Talluri et al. BMC Genetics 2014, 15:75
http://www.biomedcentral.com/1471-2156/15/75

BMC Genetics

METHODOLOGY ARTICLE Open Access

Calculation of exact p-values when SNPs are tested using multiple genetic models

Rajesh Talluri[1], Jian Wang[1] and Sanjay Shete[1,2*]

Abstract

Background: Several methods have been proposed to account for multiple comparisons in genetic association studies. However, investigators typically test each of the SNPs using multiple genetic models. Association testing using the Cochran-Armitage test for trend assuming an additive, dominant, or recessive genetic model, is commonly performed. Thus, each SNP is tested three times. Some investigators report the smallest p-value obtained from the three tests corresponding to the three genetic models, but such an approach inherently leads to inflated type 1 errors. Because of the small number of tests (three) and high correlation (functional dependence) among these tests, the procedures available for accounting for multiple tests are either too conservative or fail to meet the underlying assumptions (e.g., asymptotic multivariate normality or independence among the tests).

Results: We propose a method to calculate the exact p-value for each SNP using different genetic models. We performed simulations, which demonstrated the control of type 1 error and power gains using the proposed approach. We applied the proposed method to compute p-value for a polymorphism *eNOS -786T>C* which was shown to be associated with breast cancer risk.

Conclusions: Our findings indicate that the proposed method should be used to maximize power and control type 1 errors when analyzing genetic data using additive, dominant, and recessive models.

Keywords: Genetic association, Multiple testing, Cochran-Armitage trend test, Genetic models, Exact p-value

Background

Genome-wide association studies (GWAS) and candidate gene association studies are commonly performed to test the association of genetic variants with a particular phenotype. Typically, hundreds of thousands of single-nucleotide polymorphisms (SNPs) are tested for association in these studies. Associations between the SNPs and the phenotypes are determined on the basis of differences in allele frequencies between cases and controls [1]. Several statistical methods have been proposed to control the family-wise error rate (FWER) for multiple comparison testing.

A simple approximation can be used to obtain a FWER of α by utilizing the Bonferroni adjustment [2] of $\alpha^* = \frac{\alpha}{n}$ and using α^* as the threshold for significance for each test. Bonferroni adjustment tends to be conservative when the tests are correlated. In genetic association studies, the SNPs being tested are typically in linkage disequilibrium (LD), which leads to correlation among the tests. An alternative approximation to the Bonferroni adjustment is Sidak's correction [3,4], $\alpha^* = 1-(1-\alpha)^{\frac{1}{n}}$ which assumes independence among tests. Conneely and Boehnke [5] proposed a correction that does not assume independence among tests but assumes joint multivariate normality of all test statistics. Other methods to control the FWER include using the false discovery rate (FDR) [6,7].

In genetic association studies, three genetic models–additive, dominant, and recessive–are generally used to test each SNP using the Cochran-Armitage (CA) trend test [8-12]. In association studies the true underlying genetic model is unknown. Some investigators report the smallest p-value obtained from the three tests corresponding to the three genetic models. However, such a procedure inherently leads to an inflated type 1 error rate. Also, FDR-based methods to control FWER are not applicable in this situation because the hypotheses are

* Correspondence: sshete@mdanderson.org
[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[2]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

highly correlated, as the same SNP is tested using different genetic models.

Thus, there is a need to correct for multiple comparisons corresponding to the three genetic tests performed for testing the association of a single SNP. These three tests are not only correlated but also functionally dependent. The standard methods for correcting for multiple testing referred to above are either too conservative or fail to meet the assumptions underlying these methods (e.g., asymptotic multivariate normality, independence among tests). Several approaches have been proposed to account specifically for the multiple comparisons of these three genetic models [13-15]. However, these approaches assume asymptotic trivariate normality for the additive, dominant and recessive test statistics. While this is a reasonable approximation to correct for multiple comparisons, preliminary investigations regarding the joint distribution of the three test statistics revealed the following insights: 1) the joint distribution of the test statistics is discrete and the grids at which the probability mass function is positive is few and far between; 2) The distribution is highly multimodal in most of the situations, particularly, when the number of cases and controls are different and unimodal only in a handful of situations (e.g. when the number of cases and controls are equal). Therefore, we propose a method to compute the exact joint distribution of the three CA trend tests corresponding to the additive, dominant, and recessive genetic models. We used this joint distribution to compute the exact p-value for testing each SNP using the different genetic models. We performed simulations to demonstrate control of type 1 errors and power gains using the proposed approach. Finally, we applied the proposed approach to assess the significance of the association between a promoter polymorphism, *eNOS-786T>C* and breast cancer risk.

Methods

Consider a di-allelic SNP locus. The minor (deleterious) allele is labeled as *a,* and the major (normal) allele is labeled as *A*. The deleterious allele *a* is assumed to affect a phenotype *Z*, which takes the values of 0 or 1: $Z = 1$ indicates cases (affected) and $Z = 0$ indicates controls (unaffected). The observed genotype data for the SNP is one of three genotypes (A, A), (A, a), or (a, a). Let R_X denote the number of cases and R_Y denote the number of controls, with $R_X + R_Y = N$. Let X_1, X_2, X_3 and Y_1, Y_2, Y_3 be the number of individuals with genotypes *AA*, *Aa*, and *aa* in cases and controls, respectively. The data can be formulated in a 2×3 contingency table, as shown in Table 1. Let p_1, p_2, p_3 be the frequencies of genotypes, *AA*, *Aa* and *aa* in cases and q_1, q_2, q_3 be the frequencies of these three genotypes in controls. The values of p_i, q_i, i=1,2,3 can be estimated from the data as $p_i = \frac{X_i}{R_X}$ and $q_i = \frac{Y_i}{R_Y}$.

