Comparison of several sequence-based association methods in pedigrees

George Mathew  
*Missouri State University*

Varghese George

Hongyan Xu

Follow this and additional works at: [https://bearworks.missouristate.edu/articles-cnas](https://bearworks.missouristate.edu/articles-cnas)

**Recommended Citation**

This article or document was made available through BearWorks, the institutional repository of Missouri State University. The work contained in it may be protected by copyright and require permission of the copyright holder for reuse or redistribution.
For more information, please contact bearworks@missouristate.edu.
Comparison of several sequence-based association methods in pedigrees

George Mathew1*, Varghese George2, Hongyan Xu2

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

Abstract
Genome-wide association studies are very powerful in determining the genetic variants affecting complex diseases. Most of the available methods are very useful in detecting association between common variants and complex diseases. Recently, methods to detect rare variants in association with complex diseases have been developed with the increasingly available sequencing data from next-generation sequencing. In this paper, we evaluate and compare several of these recent methods for performing statistical association using whole genome sequencing data in pedigrees. Specifically, functional principal component analysis (FPCA), extended combined multivariate and collapsing (CMC) method for families, a generalized $T^2$ method, and chi-square minimum approach were compared by analyzing all the genetic variants, common and rare, of both the real data set and the simulated data set provided as part of Genetic Analysis Workshop 18.

Background
With advances in genotyping technologies, genome-wide association studies (GWAS) became a very popular procedure to identify disease genes and other traits by conducting statistical tests on many thousands of single-nucleotide polymorphisms (SNPs). The procedure has great potential for discovering genetic variants influencing complex diseases. However, these procedures have discovered loci that account only for a small percentage of phenotypic variance [1]. One of the reasons for this difficulty may be that rare variants might explain disease susceptibility [2-4]. Recently, several methods have been developed to determine the influence of rare variants on complex diseases. These methods differ from the traditional methods of testing where the focus has been on individual common variants. It is understood that those variants with a population frequency greater than 5% are considered to be common variants, those with less than 1% population frequency as rare variants, and the rest as low-frequency variants [4]. The common variants are believed to be from distant ancestors, whereas rare variants are from recent ancestors [5]. Most of these methods assume the individuals are independent and are designed for population-based data. Only recently have several methods been developed that can perform statistical association of sequence data in pedigrees. In this paper, we used functional principal component analysis (FPCA) [4], the generalized $T^2$ approach [4], the combined multivariate and collapsing (CMC) test for family data [2,4], and the chi-square minimum approach for family data [4] to analyze association of the dichotomous hypertension trait with all genetic variants, common and rare, of the real data set and all replicates of chromosome 3 of the simulated data set provided by Genetic Analysis Workshop 18 (GAW 18) [6]. We compared the results to assess the merits of these methods.

Methods
An extension of the generalized $T^2$ test [7] for family-based association studies is provided by Zhu and Xiong [4]. The test statistic is given by $T^2 = \frac{T^2}{P_{corr}}$, where $T^2$ is the generalized $T^2$ statistic [7], and $P_{corr}$ [4, p. 1030] is the correction factor to account for the familial correlation in the pedigree data. A similar extension of the CMC test is also developed and is provided by equation

* Correspondence: georgemathew@missouristate.edu
1 Department of Mathematics, Missouri State University, 901 South National Avenue, Springfield, Missouri 65897, USA
Full list of author information is available at the end of the article

© 2014 Mathew et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
FPCA method finds the fewest number of significant genes compared to the other 3 methods. Chi_min finds the highest number of significant genes at 0.05 and 0.01 levels. However, T^2 finds more significant genes at lower significance levels (0.001, 0.0001, and 4.7 × 10^{-6}). The number of genessignificant at the 4.7 × 10^{-6} level by the FPCA, Chi_min, CMC, and T^2 methods are 0, 15, 598, and 1794, respectively. Figure 1 is a Venn diagram showing overlaps of the significant genes from Chi_min, CMC, and T^2 at 4.7 × 10^{-6} level. It is interesting to note that all 598 significant genes found by CMC overlap with those found by T^2.

