The Author(s) *BMC Proceedings* 2016, **10**(Suppl 7):22
DOI 10.1186/s12919-016-0029-6

BMC Proceedings

PROCEEDINGS Open Access

A novel statistical method for rare-variant association studies in general pedigrees

Huanhuan Zhu[1], Zhenchuan Wang[1], Xuexia Wang[2] and Qiuying Sha[1*]

From Genetic Analysis Workshop 19
Vienna, Austria. 24-26 August 2014

Abstract

Both population-based and family-based designs are commonly used in genetic association studies to identify rare variants that underlie complex diseases. For any type of study design, the statistical power will be improved if rare variants can be enriched in the samples. Family-based designs, with ascertainment based on phenotype, may enrich the sample for causal rare variants and thus can be more powerful than population-based designs. Therefore, it is important to develop family-based statistical methods that can account for ascertainment. In this paper, we develop a novel statistical method for rare-variant association studies in general pedigrees for quantitative traits. This method uses a retrospective view that treats the traits as fixed and the genotypes as random, which allows us to account for complex and undefined ascertainment of families. We then apply the newly developed method to the Genetic Analysis Workshop 19 data set and compare the power of the new method with two other methods for general pedigrees. The results show that the newly proposed method increases power in most of the cases we consider, more than the other two methods.

Background

There is increasing interest in detecting associations between rare variants and complex traits. Although statistical methods to detect common variant associations are well developed, these variant-by-variant methods may not be optimal for detecting associations with 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 were developed for population-based designs, including the cohort allelic sums test [2], the combined multivariate and collapsing method [1], the weighted sum statistic [3], the variable minor allele frequency threshold method [4], the adaptive sum test [5], the step-up method [6], the sequence kernel association test [7], and the test for optimally weighted combination of variants [8].

Meanwhile, quite a few statistical methods for rare-variant association studies have been developed for family-based designs. For any type of study design, the statistical power will be improved if rare variants can be enriched in the samples. If one parent has a copy of a rare allele, half of the offspring are expected to carry it, and, hence, variants that are rare in the general population could be very common in certain families [9]. Therefore, family-based designs may play an important role in rare-variant association studies. Because of the importance of family-based designs in rare-variant association studies, several family-based rare-variant association methods for quantitative traits [10–12] and for qualitative traits [13–15] have been developed. However, most of these methods were developed under the assumption of random ascertainment and family-based designs with random ascertainment may not yield enrichment of rare variants. To analyze the sequencing data in general pedigrees provided by Genetic Analysis Workshop 19 (GAW19), we proposed a novel method to test rare-variant association in general pedigrees for quantitative traits. Applying the proposed method to the GAW19 data set, we compared the power of the proposed method with that of two popular methods for family-based designs.

* Correspondence: qsha@mtu.edu
[1]Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA
Full list of author information is available at the end of the article

Methods

Consider a sample of n pedigrees with n_i members in the i^{th} pedigree and a genomic region with M variants. Let y_{ij} and $g_{ij} = (g_{ij1}, \ldots, g_{ijM})^T$ denote the trait value and genotypes of the M variants in the genomic region for the j^{th} individual in the i^{th} pedigree. Let $x_{ij} = \Sigma_{m=1}^{M} w_m g_{ijm}$ denote the weighted combination of genotypes at the M variants, where $w = (w_1, \ldots, w_M)^T$ is a weight function.

For given genotypes, we assume that $y_{ij} \sim N(a + x_{ij}\beta, \sigma^2)$. Using the notation $g_i = \left(g_{i1}, \ldots, g_{in_i}\right)^T$, the retrospective likelihood is given by

$$\begin{aligned} RL &= \prod_{i=1}^{n} \Pr\left(g_i \middle| y_{i1}, \ldots, y_{in_i}\right) \\ &= \prod_{i=1}^{n} \frac{\Pr\left(y_{i1}, \ldots, y_{in_i} \middle| g_i\right) \Pr\left(g_i\right)}{\sum_{g_i^*} \Pr\left(y_{i1}, \ldots, y_{in_i} \middle| g_i^*\right) \Pr\left(g_i^*\right)} \\ &= \prod_{i=1}^{n} \frac{\exp\left(-\sum_{j=1}^{n_i} \left(y_{ij} - a - x_{ij}\beta\right) 2/2\sigma^2\right) \Pr\left(g_i\right)}{\sum_{g_i^*} \exp\left(-\sum_{j=1}^{n_i} \left(y_{ij} - a - x_{ij}^*\beta\right)^2 /2\sigma^2\right) \Pr\left(g_i^*\right)}, \end{aligned}$$

