Family-based association test using normal approximation to gene dropping null distribution

Yuan Jiang, Sarah Emerson, Lu Wang, Lujing Li, Yanming Di*

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

Abstract

We derive the analytical mean and variance of the score test statistic in gene-dropping simulations and approximate the null distribution of the test statistic by a normal distribution. We provide insights into the gene-dropping test by decomposing the test statistic into two components: the first component provides information about linkage, and the second component provides information about fine mapping under the linkage peak. We demonstrate our theoretical findings by applying the gene-dropping test to the simulated data set from Genetic Analysis Workshop 18 and comparing its performance with existing population and family-based association tests.

Background

When testing genotype-phenotype association using individuals from extended families, one has to account for correlations in genotypes and/or phenotypes between related individuals. One simple and effective method to account for genotype correlations is to simulate the null genotype distribution by gene dropping [1], which is simulating founder alleles according to estimated allele frequencies and dropping these alleles down the pedigrees according to random segregation of gametes (i.e., Mendel's first law). The gene-dropping method is straightforward to implement (e.g., implemented in by Allen-Brady et al [2]) and applies to all pedigree structures, but it is computationally intensive and thus is impractical to use when dealing with millions of single-nucleotide polymorphisms (SNPs).

In this article, we derive the analytical mean and variance of the score test statistic under the gene-dropping setting and approximate the gene-dropping null distribution of the test statistic by a normal distribution with the analytically derived mean and variance. Using this normal approximation, the gene-dropping test becomes computationally efficient and can be easily applied to millions of SNPs.

Furthermore, we provide insights into the gene-dropping test by decomposing the test statistic into two components: the first component resembles a quantity frequently used in variance-component based linkage tests and provides information for linkage, and the second component provides information for fine mapping under the linkage peak. Rabinowitz and Laird [3], among others, have pointed out the subtle distinction between two types of null hypotheses in family-based association analysis: the null hypothesis of no linkage and no association versus the null hypothesis of no association in the presence of linkage. To test the latter, one needs to condition on the inheritance $S\tau$ vector at the test locus [3]. Our decomposition provides an explicit separation of linkage and association information in a family-based study.

We compare the performance of the gene-dropping test (using normal approximation) to association tests using only unrelated individuals and to the family-based association test in the software program FBAT [3] by analyzing Genetic Analysis Workshop 18 (GAW18) simulated data set.

Methods

Preprocessing of genotype data

We analyzed SNPs from chromosome 3 only. At each of the SNPs, we performed Pearson’s chi-squared test for the Hardy-Weinberg equilibrium using 142 unrelated individuals. We excluded SNPs that yielded a $p$-value smaller than $10^{-4}$ from our analysis. In the gene-dropping
test, we excluded SNPs with estimated minor allele frequency (MAF) smaller than 0.001.

Preprocessing of phenotype data
We focused on the analysis of the quantitative trait systolic blood pressure (SBP) in the simulated data set 1. The true simulation model was known to us [4]. When testing association between genotype doses and trait values (see later discussion), we include factors AGE, SEX, and AGE by SEX interaction as covariates (Z_k’s in equation [1]). Including BPMED as a covariate will overcompensate because BPMED is a consequence of SBP level. Instead, we estimated the effect of BPMED from a regression model with only individuals with hypertension. Because BPMED was randomly assigned to individuals with hypertension, the BPMED effect estimated this way will not be biased by its correlation with SBP. We then adjusted the trait values Y by subtracting the estimated BPMED effect.

Score tests of genotype-phenotype association using unrelated individuals
At locus \( \tau \), we consider a quantitative trait model

\[
E(Y) = \mu + \sum_{k=1}^{K} \alpha_k Z_k + X_\tau \beta_\tau, \tag{1}
\]

and test the null hypothesis \( \beta_\tau = 0 \). In equation (1), \( Y \) is the vector of trait values (SBP adjusted for the BPMED effect), \( \mu \) is a constant vector of baseline mean trait values, coefficients \( \alpha_k \) represent the effects of the covariates \( Z_k, k = 1, \ldots, K \), (e.g., AGE, SEX and AGE by SEX interaction) on trait values, \( X_\tau \) is the vector of genotype doses (the number of minor alleles possessed by each individual) at locus \( \tau \), and the coefficient \( \beta_\tau \) represents the effect size of a single allele. The fitted value of \( \beta_\tau \) will reflect the collective effect of all causal SNPs that are in linkage disequilibrium (LD) with the test SNP \( \tau \) [5].

