BMC Genetics

RESEARCH Open Access

An adaptive gene-level association test for pedigree data

Jun Young Park, Chong Wu and Wei Pan*

From Genetic Analysis Workshop 20
San Diego, CA, USA. 4-8 March 2017

Abstract

Background: We propose a gene-level association test that accounts for individual relatedness and population structures in pedigree data in the framework of linear mixed models (LMMs). Our method data-adaptively combines the results across a class of score-based tests, only requiring fitting a single null model (under the null hypothesis) for the whole genome, thereby being computationally efficient.

Results: We applied our approach to test for association with the high-density lipoprotein (HDL) ratio of post- and pretreatments in GAW20 data. Using the LMM similar to that used by Aslibekyan et al. (PLos One, 7:48663, 2012), our method identified 2 nearly significant genes (*APOA5* and *ZNF259*) near rs964184, whereas neither the other gene-level tests nor the standard test on each individual single-nucleotide polymorphism (SNP) detected any significant gene in a genome-wide scan.

Conclusions: Gene-level association testing can be a complementary approach to the SNP-level association testing and our method is adaptive and efficient compared to several other existing gene-level association tests.

Keywords: aSPU, GWAS, HDL, Linear mixed models, Score test

Background

Genome-wide association studies (GWASs) are considered to be the standard approach to use to detect common genetic variants associated with complex traits. It has become popular to extend the most popular single-nucleotide polymorphism (SNP)-level analysis to gene-level analysis by aggregating multiple SNPs in a gene or other functional unit. As a complement to the standard single SNP-based approach, the gene-level approach can achieve higher reproducibility and power. An additional benefit of the gene-level approach is that a decreased number of hypotheses need to be tested, thereby reducing the burden of multiple testing.

The goal of this work is to perform a gene-level association test to detect genes significantly associated with a single trait using the GAW20 data while effectively controlling for the false-positive rate. Note that the candidate gene approach conducted by Aslibekyan et al. was based on the 95 loci drawn from previous studies based on SNP-level association testing [1], and found SNP rs964184 to be strongly associated with the high-density lipoprotein (HDL) ratio of post- and pretreatments. We are interested in determining whether a gene-level analysis can lead to uncovering significantly associated genes, and, in particular, whether the genes near rs964184 are significantly associated in a genome-wide scan. Specifically, we apply the adaptive sum of powered score (aSPU) test [2], which is motivated to account for unknown and varying association patterns (eg, varying numbers or proportions of associated SNPs) across the genes, thus maintaining higher power than other nonadaptive gene-level tests. The aSPU test is computationally feasible as it does not require to fit separate models for each SNP or gene, and it satisfactorily controls false-positive rates. Note that the aSPU test was originally proposed for generalized linear models, and extended to generalized estimating equations and generalized linear mixed models (GLMM) [3–5]. Its application to and empirical performance in linear mixed

* Correspondence: weip@biostat.umn.edu
Division of Biostatistics, University of Minnesota, 420 Delaware Street SE, Minneapolis, MN 55455, USA

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. 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.

models (LMMs), especially with large pedigree data, have not been discussed in previous studies.

The Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study collected pedigree data, motivating the use of LMMs to account for population structures and relatedness as adopted by Aslibekyan et al. [1]. In our LMM, we account for genetic relatedness among subjects as a random effect with a covariance matrix calculated based on individual-level SNP data. We also adjusted for covariates such as age gender, and study center. In this paper, we present the results of the aSPU test based on LMM and compare with other existing gene-level tests and individual SNP analysis.

Methods

Suppose that y_i denotes a quantitative trait for individual $i = 1, \cdots, n$, $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{iq})'$ is a vector of q covariates, and $\boldsymbol{G}_i = (G_{i1}, \cdots, G_{ip})'$ is a vector of p SNPs in a gene for individual i. A LMM is constructed as

$$y_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{G}_i\boldsymbol{\beta} + b_i + \varepsilon_i \tag{1}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the unknown regression coefficient vectors for the corresponding covariates and SNPs, b_i and ε_i are a random intercept and an error term that are independent with each other. We further assume that the error terms ε_is are independently distributed, but b_is are not. Specifically,

