Jiang et al. BMC Proceedings 2014, 8(Suppl 1):S18
http://www.biomedcentral.com/1753-6561/8/S1/S18

BMC Proceedings

PROCEEDINGS Open Access

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_τ 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

* Correspondence: diy@stat.oregonstate.edu
Department of Statistics, Oregon State University, Corvallis, OR 97331, USA

BioMed Central

© 2014 Jiang 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.

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 τ, we consider a quantitative trait model

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

and test the null hypothesis $\beta_\tau = 0$. In equation (1), Y is the vector of trait values (SBP adjusted for the BPMED effect), μ is a constant vector of baseline mean trait values, coefficients α_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_τ is the vector of genotype doses (the number of minor alleles possessed by each individual) at locus τ, and the coefficient β_τ represents the effect size of a single allele. The fitted value of β_τ will reflect the collective effect of all causal SNPs that are in linkage disequilibrium (LD) with the test SNP τ [5].

Let $\hat{Y}$ and $\hat{X}_\tau$ be the vectors of fitted values after regressing the Y and X_τ on measured covariates Z_k's. The score statistic [6,7] for testing genotype-trait association at a single SNP τ is $u = X_\tau' R$, where $R = Y - \hat{Y}$ is the vector of residuals. Under the null hypothesis of no association, the variance of u is estimated by

$$v = s_{YY} X_\tau' (X_\tau - Z(Z'Z)^{-1} Z' X_\tau) = s_{YY} X_\tau' \left(X_\tau - \hat{X}_\tau\right), \quad (2)$$

where $Z = (1, Z_1, \ldots, Z_K)$ and s_{YY} is the sample variance of the residual trait values (1 is a vector of ones) [6]. To test association, u^2/v is compared with a χ_1^2 distribution.

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_τ can be dependent, and the variance estimator (2) is no longer valid. One can account for correlations between components in X_τ by simulating the null distribution of X_τ 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_τ can be viewed as either random or fixed. In a gene-dropping simulation, R is held fixed, and X_τ is random.

Let i, j index individuals $(i, j = 1, \ldots, n)$ and let $X_\tau = (X_1, \ldots, X_n)'$ and $R = (R_1, \ldots, R_n)'$. The expected value of u is $\sum_{i=1}^{n} E(X_i) R_i$ and $X_i = P_i + M_i$, where (M_i) (M_i) is 1 if the paternal (maternal) allele is the minor allele and 0 otherwise. So $E(X_i)$ is twice the MAF f_τ at SNP τ 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 $E(u^2) = R' E(X_\tau X_\tau') R$. The (i, j) th element in $E(X_\tau X_\tau')$ is $E(X_i X_j) = E(P_i P_j + P_i M_j + M_i P_j + M_i M_j)$. P_i, M_i, P_j, M_j are all Bernoulli random variables with probability f_τ, and any two of them are identical if the corresponding alleles are identity-by-descent (IBD) and are independent otherwise [8]. Let ϕ_{ij} be the number of IBD pairs among the four pairs of alleles $P_i P_j, P_i M_j, M_i P_j, M_i M_j$. The value of ϕ_{ij} at locus τ is determined by the inheritance vector S_τ, 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_τ,

$$E(X_i X_j | S_\tau) = \phi_{ij}(S_\tau) f_\tau + (4 - \phi_{ij}(S_\tau)) f_\tau^2 = \phi_{ij}(S_\tau)(f_\tau - f_\tau^2) + 4f_\tau^2,$$

(e.g., $E(P_i P_j) = E(P_i^2) = f_\tau$ if P_i and P_j correspond to IBD alleles and $E(P_i P_j) = E(P_i) E(P_j) = f_\tau^2$ if P_i and P_j correspond to non-IBD alleles). In a gene-dropping simulation, the inheritance vector S_τ is randomly sampled among all possible inheritance vectors. The expected number of IBD alleles shared between i and $\mathrm{E}(\phi_{ij}(S_\tau))$,, $\mathrm{E}(\phi_{ij}(S_\tau))$, over all possible inheritance vectors is four times the kinship coefficient ψ_{ij}:$\mathrm{E}(\phi_{ij}(S_\tau)) = 4\psi_{ij}$. The kinship coefficients are determined by pedigree structures. The expected value of $X_i X_j$ in a gene-dropping simulation is thus $E(E(X_i X_j | S_\tau)) = 4\psi_{ij}(f_\tau - f_\tau^2) + 4f_\tau^2$. Letting $\Phi(S_\tau) = (\phi_{ij})$ be the matrix of IBD counts and Ψ be the matrix of kinship coefficients, we can rewrite the above as:

$$E(X_\tau X'_\tau | S_\tau) = \Phi(S_\tau)(f_\tau - f_\tau^2) + 4Jf_\tau^2,$$

$$E(X_\tau X_\tau') = E(E(X_\tau X_\tau' | S_\tau)) = 4\Psi(f_\tau - f_\tau^2) + 4Jf_\tau^2,$$

where J is a matrix of all ones. Because $R'JR = 0$ for residuals from a linear regression model with an

intercept, the variance of u under gene dropping is $v_{gd} = R'E\left(X_\tau X_\tau'\right)R = 4R'\Psi R\left(f_\tau - {f_\tau}^2\right)$ if unconditional on the inheritance vector S_τ, and is $v_\tau = R'E\left(X_\tau X_\tau'|S_\tau\right)R = R'\Phi\left(S_\tau\right)R\left(f_\tau - {f_\tau}^2\right)$ if conditional on the inheritance vector S_τ (holding S_τ fixed). We can approximate the gene-dropping null distribution of u by a normal distribution with mean 0 and variance v_{gd}, and compute the gene-dropping p-value by comparing $t = u^2/v_{gd}$ with a χ_1^2 distribution. To test association in the presence of linkage, one needs to condition on the inheritance S_τ vector at τ [3] and use v_τ. In practice, S_τ is not observable, but we estimate v_τ by drawing Markov chain Monte Carlo (MCMC) samples of S_τ 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 $u = {X_\tau}'R$ is 0. The variance of the score statistic is $R'\Phi\left(S_\tau\right)R\left(f_\tau - f_\tau^2\right)$ if conditional on the inheritance vector (i.e., holding the inheritance vector fixed during gene-dropping simulation) and is $4R'\Psi R\left(f_\tau - f_\tau^2\right)$ 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:

$$\frac{u^2}{4R'\Psi R\left(f_\tau - f_\tau^2\right)} = \left[\frac{u^2}{R'\Phi\left(S_\tau\right)R\left(f_\tau - f_\tau^2\right)}\right]\left[\frac{R'\Phi\left(S_\tau\right)R}{4R'\Psi R}\right].$$

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 Ψ are determined by pedigree structure, so $R'\Psi R$ is a constant in a gene-dropping simulation. $R'\Phi\left(S_\tau\right)R = \sum_{ij} r_i r_j \phi_{ij}\left(S_\tau\right)$ measures the correlation between trait value similarity $\left(r_i r_j\right)$ and IBD sharing $\left(\phi_{ij}\right)$ at locus τ across all pairs of individuals in a pedigree. This correlation is expected to be stronger if there is stronger linkage between τ and a true causal locus. Therefore, $R'\Phi\left(S_\tau\right)R$ 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. $R'\Phi\left(S_\tau\right)R$ 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 log 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

GWAS			FBAT			Gene dropping		
Rank	Relative rank (%)	SNP position	Rank	Relative rank (%)	SNP position	Rank	Relative rank (%)	SNP position
22	0.00212	47957996	1433	0.25642	47956424	1.5	0.00012	48040283
27	0.00260	48040283	2,903	0.51947	47958037	3.5	0.00029	47957996
1,561	0.15024	141693906	2,913	0.52126	50185967	202.5	0.01686	47958037
5,901	0.56796	47467805	5,536	0.99062	48040283	232	0.01932	47956424
11,415	1.09868	58161774	9,086.5	1.62595	47957996	3,937	0.32787	48040284
21,148.5	2.03552	47958037	15,860.5	2.83810	141093285	13,668	1.13826	58109162
23,791	2.28985	196597635	17,341.5	3.10311	141162128	19,870.5	1.65480	123170592
28,783	2.77033	135789360	22,148.5	3.96328	139276557	37,071	3.08725	141162128
30,761.5	2.96075	47956424	23,778	4.25487	141160882	40,497.5	3.37261	47913455
34,720.5	3.34180	58190853	32,483	5.81256	58192585	42,740	3.55936	141160882

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.

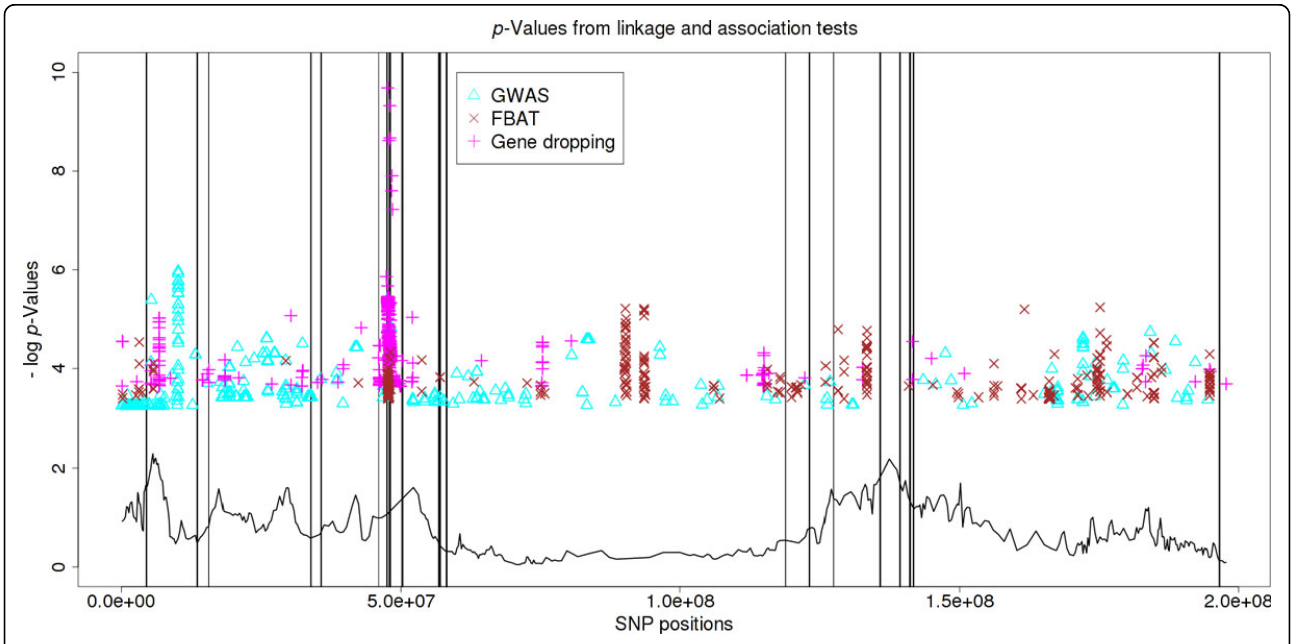

Figure 1 *p*-Values from linkage and association tests. Here we present the chromosome locations of the 500 most significant single-nucleotide polymorphisms (SNPs) on chromosome 3 identified by each method and their corresponding -log p-values. The triangles are the 500 most significant SNPs identified by the genome-wide association studies (GWAS) score test using unrelated individuals; the crosses are those identified by FBAT, and the plusses are those identified by the gene-dropping test. The solid curve shows the -log p-values from the linkage test at 449 evenly spaced SNPs (by comparing the linkage component of the gene-dropping test statistic with its Monte Carlo null distribution from gene dropping). Solid vertical lines indicate the positions of truly influential SNPs on chromosome 3.

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 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 GAW 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

1. MacCluer JW, Vandeburg JL, 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. Rabinowitz 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, Blangero J: **Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees.** *BMC Proc* 2014, **8(suppl 2)**:S2.
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)**:S63.
6. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: **Score tests for association between traits and haplotypes when linkage phase is ambiguous.** *Am J 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;Liang F, Wang J-S, Kendall W 183-216.
11. Elston RC, Buxbaum S, Jacobs KB, Olson JM: **Haseman and Elston revisited.** *Genet Epidemiol* 2000, **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, **26**: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**:904-909.

doi:10.1186/1753-6561-8-S1-S18
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.

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

BioMed Central