Table 1 Genotypic counts, parameterizations, and notations for various parameters used in the model formulation

	Genotype			
	AA	*Aa*	*aa*	Sum
Cases (*X*)	X_1	X_2	X_3	R_X
Controls (*Y*)	Y_1	Y_2	Y_3	R_Y
Sum	C_1	C_2	C_3	N

There have been many approaches in the literature for testing the association between a SNP and disease status. The CA test for trend [8] is generally the most popular and is available in most genetic analysis software packages, such as PLINK [16]. The test statistic for the CA test is as follows:

$$W = \sum_{i=1}^{3} t_i (R_Y X_i - R_X Y_i),$$

where the weight, t_i, is chosen on the basis of the genetic model considered: additive, dominant, or recessive. The additive model assumes the deleterious effect is linearly related to the number of deleterious alleles. The dominant model assumes the deleterious effect is related to the presence of the deleterious allele. And the recessive model assumes the deleterious effect is related to the presence of both the deleterious alleles. The weights $t = [t_1, t_2, t_3]$ corresponding to each of the models are as follows: additive model: $t = [0, 1, 2]$, dominant model: $t = [0, 1, 1]$, and recessive model: $t = [0, 0, 1]$ for genotypes *AA*, *Aa*, and *aa*, respectively. Let the three test statistics corresponding to the additive, dominant, and recessive models be T_1, T_2, and T_3, respectively.

The joint distribution

Each test statistic, T_1, T_2 and T_3, has an asymptotically normal univariate distribution. Therefore, the p-values for each of these tests can be obtained from their asymptotic distributions. However, reporting the smallest p-value obtained from testing T_1, T_2 and T_3, individually leads to an inflated type 1 error rate. If the exact joint distribution of the three tests is known, one can compute the exact p-value for the SNP that will account for the multiple correlated tests. We proceed to derive the joint distribution of the three test statistics, $T_1 = (R_Y X_2 - R_X Y_2) + 2(R_Y X_3 - R_X Y_3)$, and $T_2 = (R_Y X_2 - R_X Y_2) + (R_Y X_3 - R_X Y_3)$, and $T_3 = (R_Y X_3 - R_X Y_3)$. As $T_3 = T_1 - T_2$, we only need to derive the joint distribution of T_1 and T_2. It is reasonable to assume that the three genotype counts in cases (X_1, X_2, X_3) and the three genotype counts in controls (Y_1, Y_2, Y_3) follow a multinomial distribution, with probabilities (p_1, p_2, p_3)

and (q_1, q_2, q_3) respectively. Let $T = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}$, $X = \begin{pmatrix} X_2 \\ X_3 \end{pmatrix}$ and $Y = \begin{pmatrix} Y_2 \\ Y_3 \end{pmatrix}$. The test statistics can be written as $T = AX + BY$, where $A = \begin{bmatrix} R_Y & 2R_Y \\ R_Y & R_Y \end{bmatrix}$ and $B = \begin{bmatrix} -R_X & -2R_X \\ -R_X & -R_X \end{bmatrix}$. Then the joint probability mass function (pmf) of T_1, T_2 is given by

$$f_T(T_1, T_2) = \sum_{X_2=0}^{R_X} \sum_{X_3=0}^{R_X - X_2} f_X(X_2, X_3) f_Y(h(X_2, X_3, T_1, T_2))$$

where f_x,f_y are trinomial probability mass functions and $h(X, T) = B^{-1}T - B^{-1}AX$. The derivation of the joint pmf of T_1, T_2 is detailed in the Appendix. The p-value corresponding to the test statistic (t_1, t_2) can be computed by summing up the probabilities of the test statistics that are equally or less probable than the observed test statistic, which can be written as

$$pvalue(t_1, t_2) = \sum_{\langle T_1, T_2 : f_T(T_1, T_2) \leq f_T(t_1, t_2) \rangle} \sum_{T_1} \sum_{T_2} f_T(T_1, T_2)$$

The computation of the p-value using the above formula is nontrivial; however, there are a variety of computational optimizations and parallels to Fisher's exact test that can be used to drastically reduce the computational complexity (see details in the Appendix). Briefly, the CA trend test statistics form a system of constrained linear Diophantine equations. The computational optimizations presented in the Appendix are based on exploiting the properties of the linear Diophantine equations with trinomial constraints. The solution space of these equations corresponds to the discrete space of nonzero probabilities for the joint pmf. This discrete space has a pattern of overlapping triangles that can be enumerated based on R_X and R_Y counts (See Figures 1, 2, 3 and 4). To reduce the number of computations in the discrete space we first transformed the test statistics to be symmetric. The pattern of overlapping triangles depends on three different scenarios based on the greatest common divisor (GCD) of R_X and R_Y: 1. $GCD(R_X, R_Y) = 1$, 2. $GCD(R_X, R_Y) = R_X = R_Y$ and 3. $1 < GCD(R_X, R_Y) < \min(R_X, R_Y)$. In scenario 1 the triangles do not overlap, therefore the p-value can be evaluated most efficiently (Figures 1 and 2). In scenario 2 most of the triangles overlap and the discrete space of nonzero probabilities is sparse (Figure 4). In this scenario, we proposed an algorithm to exploit this aspect to calculate the exact p-value more efficiently. Scenario 3 is the most general case which uses the general optimizations of symmetricity and the triangle pattern (Figure 3). The algorithms to compute the exact p-values for each of the scenarios are detailed in the Appendix.