$$\begin{aligned}\boldsymbol{b} &= (b_1, \ldots, b_n)' \sim \mathcal{N}(\boldsymbol{0}, \tau \cdot \boldsymbol{\Psi}) \text{and } \boldsymbol{\varepsilon} \\ &= (\varepsilon_1 \ldots, \varepsilon_n)' \sim \mathcal{N}(\boldsymbol{0}, \phi \cdot \boldsymbol{I})\end{aligned} \tag{2}$$

where $\boldsymbol{\Psi}$ is a known $n \times n$ genetic relationship matrix, which reflects the genetic relatedness among the subjects in the data. The null hypothesis to be tested for association between the group of the SNPs and the trait is $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$.

Fitting (generalized) LMMs can be computationally demanding. However, using penalized quasi-likelihood (PQL) to fit the model enables us to extract the test statistic for score-based tests including the aSPU test [6]. It is known that maximizing PQL is equivalent to maximizing the likelihood for quantitative traits. Specifically, we first need to fit the LMM under the null hypothesis.

$$y_i = \boldsymbol{X}_i\boldsymbol{\alpha} + b_i + \varepsilon_i, \tag{3}$$

from which, the score vector $\boldsymbol{U} = (U_1, \cdots, U_p)'$, to be used to construct various gene-level score-based tests, can be expressed as

$$U_j = \sum\nolimits_{i=1}^{n} G_{ij}\left(\frac{y_i - (\boldsymbol{X}_i\hat{\boldsymbol{\alpha}} - \hat{b}_i)}{\hat{\phi}}\right) \tag{4}$$

The aSPU test statistic can be obtained using the score vector $\boldsymbol{U}$ and its covariance matrix $\boldsymbol{V}$ under the null hypothesis, which can also be written in a closed form. Because the score vector follows asymptotic normal distribution with mean zero under the null hypothesis, one can use the Monte Carlo method to compute p-values. Note that both $\boldsymbol{U}$ and $\boldsymbol{V}$ depend only on the null model (3), which provides computational efficiency when the number of tests is large as in a genome-wide scan. We can use an R package *GMMAT* to derive $\boldsymbol{U}$ and $\boldsymbol{V}$ [7].

We briefly introduce the idea of the aSPU test here. All score-based association tests require $\boldsymbol{U}$ and $\boldsymbol{V}$, and each nonadaptive test has its own advantages and disadvantages. For example, consider these 2 cases: (a) every SNP encoded in a gene is associated with an equal effect size and direction, and (b) only one or a small proportion of the SNPs are associated. The burden test, which takes $\sum_{j=1}^{p} U_j$ as a test statistic, is desired in the first case, but it will lose power in the second case. On the other hand, the UminP test, which takes $\max\{|U_1|, \cdots, |U_p|\}$ as a test statistic when the variances of the score elements are the same, is advantageous in the second case but not in the first case. Thus, applying a single and nonadaptive score-based test might not be powerful in gene-level analysis. The aSPU test offers a way to combine various score-based tests; it is based on a class of the sum of powered score (SPU) tests indexed by a positive integer γ. Specifically, the SPU(γ) test statistic is.

$$T_{SPU(\gamma)} = \sum\nolimits_{j=1}^{p} U_j^{\gamma} \text{and } T_{SPU(\infty)} = \max\left\{|U_1|, \ldots, |U_p|\right\} \tag{5}$$

It is easy to see that the burden test and the sum of squared score (SSU) test are equivalent to the SPU(1) and SPU(2) tests respectively. It was also shown that SPU(2) is equivalent to sequence kernel association test (SKAT) with the linear kernel and to Multivariate Distance Matrix Regression (MDMR) with the Euclidean distance (under the framework of LMM) [8]. Furthermore, assuming the equal variance of the score elements, the UminP test is equal to SPU test with $\gamma = \infty$. One can treat γ as a factor that decides the weight on each score element. The aSPU test uses the minimum p value of the SPU tests as the test statistic, which provides a general data-adaptive method to test for associations. The set of $\gamma \in \{1, 2, \cdots, 8, \infty\}$ was proposed by Pan et al. based on experiences [2].