The number of significant genes presented in Table 1 will contain false positives, as with any statistical test. To get an idea of the number of “true findings,” we compared our results with those findings of GWAS for blood-pressure-associated genes. We performed a comprehensive literature review, and 84 genes were identified as being associated with blood pressure from GWAS. Table 2 shows the number of overlapped genes between our analysis and the GWAS findings.

We analyzed chromosome 3 of the simulated data set. There are a total of 1120 genes on chromosome 3, of which 30 were used for causal variants of hypertension in the simulation model. The remaining 1090 were assumed to be unrelated to the disease and are used only for calculating type I error rate. The linkage disequilibrium (LD) between the genetic variants from these groups of 1090 genes and 30 genes were analyzed with Haploview [9] and no significant LD was found. Table 3 lists the type I error rates from the analysis of all 200 replicates by all 4 methods at various significance levels.

The analysis of the 30 positive genes is used to calculate the power of the various methods. Table 4 lists the estimates of the power by the various methods.

Discussion
With the increasingly available sequence data from the next-generation sequencing technologies, it is important for a statistical association method to handle both common and rare genetic variants. It is also important for these methods to handle data from pedigrees because rare genetic variants are enriched in families with multiple affected individuals, which could confer more statistical power. From our analysis of the real data, T^2 seems to be a better method than the other 3 methods because it finds more significant genes at low significance levels. At the Bonferroni corrected p-value of 4.7 × 10^{-6}, T^2 identified the genes CASZ1, ADAMTS8, NUCB2, ABCC8, SLC4A7, MAP4, CASR, EBF1, PLEKHA7, SOX6, ULLK4, and MECOM. The last 4 genes were also identified by the CMC method. All the genes mentioned above were found to be associated with blood pressure, in particular ULLK4 and PLEKHA7 by Levy et al [10], and MAP4 by Wain et al [11].
As with GWAS, we need to keep a low significance level to account for multiple testing. We note from the analysis of the simulated data that FPCA has empirical type I error rate much less than the nominal value, making it very conservative. The Chi_min method has inflated type I error rate. The type I error rates by $T^2$ and CMC are close to the nominal value. Also, $T^2$ has better power than CMC, which is consistent with the result from the real data.

From the analysis of the data sets we find that $T^2$ is a better method, which is different from the findings of Zhu and Xiong [4], which suggest that FPCA is a better procedure. There are 2 possible reasons why FPCA performs less well here. One reason may be that the SNPs in the genes are sparse. If there are too few SNPs in 1 gene, the FPCA may not perform well because the number of SNPs is not enough to estimate the function describing the allele counts across the SNPs in the gene. A second reason may be that the assumption of a smooth function of the allele counts across the SNPs for the FPCA may not hold for the GAW 18 data sets. We observed a large overlap between the results of CMC and $T^2$. This mainly comes from the fact that CMC uses the $T^2$ approach with common variants. There is also a tendency to pick up more genes with more variants for both CMC and $T^2$ methods.

**Conclusions**

From the analysis results of both real and simulated data, $T^2$ is a preferable method for pedigree-based association studies with whole-genome sequencing data because it controls the false positive rate and is more powerful than the other two methods with similar type I error rates.
Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
HX conceived of the study, performed the analysis, and helped to draft the manuscript. GM performed the analysis and helped to draft the manuscript. VG participated in the design and coordination of the study. All authors read and approved the manuscript.

Acknowledgements
The GAW18 whole genome sequence data were provided by the T2D-GENES 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 R01 HL045222, R01 DK047482, and R01 DK053889. The Genetic Analysis Workshop 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

doi:10.1186/1753-6561-8-S1-S48
Cite this article as: Mathew et al.: Comparison of several sequence-based association methods in pedigrees. BMC Proceedings 2014 8(Suppl 1):S48.

Submit your next manuscript to BioMed Central and take full advantage of:
• Convenient online submission
• Thorough peer review
• No space constraints or color figure charges
• Immediate publication on acceptance
• Inclusion in PubMed, CAS, Scopus and Google Scholar
• Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit