M$ are unknown constants and their values are determined later using some optimal criteria. For longitudinal data, we assume that phenotypes and covariates are collected at K time points. Let y_{ij} and $z_{ij} = (z_{ij1}, \ldots, z_{ijP})^T$ denote the trait values and the covariates of the ith individual at the jth time point. For longitudinal data, we propose a mixed linear model to model the relationship between phenotype, covariates, and genotypic scores:

$$y_{ij} = Z_{ij}^T \alpha + \beta x_i + v_{ij} + e_{ij},$$

where $y = (y_{11}, ..., y_{1K}, \cdots, y_{n1}, ..., y_{nK})^T$, $x = (x_1, ..., x_1, \cdots, x_n, ..., x_n)^T$, v is the vector form of v_{ij}, and e is the vector form of e_{ij}. We assume that e follows normal distribution $N(0, \sigma_e^2 I)$ and v also follows normal distribution $N(0, \sigma_v^2 D)$, where $D = diag(D_0, \ldots, D_0)$ and D_0 depends on the level of correlation of phenotypes between time points. The total variance of y is $\Sigma = \sigma_v^2 D + \sigma_e^2 I$. Following Furlotte et al [10], we use sample correlation coefficients of phenotypes between time points to estimate D_0. For variance components σ_v^2 and σ_e^2, we use maximum likelihood estimates (MLEs) under null hypothesis $H_0 : \beta = 0$ as estimates of σ_v^2 and σ_e^2 and impute the estimated values of σ_v^2 and σ_e^2 into model (1). Let $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ denote the MLEs under null hypothesis of σ_v^2 and σ_e^2, and let $\hat{\Sigma} = \hat{\sigma}_v^2 D + \hat{\sigma}_e^2 I$. After imputing the estimated values of σ_v^2 and σ_e^2, model (1) becomes

$$y = Z\alpha + x\beta + \epsilon, \quad (2)$$

where ϵ follows $N(0, \hat{\Sigma})$.

Let $y_T = \hat{\Sigma}^{-1/2} y$, $x_T = \hat{\Sigma}^{-1/2} x$, and $Z_T = \hat{\Sigma}^{-1/2} Z$. Then model (2) is equivalent to

$$y_T = Z_T \alpha + x_T \beta + \epsilon_T, \quad (3)$$

where ε_T follows $N(0, I)$. The score test statistic under model (3) to test null hypothesis $H_0 : \beta = 0$ is given by $T_{score} = \frac{(y^{*T} x^*)^2}{\hat{\sigma}^2 x^{*T} x^*}$,

where y^* and x^* are the residuals under models $y_T = Z_T \alpha + \varepsilon_T$ and $x_T = Z_T \alpha + \varepsilon_T$, respectively, and $\hat{\sigma}^2 = \frac{1}{nK} y^{*T} y^*$. Let $X_m = (x_{1m}, ..., x_{1m}, \cdots, x_{nm}, ..., x_{nm})^T$, X_m^*, X_m^* is the residuals under the model $X_{mT} = Z_T \alpha + \varepsilon_T$, and $X^* = (X_1^*, \ldots, X_M^*)$. Then $T_{score} = \frac{w^T X^{*T} y^* y^{*T} X^* w}{\hat{\sigma}^2 w^T A w}$, where $A = X^{*T} X^*$.

One potential problem with the score test T_{score} is that for genotype data of rare variants, it will be problematic to use A to estimate the covariance matrix because of sparse data. Following Pan [12] and Sha et al [5], we replace A by $A_0 = diag(A)$. Then, the score test statistic is equivalent to $T_0(w) = \frac{w^T X^{*T} y^* y^{*T} X^* w}{w^T A_0 w}$.