Let \( \hat{Y} \) and \( \hat{X}_\tau \) be the vectors of fitted values after regressing the \( Y \) and \( X_\tau \) on measured covariates \( Z_k \)’s. The score statistic [6,7] for testing genotype-trait association at a single SNP \( \tau \) is

\[
E(X_\tau|S_\tau) = \phi_{ij} (S_\tau) f_\tau + (4 - \phi_{ij} (S_\tau)) f_\tau^2 = \phi_{ij} (S_\tau) (f_\tau - f_\tau^2) + 4f_\tau^2,
\]

(1.e.g., \( E(P_1P_2) = E(P_1^2) = f_\tau \) if \( P_1 \) and \( P_2 \) correspond to IBD alleles and \( E(P_1P_2) = E(P_1) E(P_2) = f_\tau^2 \) if \( P_1 \) and \( P_2 \) correspond to non-IBD alleles). In a gene-dropping simulation, the inheritance vector \( S_\tau \) is randomly sampled among all possible inheritance vectors. The expected number of IBD alleles shared between \( i \) and \( E(\phi_{ij} (S_\tau)) \), \( E(\phi_{ij} (S_\tau)) \), over all possible inheritance vectors is four times the kinship coefficient \( \psi_{ij}; E(\phi_{ij} (S_\tau)) = 4\psi_{ij} \). The kinship coefficients are determined by pedigree structures. The expected value of \( X_\tau X_j \) in a gene-dropping simulation is thus

\[
E(X_\tau X_j|S_\tau) = 4\psi_{ij} (f_\tau - f_\tau^2) + 4f_\tau^2.
\]

Letting \( \Phi(S_\tau) = (\phi_{ij}) \) be the matrix of IBD counts and \( \Psi \) be the matrix of kinship coefficients, we can rewrite the above as:

\[
E(X_\tau X_j|S_\tau) = \Phi(S_\tau) (f_\tau - f_\tau^2) + 4f_\tau^2.
\]

Family-based association test by gene dropping
When related individuals are used to compute the score test statistic \( u = X_\tau R \), components of \( X_\tau \) can be dependent, and the variance estimator (2) is no longer valid. One can account for correlations between components in \( X_\tau \) by simulating the null distribution of \( X_\tau \), using gene dropping. We now derive the analytical mean and variance of \( u \) under the gene-dropping setting. In the score test using unrelated individuals, we treat \( R \) as random, and \( X_\tau \) can be viewed as either random or fixed. In a gene-dropping simulation, \( R \) is held fixed, and \( X_\tau \) is random.

Let \( i,j \) index individuals \( (i,j = 1, \ldots, n) \) and let \( X_\tau = (X_{\tau1}, \ldots, X_{\tau n}) \) and \( R = (R_1, \ldots, R_n) \). The expected value of \( u \) is \( \sum_{i=1}^{n} E(X_{\tau i}) R_i \) and \( X_i = P_i + M_i \), where \( (M_i) \) is 1 if the paternal (maternal) allele is the minor allele and 0 otherwise. So \( E(X_{\tau}) \) is twice the MAF \( f_\tau \) at SNP \( \tau \) and is the same for all individuals and thus \( E(u) = 2f_\tau \sum R_i = 0 \) because \( R_i \)’s are residuals from a linear regression model with intercept. The variance of \( u \) is \( E(u^2) = R E(X_\tau X_\tau) R \). If \( (i,j) \)th element in \( E(X_\tau X_\tau) \) is