Simulations

We performed simulations to evaluate the performance of the proposed method and compared our approach with standard approaches used in the literature. All the simulation results were based on 1000 replicate data sets. Each replicate dataset comprised 1000 cases and 1000 controls. The disease status for each data set was obtained using the logistic regression model $logit(P(Z = 1)) = \beta_0 + \beta_1 X$,

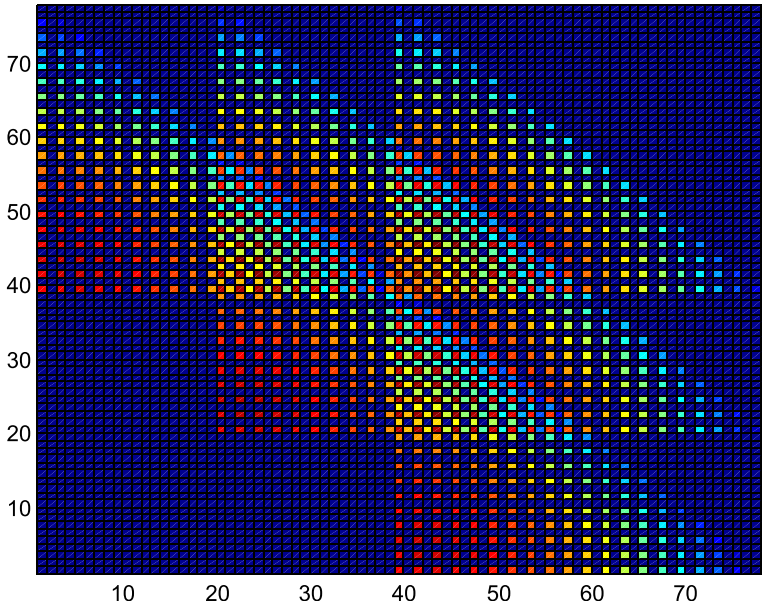

Figure 1 This figure depicts the probability mass function of the scenario with R_X = 19 and R_Y = 2. A pattern of six triangles can be visualized.

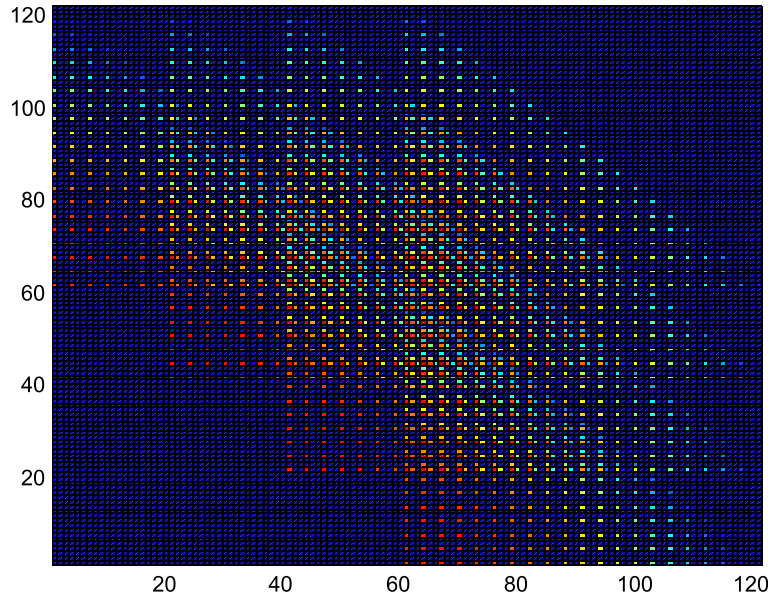

Figure 2 This figure depicts the probability mass function of the scenario with R_X = 20 and R_Y = 3. A pattern of ten triangles can be visualized.

where X is the indicator for genotype, Z is the disease status, β_0 is the intercept, and β_1 is the log odds ratio for the SNP. The genotype data for a SNP were simulated using a minor allele frequency (MAF) of 40% for the null hypothesis and two MAFs of 40% and 20% for the power comparisons. For the type 1 error comparisons, we simulated 1000 replicate datasets from the null hypothesis (i.e., the SNP was not associated with disease status), with $\beta_0 = -2.5$ and $\beta_1 = \log(1)$. For the power comparisons, we simulated 1000 replicate datasets for 40% and 20% MAFs from the alternate hypothesis (i.e., the SNP was associated with disease status) for each of the three scenarios: (1) additive model with odds ratio of 1.2, (2) dominant model with odds ratio of 1.3, and (3) recessive model with odds ratio of 1.3. The methods we compared were as follows: performing only

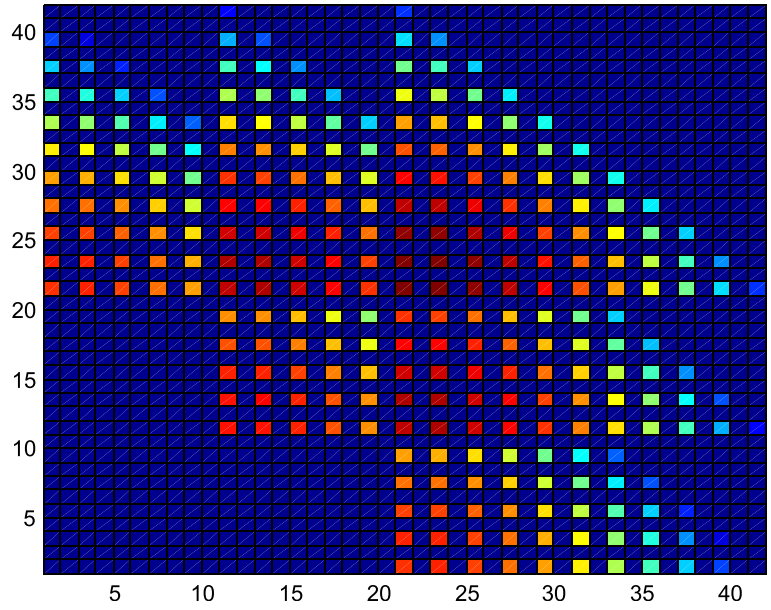

Figure 3 This figure depicts the probability mass function of the scenario with R_X = 10 and R_Y = 2 and a pattern of six overlapping triangles can be visualized.