As a function of w, $T_0(w)$ reaches its maximum when $w = w^o = A_0^{-1} X^{*T} y^*$ and the maximum value of $T_0(w)$ is $y^{*T} X^* A_0^{-1} X^{*T} y^*$. Based on longitudinal data, we define the statistic to test the effect of the optimally weighted combination (L-TOW) of variants, $\sum_{m=1}^{M} w_m^o x_{im}$, as $T_{L-TOW} = y^{*T} X^* A_0^{-1} X^{*T} y^* = \sum_{m=1}^{M} \frac{(y^{*T} X_m^*)^2}{X_m^{*T} X_m^*}$.

We use a permutation test to evaluate the p-value of T_{L-TOW}. In each permutation, we randomly shuffle the elements of y^*.

Results

We chose 157 genetically unrelated individuals from the file UNREL.txt. These individuals were extracted from 20 pedigrees in GAW18. We extracted genotypes for those individuals from files named chrN-dose.csv.gz. These files provided the estimated number of minor alleles carried for each variant. We used 200 replicates of simulated phenotype data in files PHEN.#.csv, where # is replicate number 1 to 200. Sex, age, medication use, and tobacco smoking were considered as covariates in this study. The phenotype data have been simulated at three time points with no missing data. There are 15 individuals without phenotype values in the simulated phenotype data, so the actual number of individuals used in this study is 142. To get reasonable powers for the power comparison, we merged 2 replicates to form a new replicate, so the total number of replicates for power comparison in this study was 100. We know the answers of the simulated data set in this study.

There are 2 related phenotypes, systolic blood pressure (SBP) and diastolic blood pressure (DBP) at three

time points. Based on the 2 related phenotypes, we consider 4 phenotype measurements: SBP, DBP, the first principal component of SBP and DBP, and the summation of SBP and DBP. For each phenotype measurement, we consider five tests: (a) L-TOW, which uses phenotype measurements at three time points; (b) TOW-1, TOW based on phenotype measurement at the first time point; (c) TOW-2, TOW based on phenotype measurement at the second time point; (d) TOW-3, TOW based on phenotype measurement at the third time point; and (e) TOW-Ave, TOW based on the average phenotype measurements over three time points. Based on each of the 4 phenotype measurements, we compare the power of L-TOW, TOW-Ave, and TOW-Single (average power of TOW-1, TOW-2, and TOW-3) to detect association between each of the top 17 genes that influence only DBP, only SBP, or both DBP and SBP. The power comparisons based on phenotype measurement DBP are given in Figure 1. This figure shows that in 15 of 17 genes, L-TOW is the most powerful test, TOW-Ave is the second most powerful test, and TOW-Single is the least powerful one. Power comparisons based on other three phenotype measurements show similar patterns. (Results are not showed.)

We also evaluated type I error rates of the proposed test, L-TOW. To evaluate the type I error we chose 200 blocks (100 variants in each block) from chromosome 21 that are far from causal variants. In each block, we applied L-TOW to each of the 100 replicates to test association between genotypes and the trait SBP. We obtained one p-value for each replicate and each block. The histogram of the 20,000 p-values is given in Figure 2. This figure shows that the distribution of p-values is very close to the uniform distribution, which indicates that L-TOW has correct type I error.

Discussion

We have developed TOW to detect association of rare and common variants [5]. Because the GAW18 data set provided phenotype measurements at multiple time points, similar to most of the existing methods for rare-variant association studies, TOW can only be applied to this data set by either using phenotype measurement at a single time point or using the average phenotype measurements over all time points. It is likely that a method jointly modeling phenotype measurements at all time points may increase power. This motivated us to extend our previously developed method, TOW, from phenotype measurement at one time point to phenotype measurements at multiple time points. By applying our newly developed method L-TOW to the GAW18 simulated data set, we showed that L-TOW has increased power over TOW by using either phenotype measurement at one time point or average phenotype measurements over multiple time points.

Although we describe our method using unrelated individuals, it is not difficult to extend the method to family-based data. For family data, denote $(x_{ij1}, ..., x_{ijM})$ as genotypic score of the jth member in the ith family and $x_{ij} = \sum_{m=1}^{M} w_m x_{ijm}$. Let y_{ijk} and $z_{ijk} = (z_{ijk1}, \ldots, z_{ijkP})^T$

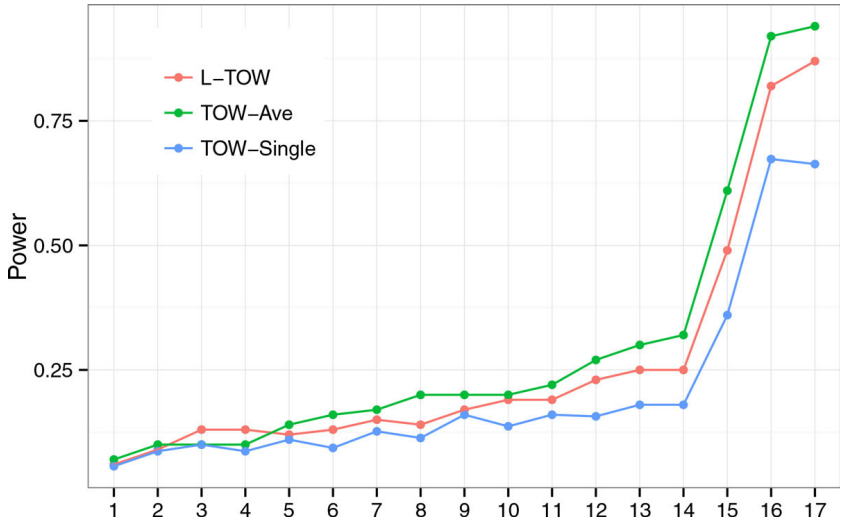

Figure 1 Power comparisons of the three tests using diastolic blood pressure as a phenotype measurement. Power of TOW-Single is the average power of TOW-1, TOW-2, and TOW-3. Numbers 1 to 17 on the x-axis refer to genes *ZNF443, ABTB1, FLNB, SLC35E2, TNN, CGN, ZFP37, LRP8, RAI1, ZNF544, LEPR, MTRR, NRF1, REPIN1, PTTG1IP, FLT3*, and *MAP4*, respectively. *TOW*, statistic to test the effect of the optimally weighted combination.

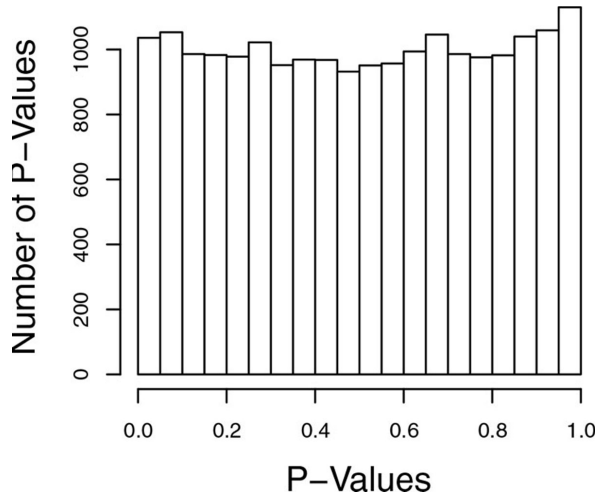

Figure 2 Histogram of *p*-values. Two hundred blocks (100 variants in each block) that are far from causal variants in chromosome 21 are chosen. In each block, the statistic to test the effect of the optimally weighted combination (L-TOW) is applied to each of the 100 replicates to test association between genotypes and the trait systolic blood pressure. One *p*-value is obtained for each replicate and each block.

denote the trait values and the covariates of the jth member in the ith family at the kth time point. For family data, we can use the following mixed linear model $y_{ijk} = Z_{ijk}^T\alpha + \beta x_{ij} + u_{ij} + v_{ijk} + e_{ijk}$,

where u_{ij} is a random variable modeling the correlation between family members, v_{ijk} is a random variable modeling the correlation of phenotype measurements between time points, and e_{ijk} is a random error term. Based on this model, using a similar argument to that in the Methods section, we can test association between the phenotype and the genomic region.

Comparing our method with functional mapping, whereas our proposed method uses age as a covariate and uses a single parameter β as the average effect over time of genotypes after adjusting for age effects, functional mapping uses mathematical models to connect gene actions and growth or development of a trait. Our proposed method has fewer parameters than the functional mapping method and uses less information. Our proposed method can easily incorporate the combination of rare variants. Incorporating the combination of rare variants to functional mapping requires further investigation.

Conclusions

We propose a novel method to test rare-variant association with longitudinal phenotypes by extending TOW, our previously proposed method. Application to the GAW18 data set shows that the newly proposed method jointly modeling phenotype measurements at all time points has increased power over TOW, which uses only one phenotype measurement.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SZ designed the overall study, SW and SF conducted statistical analyses, and QS and SZ drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Research reported in this article was supported by the National Human Genome Research Institute of the National Institutes of Health (NIH) under Award Number R03HG006155. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The GAW18 whole genome sequence data were provided by the T2D-GENES (Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples) Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. The GGAW is supported by NIH grant R01 GM031575.

This article has been published as part of *BMC Proceedings* Volume 8 Supplement 1, 2014: Genetic Analysis Workshop 18. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcproc/supplements/8/S1. Publication charges for this supplement were funded by the Texas Biomedical Research Institute.

Published: 17 June 2014

References

1. Li B, Leal SM: **Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.** *Am J Hum Genet* 2008, **83**:311-321.
2. Morgenthaler S, Thilly WG: **A strategy to discover genes that carry multiallelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST).** *Mutat Res* 2007, **615**:28-56.
3. Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet* 2009, **5**:e1000384.
4. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR: **Pooled association tests for rare variants in exon-resequencing studies.** *Am J Hum Genet* 2010, **86**:832-838.
5. Sha Q, Wang X, Wang X, Zhang S: **Detecting association of both rare and common variants by testing an optimally weighted combination of variants.** *Genetic Epidemiol* 2012, **36**:561-571.
6. Ma CX, Casella G, Wu RL: **Functional mapping of quantitative trait loci underlying the character process: a theoretical framework.** *Genetics* 2002, **161**:1751-1762.
7. Wu RL, Lin M: **Functional mapping–how to map and study the genetic architecture of dynamic complex traits.** *Nat Rev Genet* 2006, **7**:229-237.
8. Wu S, Yang J, Wu RL: **Semiparametric functional mapping of quantitative trait loci governing long-term HIV dynamics.** *Bioinformatics* 2007, **23**: i569-i576.
9. Das K, Li JH, Wang Z, Fu G, Li Y, Mauger D, Li R, Wu RL: **A dynamic model for genome-wide association studies.** *Hum Genet* 2011, **129**:629-639.
10. Furlotte N, Eskin E, Eyheramendy S: **Genome-wide association mapping with longitudinal data.** *Genetic Epidemiol* 2012, **36**:463-471.
11. Almasy L, Dyer TD, Peralta JM, Jun G, Fuchsberger C, Almeida MA, Kent JW Jr, Fowler S, Duggirala R, Blangero J: **Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees.** *BMC Proc* 2014, **8(suppl 2)**:S2.
12. Pan W: **Asymptotic tests of association with multiple SNPs in linkage disequilibrium.** *Genet Epidemiol* 2009, **33**:497-507.

doi:10.1186/1753-6561-8-S1-S91
Cite this article as: Wang *et al.*: Detecting association of rare and common variants by testing an optimally weighted combination of variants with longitudinal data. *BMC Proceedings* 2014 **8**(Suppl 1):S91.