

PROCEEDINGS

Open Access

Adjustment of familial relatedness in association test for rare variants

Cong Li¹, Can Yang², Mengjie Chen¹, Xiaowei Chen¹, Lin Hou², Hongyu Zhao^{2*}

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

Abstract

High-throughput sequencing technology allows researchers to test associations between phenotypes and all the variants identified throughout the genome, and is especially useful for analyzing rare variants. However, the statistical power to identify phenotype-associated rare variants is very low with typical genome-wide association studies because of their low allele frequencies among unrelated individuals. In contrast, a family-based design may have more power because rare variants are more likely to be enriched in families than among unrelated individuals. Regardless, an analysis of family-based association studies needs to account appropriately for relatedness between family members. We analyzed the observed quantitative trait systolic blood pressure as well as the simulated Q1 data in the Genetic Analysis Workshop 18 data set using 4 tests: (a) a single-variant test, (b) a collapsing test, (c) a single-variant test where familial relatedness was accounted for, and (d) a collapsing test where familial relatedness was accounted for. We then compared the results of the 4 methods and observed that adjusting for familial relatedness could appropriately control the false-positive rate while maintaining reasonable power to detect several strongly associated variants/genes.

Background

Current platforms for genome-wide association studies are limited to scanning common variants. Although rare variants may contribute a significant proportion of heritability, the statistical power to detect rare variants is low because of their low allele frequencies. Recent advances in high-throughput sequencing technologies have provided us great opportunities to delve deeper into the genetic components of complex traits by identifying millions of rare variants in the human genome [1] and allowing them to be tested for associations with complex traits.

In an effort to increase the power to detect rare-variant associations, many methods have been proposed to aggregate the effects of multiple rare variants within a specific functional unit, for example, a gene [2]. Among those methods, the kernel score test has enjoyed great popularity thanks to its flexibility and computational efficiency [3]. Family-based designs may offer more power, however,

because related individuals are more likely than unrelated individuals to be enriched for rare variants [4]. One challenge of analyzing family data is that familial relatedness needs to be carefully adjusted for in the association analysis. Linear mixed models have been used to account for population stratification in the context of genome-wide association studies (GWAS) [5], and can also be naturally adapted to a family-based design. In this paper, we adopted a linear mixed model to adjust for familial relatedness while aggregating the effects of rare variants within a gene using a kernel score test.

Methods

Data description and preprocessing

The Genetic Analysis Workshop 18 (GAW18) data consists of genotyping and sequencing data for 959 individuals from 20 extended families. Among them, 849 individuals had at least 1 blood pressure measurement. For the real data, we used the first nonmissing measurement of systolic blood pressure (SBP) as the target quantitative trait. For the 200 replicates of the simulated phenotype, we also used the first measurement of SBP as the target quantitative

* Correspondence: hongyu.zhao@yale.edu

²Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06520, USA

Full list of author information is available at the end of the article

trait. Finally, to evaluate the false-positive rate, we used the simulated Q1 data as the null phenotype.

The common variants in the genotyping data were used to estimate the genetic similarity matrix. As for the sequencing data, we selected only the nonsynonymous mutations for the association test in recognition of the fact that variants causing amino acid alterations tend to have large effects on the phenotype.

The model

We used the following linear mixed model to adjust for familial relatedness: $Y = C\gamma + X\beta + Z\alpha + e$ where Y represents the phenotype (quantitative trait), C represents the collection of the covariates (age and gender), X includes the genotypes of the variants to be tested, and Z is the design matrix for the whole-genome polygenic random effects α . Finally, γ and β are fixed effects and e is the random residual. When the random effects are integrated out, the model is marginalized as $Y \sim N(C\gamma + X\beta, K\sigma_\alpha^2 + I\sigma_e^2)$, where σ_α^2 and σ_e^2 are the variances for the random effects α and e , respectively. As for K , one could use twice the theoretical kinship matrix to represent the familial relationships. However, to account for potential cryptic relatedness between individuals across different families, we used a genetic similarity matrix that was estimated from the genotyping data as follows:

$$K_{ij} = \frac{1}{p} \sum_{m=1}^p \frac{(z_{im} - 2f_m)(z_{jm} - 2f_m)}{2f_m(1 - f_m)}$$
, where z_{im} is the number of minor alleles for i th individual at the m th marker, f_m is the allele frequency for the m th marker, and p is the total number of markers in the genotyping data.