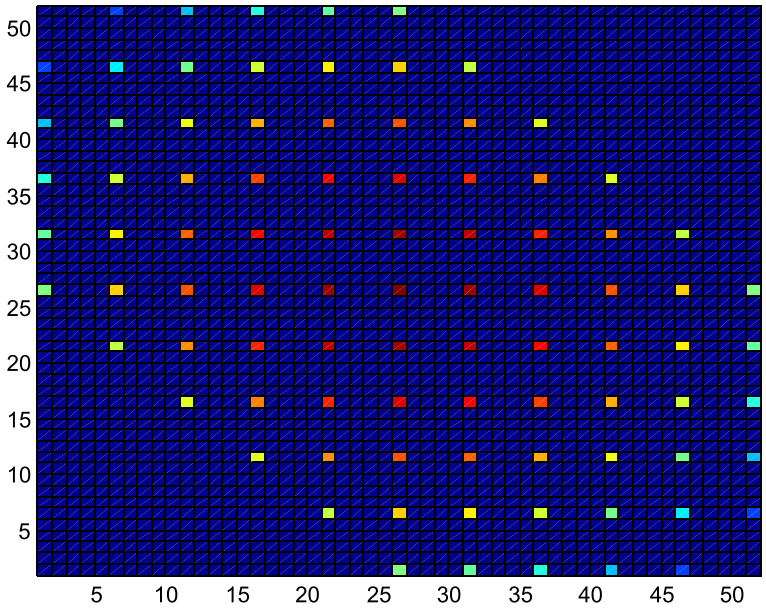

Figure 4 This figure depicts the probability mass function of the scenario with $R_X = 5$ and $R_Y = 5$. A pattern of 21 triangles can be visualized from the figure, where most of the triangles are overlapping completely or partially with one another.

additive analyses (additive-only), performing only dominant analyses (dominant-only), performing only recessive analyses (recessive-only), using the p-value based on reporting the smallest p-value of the three genetic models (min-p), using the Bonferroni correction approach, and using the proposed exact p-value method.

Results

The type 1 errors based on 1000 replicates from the null hypothesis are shown in Table 2. Analyses based on additive-only, dominant-only, and recessive-only models gave empirical type 1 errors of 0.044, 0.045, and 0.056, respectively, at the 0.05 level of significance. As expected, these models provided good control of type 1 errors because only one genetic model was tested in these analyses. The Bonferroni approach also had a well-controlled, but conservative, type 1 error (0.030 at the 0.05 level of significance). The min-p had a type 1 error of 0.105 at the 0.05 level of significance, which was very liberal and confirmed that the minimum p-value of the three genetic models is not a valid test. Finally, our proposed approach provided good control of the type 1 error (0.047 at the 0.05 level of significance).

Table 2 Type 1 error comparisons for different approaches at the 0.05 level of significance for 1000 replicates, each replicate representing a data set containing 1000 cases and 1000 controls

Method	α =0.05
Additive Only	0.044
Dominant Only	0.045
Recessive Only	0.056
Min-p	0.105
Bonferroni	0.030
Exact p-value	0.047

Min-p: p-value based on reporting the smallest p-value of the three genetic models.

The power comparisons based on 1000 replicates for the SNP data simulated using 40% and 20% MAFs for the three scenarios when the data were simulated using the additive, dominant, and recessive models, respectively, are shown in Table 3. The top and bottom panels of Table 3 depict the results for 40% and 20% MAFs, respectively. The min-p model was excluded from the comparison because of its inflated type 1 error. When the data were simulated using the additive genetic model (column 3, Table 3), and were analyzed using only the additive model, it had the highest powers (0.816 and 0.656 for 40% and 20% MAFs, respectively). However, when the data were analyzed using only the dominant model, the powers were 0.676 and 0.603 for 40% and 20% MAFs, respectively. Also, when the data were analyzed using only the recessive model the powers were 0.588 and 0.306 for 40% and 20% MAFs, respectively. The powers for the additive only analysis were the highest as expected because the true simulation model in this scenario was additive. However, the true model of disease inheritance is generally unknown and one performs analyses using all three genetic models. In this scenario, the proposed exact p-value method had powers of 0.743 and 0.584 for 40% and 20% MAFs, respectively, at the 0.05 level of significance, which were higher than the

Table 3 Power comparisons for different approaches at the 0.05 level of significance for 3 different simulation scenarios using genotypes coded as additive, dominant, and recessive, respectively, for 40% and 20% MAFs

MAF	Method	Genotype model		
		Additive model	Dominant model	Recessive model
		Odds ratio = 1.2	Odds ratio = 1.3	Odds ratio = 1.3
40%	Additive Only	0.816	0.660	0.410
	Dominant Only	0.676	0.803	0.116
	Recessive Only	0.588	0.158	0.589
	Bonferroni	0.721	0.671	0.452
	Exact p-value	0.743	0.726	0.517
20%	Additive Only	0.656	0.774	0.116
	Dominant Only	0.603	0.823	0.061
	Recessive Only	0.306	0.102	0.249
	Bonferroni	0.556	0.715	0.168
	Exact p-value	0.584	0.782	0.197

The results for each panel are based on 1000 replicates, with each replicate representing a data set containing 1000 cases and 1000 controls. MAF: Minor allele frequency.