where $\sum_{g_i^*}$ represents the summation of all possible genotypes. Based on RL, the score test statistic for testing the null hypothesis $H_0 : \beta = 0$ is given by

$$T_{score} = U^2/V \tag{1}$$

where $U = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}\right)\left(y_{ij} - \overline{y}\right)$, $V = w^T \Sigma w \Sigma_{i=1}^{n} y_i^T \Phi_i y_i$, $y_i = \left(y_{i1}, \ldots, y_{in_i}\right)^T$, $\overline{y} = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \sum_{j=1}^{n_i} y_{ij}$, Φ_i is twice the kinship coefficient of the i^{th} pedigree, and $\Sigma = \text{cov}(g_{11}, g_{11})$ is the covariance matrix of the multiple variant genotype of one individual. Σ can be estimated by $\hat{\Sigma} = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(g_{ij} - \overline{g}\right)\left(g_{ij} - \overline{g}\right)^T$, where $\overline{g} = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \sum_{j=1}^{n_i} g_{ij}$. It is worth pointing out that T_{score} is equivalent to the quantitative version of the retrospective likelihood score statistic proposed by Schaid et al [16].

Because rare variants are essentially independent, following Pan [17] and Sha et al [8], we replace $\hat{\Sigma}$ by $\hat{\Sigma}_0 = diag\left(\hat{\Sigma}\right)$. Then, the score test statistic T_{score} becomes

$$T_0(w) = w^T u u^T w / \left(w^T \hat{\Sigma}_0 w \sum_{i=1}^{n} y_i^T \Phi_i y_i\right),$$

where $u = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(g_{ij} - \overline{g}\right)\left(y_{ij} - \overline{y}\right)$. As a function of w, $T_0(w)$ reaches its maximum when $w = \hat{\Sigma}_0^{-1} u$ and the maximum value of $T_0(w)$ is $u^T \hat{\Sigma}_0^{-1} u / \sum_{i=1}^{n} y_i^T \Phi_i y_i$. We define the statistic of optimally weighted score test (OW-score) as

$$T_{OW-score} = u^T \hat{\Sigma}_0^{-1} u / \sum_{i=1}^{n} y_i^T \Phi_i y_i = \sum_{m=1}^{M} \left(u_m^2 / \sigma_{mm}\right) / \left(\sum_{i=1}^{n} y_i^T \Phi_i y_i\right),$$

where σ_{mm} is the (m, m)th element of $\hat{\Sigma}_0$ and u_m is the m^{th} element of u. Under the null hypothesis, $T_{OW\text{-}score}$ is asymptotically distributed as a mixture of independent χ^2 statistics [18, 19]. Alternatively, the distribution of $T_{OW\text{-}score}$ can be approximated by a Satterwaite approximation for the distribution of quadratic forms [7, 20, 21] or a scaled χ^2 distribution [16]. We propose to approximate the distribution of $T_{OW\text{-}score}$ by a scaled χ^2 distribution with the scale δ and degrees of freedom d estimated by the expectation and variance of $T_{OW\text{-}score}$. Note that $u \sim N(0, \Sigma \Sigma_{i=1}^{n} y_i^T \Phi_i y_i)$. We have $\hat{\mu}_T = \hat{E}(T_{OW-score}) = trace\left(\hat{\Sigma}\hat{\Sigma}_0^{-1}\right)$ and $\hat{\sigma}_T^2 = \text{va}\hat{r}(T_{OW-score}) = 2trace\left(\hat{\Sigma}\hat{\Sigma}_0^{-1}\hat{\Sigma}\hat{\Sigma}_0^{-1}\right)$. Then, the scale δ is estimated as $\hat{\delta} = \hat{\sigma}_T^2/(2\hat{\mu}_T)$ and the degree of freedom d is estimated as $\hat{d} = 2\hat{\mu}_T^2/\hat{\sigma}_T^2$

We compare the performance of our OW-score with (a) WS-score, the score test given by equation (1) with weight given by Madsen and Browning [3] and (b) famSKAT, family-based sequence kernel association test given by Chen et al [11].