Results

The LMM we used for the GAW20 data was similar to that used by Aslibekyan et al.; we used the ratio of post- and pretreatment HDL as the trait, and we used age, gender, and study center as covariates. The only difference was the covariance matrix of the random effects. Our covariance matrix $\boldsymbol{\Psi}$ of the random effects reflected the genetic relatedness, where each $\boldsymbol{\Psi}_{ij}$ was the Pearson correlation coefficient between 2 subjects i and j of 20,000 randomly selected SNPs. Our analysis was based on 821 subjects who did not have missing values in either the trait or the covariates. We only included common variants with minor allele frequencies (MAFs) greater than 0.05. Among those, we randomly imputed missing variants using MAF if the proportions of missing values were less than 1%. It resulted in a total of 595,304 SNPs included in our analysis. For the gene-level analysis, we used hg18 as a reference genome and each gene included the SNPs that were within 10,000 regions upstream or downstream of the gene's coding region. In total, we included 22,434 genes in our analysis.

We conducted the SPU(γ) and aSPU tests under the LMM. In addition to the SPU(1), SPU(2), and SPU(∞) tests where their theoretical equivalences with other existing gene-level tests are shown in the Methods section, we also performed the gene-level score test and the famSKAT (family-based sequence kernel association test) [9] using the same covariates and relationship matrix. Figure 1 shows the results of the tests. Using the Bonferroni adjustment for the genome-wide significance level ($\alpha = 0.05$), the aSPU test and the score test did not detect any significant genes, but 2 genes (*APOA4* and *ZNF259* on chromosome 11) were close to being significant. However, these 2 genes were detected by the SPU(1) test, suggesting that their association effects were not dominated by a small number of variants. We emphasize the adaptiveness property of the aSPU test by noting gene *BUD13* on chromosome 11 and *GUCD1* and *SNRPD3* on chromosome 22, whose $-\log_{10}$(p values) were not less than 3 by SPU(1), but much larger by the SPU(∞) test (as well as by a few other SPU tests and the