\[
E(X_{\tau i} X_{\tau j}) = E(P_i P_j + P_i M_j + M_i P_j + M_i M_j), \quad P_i, M_i, P_j, M_j \text{ are all Bernoulli random variables with probability } f_\tau, \text{ and any two of them are identical if the corresponding alleles are identity-by-descent (IBD) and are independent otherwise [8]. Let } \phi_{ij} \text{ be the number of IBD pairs among the four pairs of alleles } P_i P_j, P_i M_j, M_i P_j, M_i M_j. \text{ The value of } \phi_{ij} \text{ at locus } \tau \text{ is determined by the inheritance vector } S_\tau, \text{ which summarizes whether the paternal or the maternal allele is passed from the parent to the child in each meiosis [9]. Given the inheritance vector } S_\tau,
\]

\[
E(X_{\tau i} X_{\tau j}|S_\tau) = \phi_{ij} (S_\tau) f_\tau + (4 - \phi_{ij} (S_\tau)) f_\tau^2 = \phi_{ij} (S_\tau) (f_\tau - f_\tau^2) + 4f_\tau^2.
\]

\(E(X_{\tau i} X_{\tau j}|S_\tau) = 4\psi_{ij} (f_\tau - f_\tau^2) + 4f_\tau^2.
\]

where \( J \) is a matrix of all ones. Because \( R^T R = 0 \) for residuals from a linear regression model with an
intercept, the variance of $u$ under gene dropping is $\nu_{gd} = R' E(X_i X_i') R = 4 R' \psi R \left( f_i - f_i^2 \right)$ if unconditional on the inheritance vector $S_r$, and is $\nu_r = R' E(X_i X_i' | S_r) R = R' \Phi(S_r) R \left( f_i - f_i^2 \right)$ if conditional on the inheritance vector $S_r$ (holding $S_r$ fixed). We can approximate the gene-dropping null distribution of $u$ by a normal distribution with mean $0$ and variance $\nu_{gd}$, and compute the gene-dropping $p$-value by comparing $t = u^2/\nu_{gd}$ with a $\chi^2$ distribution. To test association in the presence of linkage, one needs to condition on the inheritance vector at $\tau$ [3] and use $\nu_r$. In practice, $S_r$ is not observable, but we estimate $\nu_r$ by drawing Markov chain Monte Carlo (MCMC) samples of $S_r$ based on observed genotypes in the pedigrees using MORGAN (http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml) [10].

Results

Theoretical findings

In a gene-dropping simulation, the analytical mean of the score statistic $u = X_i' R'$ is 0. The variance of the score statistic is $R' \Phi(S_r) R \left( f_i - f_i^2 \right)$ if conditional on the inheritance vector $S_r$ (i.e., holding the inheritance vector fixed during gene-dropping simulation) and is $4 R' \psi R \left( f_i - f_i^2 \right)$ if unconditional on the inheritance vector. The normal approximation is justified by the central limit theorem because the test statistic is additive over pedigrees. Its performance depends on the number, sizes, and structure of pedigrees and on MAF at the test locus. The approximation may not be accurate for extremely small $p$-values. However, the rankings of the $p$-values will not change.

We can decompose the unconditional gene-dropping test statistic into two components:

$$\frac{u^2}{4 R' \psi R \left( f_i - f_i^2 \right)} = \left[ \frac{u^2}{R' \Phi(S_r) R \left( f_i - f_i^2 \right)} \right] \left[ \frac{R' \Phi(S_r) R}{4 R' \psi R} \right].$$

The first component can be used as a test statistic for detecting association in the presence of linkage (i.e., fine mapping under a linkage peak) because the denominator is the variance of $u$ conditional upon the observed IBD sharing. The second component provides information about linkage. The kinship coefficients in $\psi$ are determined by pedigree structure, so $R' \Phi R$ is a constant in a gene-dropping simulation. $R' \Phi(S_r) R = \sum_{\eta} r_{\tau} r_{\eta} \phi_{\eta}(S_r)$ measures the correlation between trait value similarity $(r_{\tau} r_{\tau})$ and IBD sharing $(\phi_{\eta})$ at locus $\tau$ across all pairs of individuals in a pedigree. This correlation is expected to be stronger if there is stronger linkage between $\tau$ and a true causal locus. Therefore, $R' \Phi(S_r) R$ can be used as a test statistic to detect linkage, with null distribution obtained by gene-dropping simulations. In a gene-dropping simulation, the inheritance vectors are simulated as if they were from a marker unlinked to any potential causal loci. $R' \Phi(S_r) R$ resembles similar quantities that are frequently used in linkage analysis methods such as the well-known Haseman-Elston regression [11] as well as many variance components or generalized estimating equation-based methods [12].