Bonferroni method which had powers of 0.721 and 0.556 for 40% and 20% MAFs, respectively. Overall, powers of the proposed method were lower than additive model (true simulation model) but higher than those of the dominant-only, recessive-only, and Bonferroni correction approach.

When the data were simulated using the dominant model (column 4, Table 3), the additive-only, dominant-only and recessive-only analyses had powers of 0.660, 0.803, and 0.158, respectively, for 40% MAF and 0.774, 0.823, and 0.102, respectively for 20% MAF, at the 0.05 level of significance. Once again, as expected, the powers of the dominant-only analysis were the highest because the data were generated using the dominant model. The proposed exact p-value method had powers of 0.726 and 0.782 for the 40% and 20% MAFs, respectively, which were higher than the Bonferroni method which had powers of 0.671 and 0.715 for the 40% and 20% MAFs, respectively. When the data were simulated using the recessive model (column 5, Table 3), the additive-only, dominant-only and recessive-only analyses had powers of 0.410, 0.116, and 0.589, respectively, for 40% MAF and 0.116, 0.061, and 0.249, respectively, for 20% MAF. The proposed exact p-value method had powers of 0.517 and 0.197 for the 40% and 20% MAFs, respectively, which were higher than the Bonferroni method (0.452 and 0.168 for 40% and 20% MAFs, respectively).

We applied the proposed approach to assess the significance of the association between the promoter polymorphism *eNOS -786T>C* and sporadic breast cancer risk in non-Hispanic white women younger than 55 years from a breast cancer study performed by [17]. The study discovered that *eNOS -786T>C* was statistically significant for breast cancer (p=0.017) and included 421 breast cancer cases and 423 cancer free controls. The first panel in Table 4 depicts the genotype counts for TT, CT and CC genotypes in cases and controls for the *eNOS -786T>C*. The second panel in Table 4 reports the p-values for the *eNOS -786T>C* computed using the 5 different approaches: additive-only, dominant-only, recessive-only, Bonferroni and the proposed exact p-value method. The additive-only, dominant-only and recessive-only approaches had p-values of 0.0045, 0.0148 and 0.0313, respectively, and the Bonferroni adjusted p-value was 0.0135. For this SNP, the p-value computed using the proposed exact p-value method was 0.0021, which was more significant than the smallest of the three p-values obtained using the additive-, dominant-, and recessive-only analyses (Table 4).

Discussion

In this paper, we proposed a method to calculate the exact p-value for testing a single SNP using multiple genetic models. We recommend using the proposed method to maximize power and control type 1 errors when analyzing genetic data using additive, dominant, and recessive models. The proposed method is robust to model misspecifications and different SNP minor allele frequencies. Furthermore, similar to the computation of

Table 4 P-values computed using various approaches for association of *eNOS -786T> C* with breast cancer

Genotype Data for *eNOS -786T> C*			Method	p-value
	Controls	Cases	Additive Only	0.0045
Total	423	421	Dominant Only	0.0148
TT	203	167	Recessive Only	0.0313
CT	185	200	Bonferroni	0.0135
CC	35	54	Exact p-value	0.0021

Fisher's exact p-value, the proposed approach does not depend on asymptotic distributions.

In our simulation study, where replicate datasets were simulated using the null hypothesis, we found that the proposed method had well-controlled type 1 error probabilities. In contrast, the method of reporting the smallest p-value of the three genetic models tested had the highest false-positive rate and was found to be invalid. And, as expected, the type 1 error of the Bonferroni correction approach was well controlled but conservative, which typically led to a loss in power for identifying genetic variants.

We also simulated replicate datasets under an alternative hypothesis using the different genetic models: additive, dominant, and recessive. In these simulations, we observed that no single method: additive-only, dominant-only, or recessive-only, had higher power in all three scenarios. Each of these methods had higher power only when the model used to analyze the data was the same as the true model used to generate the data. However, because the true mode of disease inheritance is usually unknown, analyses using all three genetic models are necessary. In general, the Bonferroni correction approach led to higher power than using a model that did not correspond to the true model. The proposed exact p-value method was an improvement over the Bonferroni method. The conservativeness of the Bonferroni method may be due to its inability to account for the functional dependence between the three test statistics. In contrast, our proposed approach accounts for this functional dependence by computing p-values from the joint probability mass function. Finally, we analyzed breast cancer study data in which the polymorphism *eNOS -786T>C*, was found to be significant [17].

The computation time needed to obtain the exact p-value is substantial. The problem is very closely related to Fisher's exact test, and there are many patterns inherent in the structure of the problem that could be exploited to calculate the p-values more efficiently. In the Appendix, we present several novel optimization techniques to efficiently compute the test statistics in a reasonable time (e.g., approximately 15 min for a 1000 cases and 1000 controls dataset). The software to compute exact p-values is available at http://odin.mdacc.tmc.edu/~rtalluri/index.html.

Conclusions

In genetic association studies, three genetic models–additive, dominant, and recessive–are generally used to test each SNP using the Cochran-Armitage trend test. Reporting the minimum p-value of the three genetic models leads to inflated type 1 errors. We proposed an approach to compute the exact p-value when genomic data is analyzed using the three genetic models. The proposed approach leads to higher power while controlling the type 1 error.