We investigated 4 methods of association analysis: (a) a single-variant association test without adjusting for familial relatedness (UNI), that is, the component $Z\alpha$ is excluded in the model; (b) a single-variant association test adjusting for familial relatedness (UNI-ADJ); (c) a kernel score test without adjusting for familial relatedness (SKAT); and (d) a kernel score test adjusting for familial relatedness (SKAT-ADJ). The first method is a simple linear regression. The second method tests the null hypothesis of zero fixed effects in a linear mixed model. Specifically, we used the likelihood ratio test implemented in the software GEMMA [6]. The third method is the direct application of the SKAT software [3] that ignores familial relatedness. As for the fourth method, we adopted the approach described in Ref. [7]. Specifically, we estimated the covariance matrix $R = \text{cov}(Z\alpha + e) = K\sigma_\alpha^2 + I\sigma_e^2$ under the null model without genes or variants being tested. Then we applied a data transformation to calculate \tilde{Y} , \tilde{X} , and \tilde{C} , where: $\tilde{X} = R^{-1/2}X$, $\tilde{X} = R^{-1/2}X$, $\tilde{C} = R^{-1/2}C$, leading to the following transformed model, $\tilde{Y} = \tilde{C}\gamma + \tilde{X}\beta + \tilde{e}$, where $\tilde{e} = R^{-1/2}(Z\alpha + e)$. Note that each \tilde{e}_i is independent and

identically distributed; therefore, the kernel score test can be appropriately applied on the transformed data.

Results

Simulated data

For the simulated data, we analyzed the genes that explain the highest percentages of SBP variance (the percentages of SBP variance explained by each gene in the simulation were provided by the GAW18 organizers) using the 4 methods described above. Genes without nonsynonymous mutations were excluded. For the first two single-variant methods, we performed a Bonferroni correction on the smallest p value of the variants in a gene to obtain the gene-level p value. To assess the false-positive rate, we also calculated the p values using Q1 as the phenotype. The power and the false-positive rate were calculated based on the 200 replicates. Table 1 shows the results for the top 7 genes. As the table shows, ignoring familial relatedness results in inflation of the false-positive rate using the UNI and SKAT methods, in contrast with the UNI-ADJ and SKAT-ADJ methods, which still demonstrate good control of the false-positive rate. As for the power, the single-variant method does better for genes whose effects on the phenotype are dominated by a single variant (eg, *LEPR*) whereas the collapsing method is more powerful for genes containing multiple variants with comparable effect sizes.

Real data

Consistent with our observation in the simulated data, the Q-Q plot (Figure 1) shows that the UNI and SKAT methods that ignore familial relatedness result in substantial inflation of the false-positive rate. For UNI-ADJ, only 1 variant (chr11:119059358) has a Bonferroni-corrected p value smaller than 0.05 ($P_{\text{Bonferroni}} = 0.038$). It is a coding variant for gene *PDZD3*. To our knowledge, the literature has suggested no role for this gene in blood pressure. For SKAT-ADJ, there is no gene with a Bonferroni-corrected p value smaller than 1.

Discussion and conclusions

In this study, we investigated the utility of linear mixed models in adjusting for familial structure. We also attempted to combine a linear mixed model method with a kernel score test—a state-of-the-art collapsing methodology that tests for association between a group of variants and a phenotype.