Simulation results

We performed a genome-wide association studies (GWAS) score test using 142 unrelated individuals, the family-based association test using FBAT [3], and the gene-dropping test on SNPs on chromosome 3 (FBAT and the gene-dropping test used 847 individuals from 20 pedigrees). Table 1 summarizes the $p$-value ranks that each test assigns the true causal SNPs. The gene-dropping test for fine mapping (conditional on the inheritance vector) performs very similarly to the unconditional gene-dropping test, so its results are omitted. It is seen that the gene-dropping tests can quickly identify a few true causal SNPs within a short list of top findings. However, if we allow more false positives by considering a greater number of the most significant SNPs, other methods start to pick up true causal SNPs and eventually have a result similar to gene dropping.

Figure 1 shows the physical positions and negative log $p$-values of the top 500 SNPs identified by each of the three tests, as well as the negative log $p$-values of the linkage test based on the linkage component of the gene-dropping test statistic.

We also examined adjusting for population stratification by fitting the first two principal components of genetic variation [13] as covariates in the regression model (1). The $p$-values resulting from this expanded model differed negligibly from the original model. The ranks in Table 1 were essentially unchanged by this adjustment.

Discussion

Comparison between genome-wide association studies, FBAT and gene-dropping test

FBAT splits each pedigree into nuclear families. In each nuclear family, FBAT uses information from the offspring while conditioning on the parental marker genotypes. In contrast, GWAS uses information in unrelated individuals. The two methods use almost “orthogonal” sources of information. There is almost no correlation between the $p$-values from these two methods (Table 2). In contrast, the gene-dropping test applies to multigeneration pedigrees and uses information from all
Table 1 Ranks of truly influential single-nucleotide polymorphisms by genome-wide association studies, FBAT, and gene dropping

| Rank | GWAS SNP position | FBAT Rank | SNP position | Gene dropping Rank | SNP position |
|------|-------------------|-----------|--------------|-------------------|--------------|
| 22   | 47957996          | 0.00212   | 1433         | 0.25642           | 47956424     |
| 27   | 48040283          | 0.00260   | 2903         | 0.51947           | 47958037     |
| 1,561| 58161774          | 1,561     | 9,096.5      | 1.62595           | 47957996     |
| 5,901| 47467805          | 5,901     | 5,536        | 0.99062           | 48040283     |
| 11,415| 58161774         | 11,415    | 3,937        | 0.32787           | 47956424     |
| 21,148.5| 47958037      | 21,148.5  | 15,860.5     | 2.83810           | 141093285    |
| 23,791| 196597635         | 23,791    | 17,341.5     | 3.10311           | 141162128    |
| 28,783| 135789360         | 28,783    | 22,148.5     | 3.96328           | 141160882    |
| 30,761| 47956424          | 30,761    | 23,778       | 4.25487           | 141160882    |
| 34,720.5| 58190853        | 34,720.5  | 32,483       | 5.81256           | 58192585     |

Rank is raw ranks in terms of \( p \)-value significance of truly influential single-nucleotide polymorphisms (SNPs) (smaller numbers better, indicating that the method identifies a true SNP as more significant). The fractional ranks appearing in the gene-dropping column arise from ties: two SNPs being assigned exactly the same \( p \)-value. Note that it is not completely fair to compare these numbers directly because FBAT and genome-wide association studies (GWAS) produce not available (NA) results for a significant portion of the tested SNPs. Relative rank is the normalized ranks of truly influential SNPs: \( p \)-value rank divided by the total number of non-NA SNPs tested multiplied by 100. SNP position is the base-pair position of the identified truly influential SNP.