Appendix

Optimization techniques for computing the exact p-value

Recall that X_1, X_2, X_3 and Y_1, Y_2, Y_3 are the number of individuals with genotypes *AA*, *Aa*, and *aa* in cases and controls, respectively, with $X_1 + X_2 + X_3 = R_X$ and $Y_1 + Y_2 + Y_3 = R_Y$. The three genotype counts in cases (X_1, X_2, X_3) and the three genotype counts in controls (Y_1, Y_2, Y_3) follow a multinomial distribution with probabilities (p_1, p_2, p_3) and (q_1, q_2, q_3), respectively. The probability mass function (pmf) of (X_1, X_2, X_3) is $f_X(X) = \frac{R_X!}{X_2!X_3!(R_X-X_2-X_3)!} p_1^{R_X-X_2-X_3} p_2^{X_2} p_3^{X_3}$ and the pmf of (Y_1, Y_2, Y_3) is $f_Y(Y) = \frac{R_Y!}{Y_2!Y_3!(R_Y-Y_2-Y_3)!} q_1^{R_Y-Y_2-Y_3} q_2^{Y_2} q_3^{Y_3}$. The three test statistics corresponding to the additive, dominant, and recessive models are, $T_1 = (R_Y X_2 - R_X Y_2) + 2(R_Y X_3 - R_X Y_3)$, $T_2 = (R_Y X_2 - R_X Y_2) + (R_Y X_3 - R_X Y_3)$, and $T_3 = (R_Y X_3 - R_X Y_3)$ respectively. As $T_3 = T_1 - T_2$, we only need to derive the joint distribution of T_1 and T_2. Let $T = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}$, $X = \begin{pmatrix} X_2 \\ X_3 \end{pmatrix}$, and $Y = \begin{pmatrix} Y_2 \\ Y_3 \end{pmatrix}$. The test statistics can be written as $T = AX + BY$, where $A = \begin{bmatrix} R_Y & 2R_Y \\ R_Y & R_Y \end{bmatrix}$ and $B = \begin{bmatrix} -R_X & -2R_X \\ -R_X & -R_X \end{bmatrix}$. We proceed to derive the joint probability mass function of $T = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}$.

Consider an n-dimensional discrete random vector G with pmf $f_G()$. Suppose we have a transformation from $G \rightarrow H$. The pmf $f_H()$ of the transformed variables H can be expressed as follows: [18]

$$f_H(H) = f_G\left(\varnothing^{-1}(H)\right)$$

This can be extended to the case where the dimensions of G and H are different, i.e., the transformation from $(X, Y) \rightarrow T$ is a linear transformation of the form $T = AX + BY$. The pmf of T is given by

$$f_T(T) = \sum_X f_X(X) f_Y(h(X,T)), \quad h(X,T) = B^{-1}T - B^{-1}AX$$

This can be simplified as:

$$h(X,Y) = \begin{pmatrix} Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} \frac{T_1}{R_X} - \frac{2T_2}{R_X} + \frac{R_Y X_2}{R_X} \\ \frac{T_2}{R_X} - \frac{T_1}{R_X} + \frac{R_Y X_3}{R_X} \end{pmatrix},$$

$$f_T(T_1, T_2) = \sum_{X_2=0}^{R_X} \sum_{X_3=0}^{R_X - X_2} f_X(X_2, X_3) f_Y(h(X_2, X_3, T_1, T_2))$$

Computing this pmf on all the possible values of (T_1, T_2) is prohibitively time consuming. Computational optimizations can be used to speed up the computations of the probability mass function. We list several optimization techniques below. The first optimization is to transform the pmf to be symmetric in (T_1, T_2), which reduces the computational burden by half. The original test statistics T_1 and T_2 are $T_1 = (R_Y X_2 - R_X Y_2) + 2(R_Y X_3 - R_X Y_3)$

and $T_2 = (R_Y X_2 - R_X Y_2) + (R_Y X_3 - R_X Y_3)$, respectively. The joint pmf of (T_1, T_2) is a one-to-one function of the joint distribution of any two orthogonal linear combinations of T_1 and T_2. So if we transform the test statistics T_1 and T_2 into

$$Z_1 = (R_Y X_3 - R_X Y_3),$$

$$Z_2 = (R_Y X_2 - R_X Y_2),$$

the resulting pmf of (Z_1, Z_2) is a one-to-one function of the pmf of (T_1, T_2). Hence, the p-value obtained will be the same when using (Z_1, Z_2) instead of (T_1, T_2). The resulting pmf of (Z_1, Z_2) can be derived using the same method as with (T_1, T_2).

The next computational optimization is to identify the values that can be taken by (Z_1, Z_2). The number of values (Z_1, Z_2) can take are finite and represented by the solution space of the equations

$$Z_1 = (R_Y X_3 - R_X Y_3),$$

$$Z_2 = (R_Y X_2 - R_X Y_2),$$

which depends on the values of R_X and R_Y. These equations are called linear Diophantine equations and have an infinite number of solutions [19]. But in our case we have multiple constraints on the equations, which reduce the solution space to a finite number of solutions. The constraints are

1. X_3, Y_3, X_2 and Y_2 are integers
2. X_3, Y_3, X_2 and $Y_2 \geq 0$
3. $X_3 + X_2 \leq R_X$
4. $Y_3 + Y_2 \leq R_Y$

On the basis of these four constraints the solution space can be calculated. While the exact solution space could not be found, it follows a pattern that can be enumerated.

Figure 1 depicts the pmf of the scenario with $R_X = 19$ and $R_Y = 2$ where a pattern of six triangles can be visualized from the figure. Similarly, Figure 2 depicts the pmf of the scenario with $R_X = 20$ and $R_Y = 3$, where a pattern of ten triangles can be visualized from the picture. This trend can be generalized for all values of R_X and R_Y.

Generalizing the above scenario, there are $[1 + 2 + \cdots + (R_Y + 1) = \frac{(R_Y+1)(R_Y+2)}{2}]$ triangles for the solution space. In each triangle, there are $[1 + 2 + \cdots + (R_X + 1) = \frac{(R_X+1)(R_X+2)}{2}]$ elements that correspond to all possible combinations of $X_3 + X_2 \leq R_X$. In each triangle, the values of Y_3 and Y_2 are constant and the $\frac{(R_Y+1)(R_Y+2)}{2}$ triangles correspond to all possible combinations of $Y_3 + Y_2 \leq R_Y$, which make up the whole solution space.

Another important fact is that these triangles may overlap, reducing the solution space, which is depicted in Figures 3 and 4. Figure 3 depicts the pmf of the scenario with $R_X = 10$ and $R_Y = 2$ where a pattern of six triangles can be visualized from the figure. The overlap of the triangles can be observed when compared to Figure 1. Figure 4 depicts the pmf of the scenario with $R_X = 5$ and $R_Y = 5$ where a pattern of 21 triangles can be visualized from the figure, where most of the triangles are overlapping one another. The additional computational burden is to determine where the solution space triangles overlap and how many triangles are overlapping at a particular location. This is a function of the greatest common divisor (GCD) of R_X and R_Y. If R_X and R_Y are co-prime (GCD=1), only three triangles overlap at a single point ($Z_1 = 0$, $Z_2 = 0$) which requires no additional computation. When R_X and R_Y are not co-prime, the triangles overlap at multiples of the GCD of R_X and R_Y. In this scenario, multiple values of X_3, Y_3, X_2, and Y_2 contribute to the same (Z_1, Z_2).

In an ideal scenario, the total number of computations required to compute the pmf of (Z_1, Z_2) is $\frac{(R_Y+1)(R_Y+2)}{2} \frac{(R_X+1)(R_X+2)}{2} \approx \frac{R_X^2 R_Y^2}{4}$, which can be computed in approximately 15 minutes for $R_X = 1000$ and $R_Y = 1000$ using a computer with a 3.4-GHz processor and 8 GB of RAM. However, the amount of storage required for the solution space far exceeds the hardware capabilities available. In light of this limitation, computational optimizations should be employed to avoid storing the whole solution space. This limitation leads to three possible scenarios:

1. $GCD(R_X, R_Y) = 1$
2. $GCD(R_X, R_Y) = R_X = R_Y$
3. $GCD(R_X, R_Y) < \min(R_X, R_Y)$

Scenario 1

When R_X and R_Y are co-prime, the triangles only overlap at a single point ($Z_1 = 0, Z_2 = 0$); therefore, we can independently evaluate each of the possible values of the solution space. The p-value is the probability of obtaining a test statistic at least as extreme as the one observed, so we evaluate the probabilities of each of the possible values of the test statistics one at a time. Hence, the p-value is the sum of all the probabilities of test statistics that are lower than the probability of the observed test statistic. Using this procedure there is no need to store any data, which leads to faster computation of the p-value from the joint distribution.

Scenario 2

When R_X and R_Y are equal, most of the triangles overlap with each other. But a pattern has been observed in this

scenario, which is shown in Figure 5, where $R_X = 10$ and $R_Y = 10$. As seen in Figure 4, the solution space is very sparse and only requires computation of the colored cells. The possible solution space is spaced R_X apart. So if we condense the possible solution space, the solution space is as shown in Figure 5. Figure 5 shows the number of triangles overlapping at each point in the solution space. Only half of the matrix needs to be computed, as the other half is symmetric. The algorithm to compute the p-value is as follows.

Algorithm:

1. Let $R_X = R_Y = R$. The solution space can then be constrained to a matrix with $2R + 1$ rows and $2R + 1$ columns. Let the center of the matrix correspond to the test statistic $(Z_1 = 0, Z_2 = 0)$.
2. Now, as we can see from Figure 5, we need to compute the colored cells in quadrants 3 and 4. In quadrant 3, the cells with the same number of overlapping triangles are placed diagonally, and in quadrant 4, they are placed horizontally and then vertically. We exploit the pattern that follows from the same number of triangles overlapping at a particular cell.
3. For $i = 1$: R start at $(Z_1 = -(R - i), Z_2 = -1)$. Find the possible combinations of X_3, Y_3, X_2 and Y_2 that contribute to the cell corresponding to $(Z_1 = -(R - i), Z_2 = -1)$. Compute the probabilities for the cells along the diagonal path in quadrant 3, until $Z_1 = 0$. Here X_3 and X_2 remain the same; hence, it is trivial to compute the probabilities for each cell.
4. Then in quadrant 4, compute the probabilities for the cells along the horizontal path until $Z_1 = R - (i - 1)$; here X_3 remains the same and $X_{2new} = X_2 + Z_2$.
5. Then continue vertically until $Z_2 = 0$; here X_3 and X_2 remain the same.

This algorithm reduces the computational burden by computing the possible combinations of X_3, Y_3, X_2 and Y_2 that contribute to all the cells only R times, as opposed to computing once for each cell (approximately $4R^2$ times).

Scenario 3

This is the general scenario where $GCD(R_X, R_Y) < \min(R_X, R_Y)$. Several patterns that can be used to reduce the computational burden that could be applied for a particular GCD were found, but these could not be generalized to all the possible situations. We instead use a straightforward approach to determine the p-value for each of the possible solutions for (Z_1, Z_2). The algorithm is as follows:

1. For each possible (Z_1, Z_2) compute the triangles that contribute to this particular point.
2. Add up the probabilities of each of the elements of these triangles to compute the p-value of that particular (Z_1, Z_2).

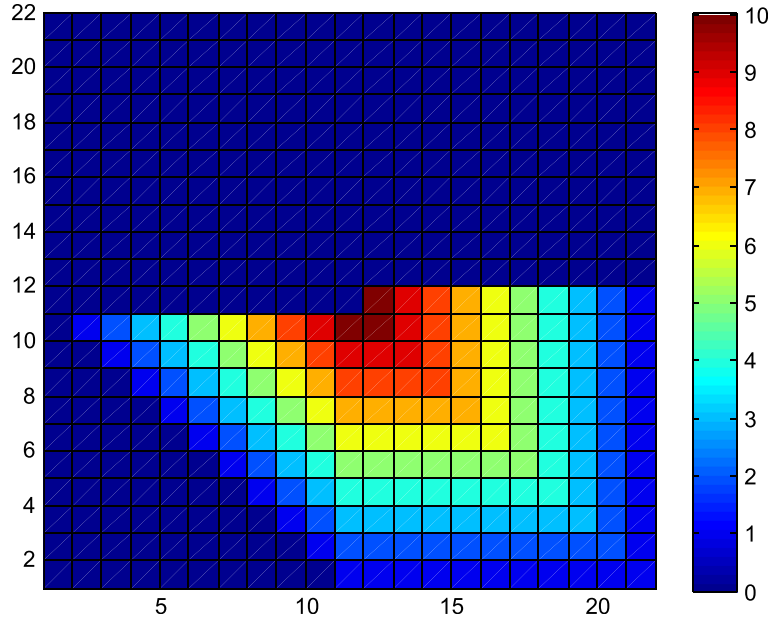

Figure 5 This figure shows the number of triangles overlapping at each point in the condensed solution space in the scenario with $R_X = 10$ and $R_Y = 10$, where most of the triangles are overlapping completely or partially with one another.

Competing interests
We declare that there are no competing interests.

Authors' contributions
RT and SS conceived and designed the study. RT implemented the method. RT and JW performed simulations. RT and SS wrote the paper. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by National Institutes of Health grants R01CA131324 (SS), NIH R25 DA026120 (SS), and R01DE022891 (SS). This research was supported in part by Barnhart Family Distinguished Professorship in Targeted Therapy (SS). This research was supported in part by a cancer prevention fellowship for Rajesh Talluri supported by a grant from the National Institute of Drug Abuse (NIH R25 DA026120).

Received: 5 March 2014 Accepted: 27 May 2014
Published: 20 June 2014

References
1. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT: **Basic statistical analysis in genetic case–control studies.** *Nat Protoc* 2011, **6**(2):121–133.
2. Dunn OJ: **Multiple comparisons among means.** *J Am Stat Assoc* 1961, **56**(293):52–64.
3. Sidak Z: **On multivariate normal probabilities of rectangles - their dependence on correlations.** *Ann Math Stat* 1968, **39**(5):1425–1434.
4. Sidak Z: **Probabilities of rectangles in multivariate student distributions - their dependence on correlations.** *Ann Math Stat* 1971, **42**(1):169–175.
5. Conneely KN, Boehnke M: **So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests.** *Am J Hum Genet* 2007, **81**(6):1158–1168.
6. Benjamini Y, Hochberg Y: **Controlling the false discovery rate - a practical and powerful approach to multiple testing.** *J Roy Stat Soc B Methods* 1995, **57**(1):289–300.
7. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Ann Math Stat* 2001, **29**(4):1165–1188.
8. Agresti A: *Categorical Data Analysis.* New York: John Wiley & Sons; 2002.
9. Armitage P: **Tests for linear trends in proportions and frequencies.** *Biometrics* 1955, **11**(3):375–386.
10. Barrett JH, Iles MM, Harland M, Taylor JC, Aitken JF, Andresen PA, Akslen LA, Armstrong BK, Avril MF, Azizi E, Bakker B, Bergman W, Bianchi-Scarra G, Bressac-de Paillerets B, Calista D, Cannon-Albright LA, Corda E, Cust AE, Debniak T, Duffy D, Dunning AM, Easton DF, Friedman E, Galan P, Ghiorzo P, Giles GG, Hansson J, Hocevar M, Hoiom V, Hopper JL, *et al*: **Genome-wide association study identifies three new melanoma susceptibility loci.** *Nat Genet* 2011, **43**(11):1108–1113.
11. Cochran WG: **Some methods for strengthening the common X2 tests.** *Biometrics* 1954, **10**(4):417–451.
12. Lewis CM, Knight J: **Introduction to genetic association studies.** *Cold Spring Harb Protoc* 2012, **2012**(3):297–306.
13. Freidlin B, Zheng G, Li ZH, Gastwirth JL: **Trend tests for case–control studies of genetic markers: power, sample size and robustness.** *Hum Hered* 2002, **53**(3):146–152.
14. Gonzalez JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X, Moreno V: **Maximizing association statistics over genetic models.** *Genet Epidemiol* 2008, **32**(3):246–254.
15. Hothorn LA, Hothorn T: **Order-restricted scores test for the evaluation of population-based case–control studies when the genetic model is unknown.** *Biometrical J* 2009, **51**(4):659–669.
16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**(3):559–575.
17. Lu JC, Wei QY, Bondy ML, Yu TK, Li DH, Brewster A, Shete S, Sahin A, Meric-Bernstam F, Wang LE: **Promoter polymorphism (−786T > C) in the endothelial nitric oxide synthase gene is associated with risk of sporadic breast cancer in non-Hispanic white women age younger than 55 years.** *Cancer* 2006, **107**(9):2245–2253.
18. Casella G, Berger RL: *Statistical Inference.* 2nd edition. Australia ; Pacific Grove, CA: Thomson Learning; 2002:46–54.
19. Mordell LJ: *Diophantine Equations.* Academic P: London, New York; 1969.

doi:10.1186/1471-2156-15-75
Cite this article as: Talluri *et al.*: Calculation of exact p-values when SNPs are tested using multiple genetic models. *BMC Genetics* 2014 **15**:75.