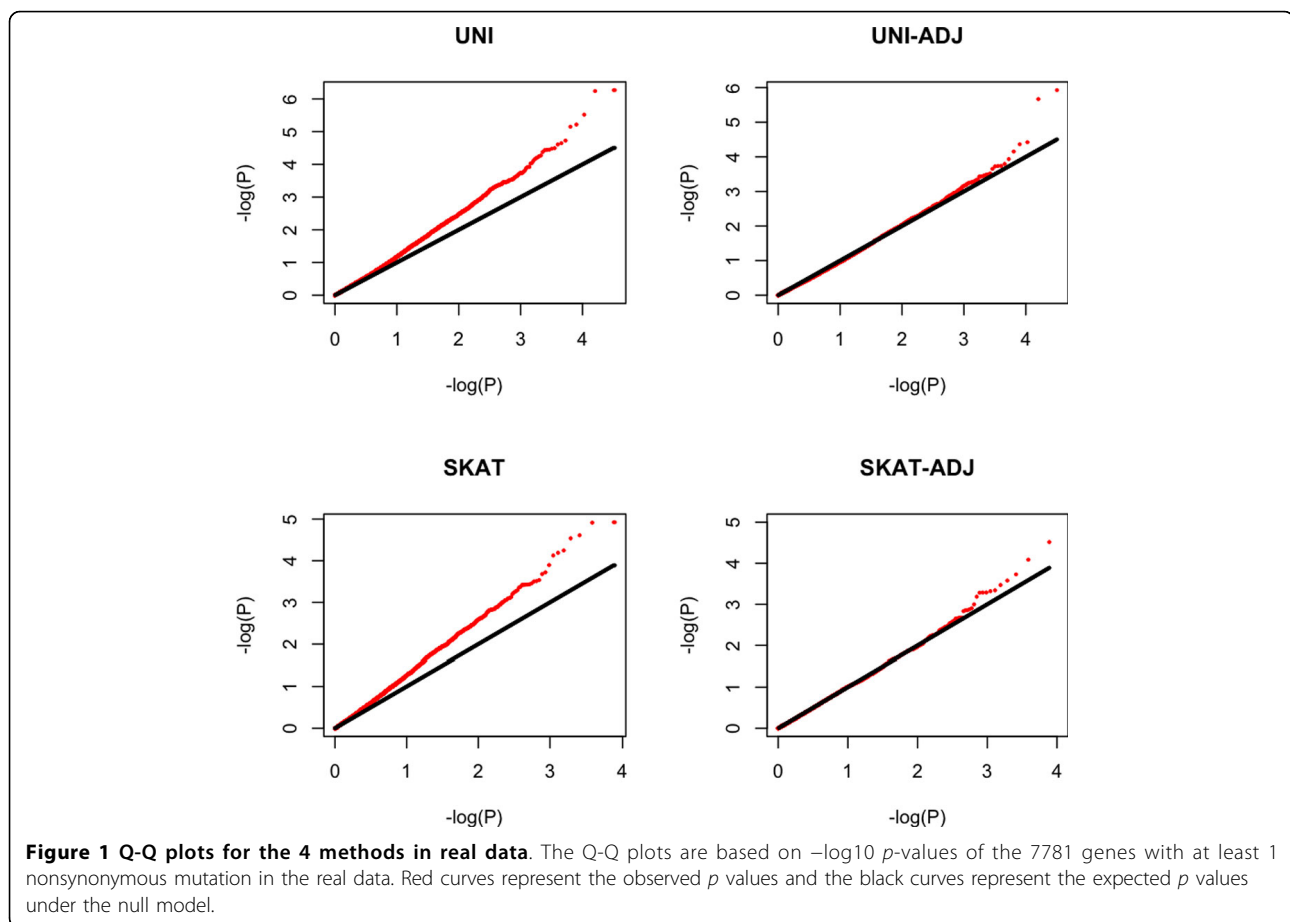
Conclusions

We found that linear mixed models are able to satisfactorily adjust for familial relatedness in the sense that false-positive rates are well controlled at the nominal significance level. Not surprisingly, in terms of the statistical power, the performance of the single-variant

Table 1 The power and false-positive rate (FPR) of the 4 methods in simulated data

Gene	# of NS variants	UNI		UNI-ADJ		SKAT		SKAT-ADJ	
		Power	FPR	Power	FPR	Power	FPR	Power	FPR
MAP4	8	1	0.12	1	0.04	1	0.13	1	0.03
TNN	20	0.9	0.225	0.74	0.055	0.895	0.38	0.825	0.04
LEPR	6	0.915	0.125	0.895	0.05	0.22	0.085	0.73	0.04
FLT3	4	0.66	0.105	0.3	0.045	0.33	0.135	0.305	0.07
FLNB	13	0.44	0.14	0.165	0.03	0.08	0.17	0.515	0.035
ZNF443	9	0.070	0.155	0.035	0.030	0.150	0.075	0.050	0.075
GSN	6	0.265	0.155	0.130	0.070	0.030	0.065	0.385	0.065

A significance level of 0.05 was used to calculate both power and FPR.



method and the collapsing method depends on the distribution of the variants' effect sizes within the gene. The single-variant method is more powerful for genes with one dominating causal variant, whereas the collapsing method is more powerful for genes with multiple causal variants of similar effect sizes.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CL and HZ designed the overall study; CY, CL, MC, XC, and LH conducted statistical analyses; and CL and HZ drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the organizers of Genetic Analysis Workshop 18 for providing the data set. We also thank Yale University Biomedical High Performance Computing Center for computing resources, and National Institutes of Health (NIH) grant RR19895, which funded the instrumentation. We would also like to acknowledge the support from the VA Cooperative Studies of the Department of Veterans Affairs. CY, LH, and HZ were supported in part

by the NIH grants R01 GM59507, R01 DA12849, and R01 DA030976. CL, MC, and XC acknowledge support from CSC Scholarship. Finally, we thank our colleague Gregory Ryslik for helping us to revise this manuscript. 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 P01 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.

Authors' details

¹Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. ²Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06520, USA.

Published: 17 June 2014

References

1. Nielsen R: **Genomics: In search of rare human variants.** *Nature* 2010, **467**:1050-1051.
2. Bansal V, Libiger O, Torkamani A, Schork NJ: **Statistical analysis strategies for association studies involving rare variants.** *Nat Rev Genet* 2010, **11**:733-785.
3. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: **Rare-variant association testing for sequence data with the sequence kernel association test.** *Am J Hum Genet* 2011, **89**:82-93.
4. Ott J, Kamatani Y, Lathrop M: **Family-based designs for genome-wide association studies.** *Nat Rev Genet* 2011, **12**:465-474.
5. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E: **Variance component model to account for sample structure in genome-wide association studies.** *Nat Genet* 2010, **42**:348-354.
6. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies.** *Nat Genet*; 2012:**44**:821-824.
7. Wiencierz A, Greven S, Küchenhoff H: **Restricted likelihood ratio testing in linear mixed models with general error covariance structure.** *Electron J Statist*; 2011:**5**:1718-1734.

doi:10.1186/1753-6561-8-S1-S39

Cite this article as: Li *et al.*: Adjustment of familial relatedness in association test for rare variants. *BMC Proceedings* 2014 **8**(Suppl 1):S39.

**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