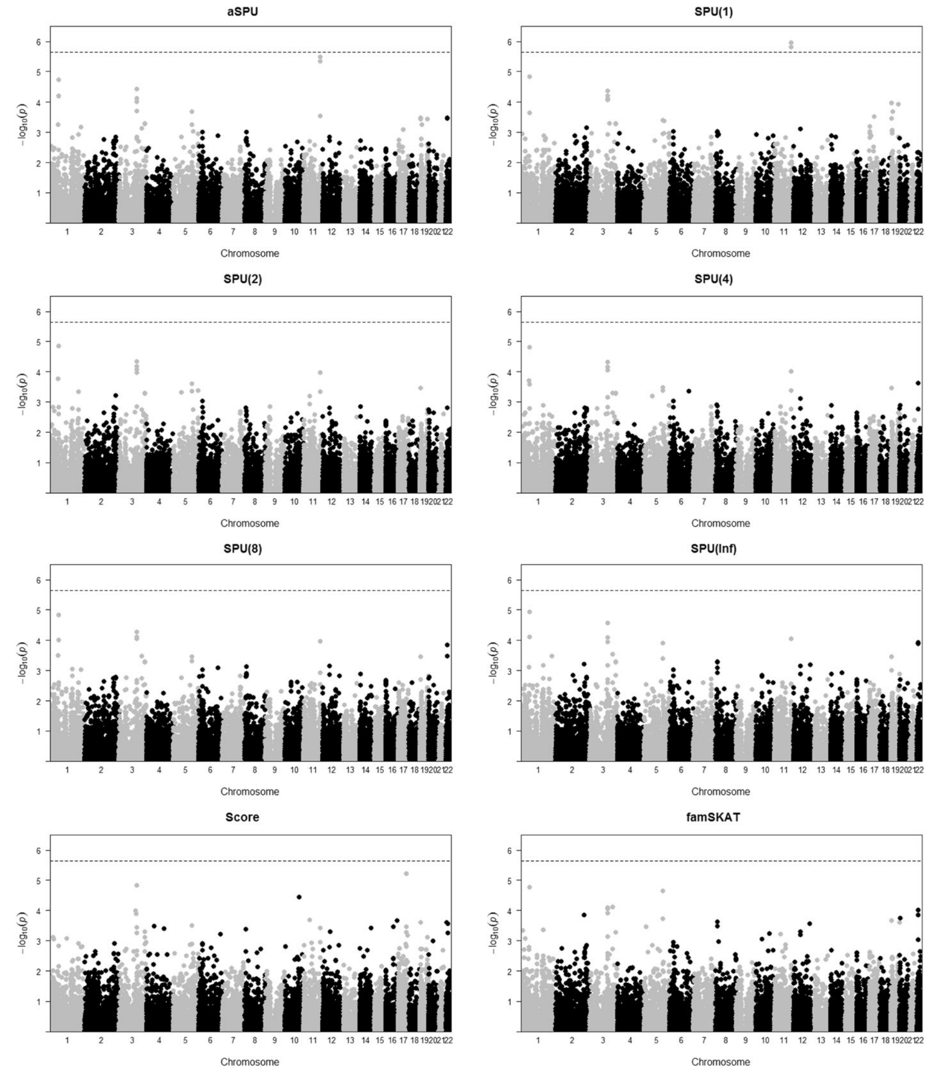

Fig. 1 The Manhattan plots for the gene-level association tests

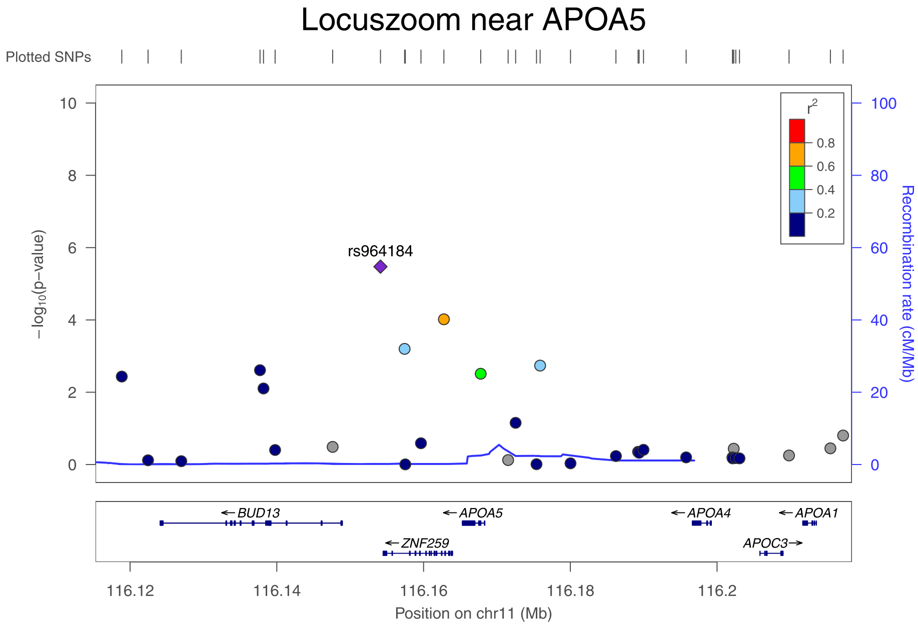

Fig. 2 The LocusZoom plot near *APOA5* gene

aSPU test). We also note that *APOA5* and *ZNF259* were located nearby as shown in Fig. 2. In particular, they shared 7 variants out of 9 SNPs in both genes. The gene-level score test yielded a gene (*DDX42* on chromosome 17) almost significant at the genome-wide significance level, but the score test did not detect any loci near rs964184. Similarly, the famSKAT did not detect any significant gene.

<insert Figure(s) 1 and 2 here>.

We also compared the gene-based tests to the score test for single variants. We used the usual 5×10^{-8} as the genome-wide significance level for the SNP-level analysis. Even though rs964184 turned out to be the one most significantly associated with the trait among all the SNPs, its p value was far away from the genome-wide significance level, as shown in Fig. 3. This example partially confirms the usefulness of gene-level testing.

Discussion

In GWAS, individuals in pedigree data are not independent, thus motivating the use of (generalized) LMMs. We considered a general LMM with a random intercept that reflects the genetic relatedness among the subjects. We then conducted the aSPU test on the genes across the whole genome based on fitting a single null model, and identified 2 genes near SNP rs964184 to be nearly significant. In contrast, none of the SNPs, including SNP rs964184, were nearly significant in a standard single SNP-based analysis.