Results

We applied our proposed method as well as the WS-score test and famSKAT to the simulated data from GAW19. All tests were conducted on 849 individuals, from 20 pedigrees, that had no missing genotypes or phenotypes. Sex, age, blood pressure medication status, and smoking status were considered as covariates in this study. We were aware of the underlying simulation model.

There are two related phenotypes, systolic blood pressure (SBP) and diastolic blood pressure (DBP), at three time points. We considered the average of DBP at three time points as the phenotype of interest in our analysis. We compared the power of the three tests (OW-score, WS-score, and famSKAT) to detect association between each of the top 14 genes that influence the phenotype of interest. We used the variants between the first functional single nucleotide polymorphism (SNP) and the last functional SNP in each gene in our analysis. We did not consider *CABP2* because the power of the three tests are essentially the same due to only one variant in this gene. To adjust the effects of the covariates on the phenotype of interest, we first applied a linear model by regressing the phenotype of interest on the covariates: sex, the average of age, the average of blood pressure medication status, and the average of smoking status. The power comparisons based on the 200 replicated data sets are given in Table 1. Significance level is assessed at 5 %. This table shows that the OW-score test identified three genes with power greater than 40 %, famSKAT identified 1 gene with power greater than 40 %, and the WS-score test could not identify any genes with power greater than 40 %. OW-score and famSKAT have different power mainly because they use different weights. Let w_m and W_m denote the weights, rescaled to the interval (0, 1), of the OW-score test and famSKAT for the m^{th} variant. Then, $w_m > W_m$ when minor allele frequency (MAF) is less than 0.01; $w_m \le W_m$ when MAF is in the interval (0.01, 0.05); $w_m > W_m$ when MAF is greater than 0.05. The OW-score test has much higher power than famSKAT for *RAI1* and *REPIN1* because none of the MAFs of the causal variants in *RAI1* and *REPIN1* are in the interval (0.01, 0.05).

We also evaluated the type I error rate of the proposed OW-score test. To evaluate the type I error, we used 1000 blocks (100 variants in each block) from chromosome 5 that are far from causal variants. In each block, we applied the OW-score test to each of the 200 replicates to test association between genotypes and the phenotype of interest. We obtained 1 p value for each replicate and each block. The type I errors of the proposed test were 0.04887, 0.00921, and 0.00131 at significance levels of 0.05, 0.01, and 0.001, respectively. We also considered the average of SBP at three time points as the phenotype of interest, which yielded similar results.

Table 1 Power comparisons of the 3 tests using the average of DBP at 3 time points as phenotypes (significance level is assessed at 5 %)

Genes	$T_{OW\text{-}score}$	$T_{WS\text{-}score}$	FamSKAT
CGN	0.135	0	0.035
FLT3	0.005	0	0.08
LEPR	0.05	0.015	0.065
MAP4	0.175	0.185	**0.425**
MTRR	**0.465**	0.005	0.06
NRF1	0	0.005	0.035
PTTG1IP	0.02	0.145	0.06
RAI1	**0.845**	0.005	0.155
REPIN1	**0.915**	0.05	0.085
SLC35E2	0.005	0	0.05
TNN	0	0	0.035
ZFP37	0	0.005	0.005
ZNF443	0.01	0.015	0.195
ZNF544	0.005	0.015	0.06

Notes: the powers greater than 40 % are in bold

Discussion

Next-generation sequencing technologies make directly testing rare variant association possible. However, the development of powerful statistical methods for rare-variant association studies is still underway. In this article, we proposed a novel statistical method for rare-variant association studies based on general pedigrees for quantitative traits. The application to the GAW19 data set showed that the proposed method has correct type I error rate and is more powerful than the other two methods against which our method was compared.

We described our method for quantitative traits. For qualitative traits, we can derive a score test similar to that given by equation (1). However, the performance of the proposed method for qualitative traits requires further investigation. Like many statistical methods for rare variant association studies, the proposed method can consider phenotype measurement at only one time point. Statistical methods based on sequence data have been developed for unrelated individuals that have phenotype measurements at multiple time points [22]. From a statistical standpoint, modeling using longitudinal phenotypes is more informative than that using phenotypes at a single time point and thus can increase the power of an association test [22, 23]. Our future work includes extension of the proposed method to longitudinal phenotypes.

Conclusions

In this article, we developed a novel statistical method for rare variant association studies in general pedigrees (randomly ascertained pedigrees or ascertained pedigrees). Application to the GAW19 data set showed that the newly proposed method is more powerful than the other two methods in most of the cases. Our new method uses a retrospective view, which allows us to account for complex and undefined ascertainment of families. The GAW19 data is based on randomly ascertained pedigrees. Results of applying our method to GAW19 data showed that the proposed method has correct type I error based on random ascertainment. When random ascertainment is violated and ascertainment is based on trait values, the proposed method is expected to have correct type I error. If pedigrees are ascertained because of extreme trait values, the proposed method is expected to have higher power than methods based on randomly ascertained pedigrees.

Acknowledgements

The GAW19 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 National Institutes of Health (NIH) grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW19 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 GAW is supported by NIH grant R01 GM031575.

Declarations
This article has been published as part of *BMC Proceedings* Volume 10 Supplement 7, 2016: Genetic Analysis Workshop 19: Sequence, Blood Pressure and Expression Data. Summary articles. The full contents of the supplement are available online at http://bmcproc.biomedcentral.com/articles/supplements/volume-10-supplement-7. Publication of the proceedings of Genetic Analysis Workshop 19 was supported by National Institutes of Health grant R01 GM031575.

Authors' contributions
QS designed the overall study, HZ and ZW conducted statistical analyses, and HZ, XW, and QS drafted the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA. [2]Department of Mathematics, University of North Texas, 1155 Union Circle #311430, Denton, TX 76203-5017, USA.

Published: 18 October 2016

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(3):311–21.
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(1-2):28–56.
3. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5(2):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(6):832–8.
5. 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.
6. Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PLoS One. 2010;5(11):e13584.
7. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare variant association testing for sequencing data with the sequence kernel association test (SKAT). Am J Hum Genet. 2011;89(1):82–93.
8. Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. Genet Epidemiol. 2012;36(6):561–71.
9. Shi G, Rao D. Optimum designs for next-generation sequencing to discover rare variants for common complex disease. Genet Epidemiol. 2011;35(6):572–9.
10. Liu D, Leal S. A unified framework for detecting rare variant quantitative trait associations in pedigree and unrelated individuals via sequence data. Hum Hered. 2012;73(2):105–22.
11. Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol. 2013;37(2):196–204.
12. Svishcheva GR, Belonogova NM, Axenovich TI. FFBSKAT: fast family-based sequence kernel association test. PLoS One. 2014;9(6):e99407.
13. Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. Genet Epidemiol. 2010;34(2):171–87.
14. Feng T, Elston R, Zhu X. Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). Genet Epidemiol. 2011;35(5):398–409.
15. Zhu Y, Xiong M. Family-based association studies for next-generation sequencing. Am J Hum Genet. 2012;90(6):1028–45.
16. Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol. 2013;37(5):409–18.
17. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol. 2009;33(6):497–507.
18. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007;63(4):1079–88.
19. Liu H, Tang Y, Zhang H. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Comput Stat Data Anal. 2009;53:853–6.
20. Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multi locus association test for quantitative traits. Am J Hum Genet. 2008;82(2):386–97.
21. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics. 2008;9:292.
22. Wang S, Fang S, Sha Q, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants with longitudinal data. BMC Proc. 2014;8 Suppl 1:S91.
23. Furlotte N, Eskin E, Eyheramendy S. Genome-wide association mapping with longitudinal data. Genet Epidemiol. 2012;36(5):463–71.