Table 2 Correlation between log \( p \)-values of genome-wide association studies, FBAT, and gene dropping

| GWAS/FBAT | Gene dropping/FBAT | Gene dropping/GWAS |
|-----------|---------------------|---------------------|
| 0.011     | 0.232               | 0.254               |

GWAS, genome-wide association studies.
individuals: the gene-dropping test extracts information from founders by resimulating founder genotypes and from offspring by resimulating inheritance vectors. It is also possible to derive the analytical mean and variance of the test statistic in the gene-dropping test where we permute the founder alleles rather than resimulate the founder alleles. FBAT is more robust to population stratification by conditioning on founder genotypes. The gene-dropping test can gain similar robustness by restricting permutations to founder alleles within each family.

It is somewhat surprising that the gene-dropping test did not outperform GWAS given that it uses more individuals. One possible interpretation is that the effect of LD is stronger when more individuals are used. As we can see in Figure 1, the signals detected by the gene-dropping test come in bigger clusters. In other words, many SNPs ranked high by the gene-dropping test might be in LD with one or more of the causal SNPs.

Separating linkage and association signals
The gene-dropping test captures both linkage and association signals. One can decompose the test statistic into a linkage component and an association component. The association component corresponds to testing association in the presence of linkage, which requires one to condition on the true inheritance vector at the test locus. Our results through MCMC approximation show that whether or not to condition on the inheritance vector actually does not make a big difference for this data set because the variance of the test statistic with conditioning only differs slightly from the variance of the test statistic without conditioning. This conclusion might be dependent on the structure of the pedigree.

The linkage component, however, clearly provides valuable information. The linkage signal is stronger in most regions containing causal SNPs. It is obvious that the linkage curve can help eliminate many of the false association signals in this study. It would be interesting to investigate how to use the linkage information more effectively in the future.

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

Authors' contributions
All authors participated in analysis of the data. YJ, SE, and YD conceived of the study and drafted the manuscript. All authors participated in the critical revision of the manuscript and gave final approval of the article.

Acknowledgements
We thank the GAW18 organizers. The GAW18 whole genome sequence data were provided by the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) Consortium, which is supported by National Institutes of Health (NIH) grants U01 DK085522, R01 DK047482, and R01 DK053889. The GAW is supported by NIH grant R01 GM031575. This article has been published as part of BMC Proceedings Volume 8 Supplement 1: 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. MacCluer JW, Vandenberg IL, Read B, Ryder OA: Pedigree analysis by computer simulation. Zoo Biol 1986, 5:149-160.
2. Allen-Brady K, Wong J, Camp NJ: PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. BMC Bioinformatics 2006, 7:209.
3. Rabowitwitz D, Laird N: A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 2000, 50:211-223.
4. Almasy L, Dyer T, Peralta J, Jun G, Fuchsberger C, Almeida M, Kent JW Jr, Fowler S, Duggirala R, Bangerter L: Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. BMC Proc 2014, 8(Suppl 2):52.
5. Di Y, Mi G, Sun L, Dong R, Zhu H, Peng L: Power of association tests in the presence of multiple causal variants. BMC Proc 2011, 5(Suppl 9):563.
6. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. Ann Hum Genet 2002, 70:425-434.
7. Clayton D, Chapman J, Cooper J: Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 2004, 27:415-428.
8. Lange K, Westlake J, Spence MA: Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet 1976, 39:485-491.
9. Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 1987, 84:2363-2367.
10. Thompson EA: MCMC in the analysis of genetic data on pedigrees. In Markov Chain Monte Carlo: Innovations and Applications. Lecture Note Series of the IMS, National University of Singapore. World Scientific Co Pte Ltd, Singapore; 1996:183-216.
11. Elston RC, Buxbaum S, Jacobs KB, Olson JM, Haseman and Elston revisited. Genet Epidemiol 2004, 19:1-17.
12. Chen WM, Broman KW, Liang KY: Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. Genet Epidemiol 2004, 27:265-272.
13. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006, 38:900-909.

Cite this article as: Jiang et al.: Family-based association test using normal approximation to gene dropping null distribution. BMC Proceedings 2014 8(Suppl 1):S18.