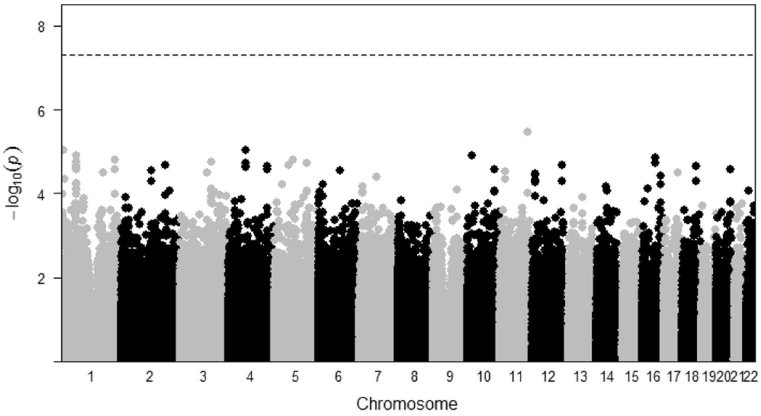

Fig. 3 The Manhattan plot for the SNP-level association test

Conclusions

We have demonstrated the applicability and usefulness of our proposed aSPU test in LMMs for association analysis of large pedigree data. Furthermore, our study has confirmed possible advantages and complementary roles of gene-level analyses with the adaptive aSPU test when compared to standard single SNP-based analyses.

Abbreviations

aSPU: Adaptive sum of powered score; GLMM: Generalized linear mixed model; GWAS: Genome-wide association study; LMM: Linear mixed model; MAF: Minor allele frequency; SNP: Single nucleotide polymorphisms; SPU: Sum of powered score.

Acknowledgements

We thank the reviewers for many helpful and constructive comments and the organizers of Genetic Analysis Workshop 20. This research was supported by the Minnesota Supercomputing Institute..

Funding

Publication of this article was supported by NIH R01 GM031575. This research was funded by NIH grants R21AG057038, R01HL116720, R01GM113250, and R01HL105397. CW was funded by the University of Minnesota Doctoral Dissertation Fellowship.

Availability of data and materials

The data that support the findings of this study are available from the Genetic Analysis Workshop (GAW), but restrictions apply to the availability of these data, which were used under license for the current study. Qualified researchers may request these data directly from GAW.

About this supplement

This article has been published as part of *BMC Genetics* Volume 19 Supplement 1, 2018: Genetic Analysis Workshop 20: envisioning the future of statistical genetics by exploring methods for epigenetic and pharmacogenomic data. The full contents of the supplement are available online at https://bmcgenet.biomedcentral.com/articles/supplements/volume-19-supplement-1

Authors' contributions

JYP, CW, and WP designed the study. JYP and CW performed the data analysis. JYP drafted the manuscript. WP helped revise the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published: 17 September 2018

References

1. Aslibekyan S, Goodarzi MO, Frazier-Wood AC, Yan X, Irvin MR, Kim E, Tiwari HK, Guo X, Straka RJ, Taylor KD, et al. Variants identified in a GWAS meta-analysis for blood lipids are associated with the lipid response to fenofibrate. PLoS One. 2012;7(10):48663.
2. Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. Genetics. 2014;197(4):1081–95.
3. Zhang Y, Xu Z, Shen X, Pan W. Alzheimer's Disease Neuroimaging Initiative: testing for association with multiple traits in generalized estimation equations with application to neuroimaging data. Neuroimage. 2014;96:309–25.
4. Kim J, Zhang Y, Pan W. Powerful and adaptive testing for multi-trait and multi-SNP associations with GWAS and sequencing data. Genetics. 2016;203(2):715–31.
5. Park JY, Wu C, Basu S, McGue M, Pan W. Adaptive SNP-set association testing in generalized linear mixed models with application to family studies. Behav Genet. 2018;48(1):55–66.
6. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993;88(421):9–25.
7. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celed'on JC, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am J Hum Genet. 2016;98(4):653–66.
8. Pan W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. Genet Epidemiol. 2011;35(4):211–6.
9. Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol. 2013;37(2):196–204.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC