Testing Genetic Association by Regressing Genotype over Multiple Phenotypes

Kai Wang*

Department of Biostatistics, University of Iowa, Iowa City, Iowa, United States of America

Abstract

Complex disorders are typically characterized by multiple phenotypes. Analyzing these phenotypes jointly is expected to be more powerful than dealing with one of them at a time. A recent approach (O'Reilly et al. 2012) is to regress the genotype at a SNP marker on multiple phenotypes and apply the proportional odds model. In the current research, we introduce an explicit expression for the score test statistic and its non-centrality parameter that determines its power. Same simulation studies as those reported in Galesloot et al. (2014) were conducted to assess its performance. We demonstrate by theoretical arguments and simulation studies that, despite its potential usefulness for multiple phenotypes, the proportional odds model method can be less powerful than regular methods for univariate traits. We also introduce an implementation of the proposed score statistic in an R package named iGasso.

Citation: Wang K (2014) Testing Genetic Association by Regressing Genotype over Multiple Phenotypes. PLoS ONE 9(9): e106918. doi:10.1371/journal.pone.0106918

Editor: Zhongxue Chen, Indiana University Bloomington, United States of America

Received June 13, 2014; **Accepted** July 26, 2014; **Published** September 15, 2014

Data Availability: The author confirms that all data underlying the findings are fully available without restriction. The statistical test proposed in the manuscript has been implemented in R language. It is freely available at http://cran.r-project.org/web/packages/iGasso/index.html. The R code is available upon request from the corresponding author.

Funding: This research is funded by the National Institute for Environmental Health Sciences through the University of Iowa Environmental Health Sciences Research Center, NIEHS/NIH P30 ES05605. The funder has no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* Email: kai-wang@uiowa.edu

Introduction

Complex human disorders are often characterized by multiple phenotypes. Some of them might be categorical while others might be continuous. For instance, patients with Bardet-Biedl syndrome often suffer from vision loss, hypertension and high cholesterol level caused by obesity, polydactyly, and other abnormalities. In order to map the genetic variants underlying such disorders, it is highly desirable to analyze all available phenotypes simultaneously. However, it is challenging to jointly model these phenotypes, especially when they are of different data types [1].

Let $\mathbf{y}$ denote a $k \times 1$ vector of k phenotypes on an individual and g his/her genotype at a marker. If all the components of $\mathbf{y}$ are continuous, one may use MANOVA given genetype g. When the components of $\mathbf{y}$ are of mixed data types, the choices are limited. One popular method is to first analyze each component individually and then combine the test statistics through a meta-analysis [2,3]. These methods model the phenotype vector $\mathbf{y}$ in terms of the genetic data g.

For a single-nucleotide polymorphism (SNP), the distribution of its genotype g is trinomial. It is appealing to model g as a function of $\mathbf{y}$ [4,5]. Furthermore, since there is a natural ordering in the three genotypes at the SNP (assuming that the possibility of over-dominance is ignorable), one can use the ordinal logistic regression (a.k.a., the proportional odds model or the cumulative logit model) [6]. One immediate advantage of using the proportional odds model is that, unlike other methods, there is no need to make assumptions on the genetic effect such as additive, dominant, or recessive. The usefulness of this approach has been demonstrated via analyses of various data [5]. It is one of the best currently available multivariate methods [7]. However, it is the slowest one [7].

The test used in [5] is the likelihood ratio test (LRT). It involves numerical maximization under both the null hypothesis and the alternative hypothesis. We introduce a score test statistic using standard statistics theory. This statistic is asymptotically equivalent to the likelihood ratio test but computationally much faster due to the availability of its explicit expression, a feature useful in genome-wide association analysis. This explicit expression also gives insight on how the proportional odds model works in the context of genetic association analysis.

This report is organized as follows. We first introduce an explicit form of the score statistic and its non-centrality parameter. The form of this score statistic provides some insights on the ability of this method to detect association. The performance of the this score statistic is evaluated using the same simulation scenarios used in [7]. Finally, we consider an important case where the phenotype vector $\mathbf{y}$ is univariate and binary to see how this test works for univariate phenotypes.

Results

The score statistic

The genetic data are assumed to come from a biallelic marker such as a single-nucleotide polymorphism (SNP). Let a denote the reference allele and A the other. For simplicity, we use 0, 1, and 2 to represent genotypes AA, Aa, and aa, respectively. Regardless of the data types of the components of $\mathbf{y}$, the genotype g follows a trinomial distribution. In most cases, the effect of an allele is monotonic. That is, as the number of a alleles increases from 0 to 2, the effect of genotypes AA, Aa, and aa is non-decreasing or non-increasing. Over-dominance effect exists but is rather rare. Given this consideration, we model the genotype g in terms of phenotype $\mathbf{y}$ using the proportional odds model [5].

Let $\pi_j(\mathbf{y}) = \Pr(g = j|\mathbf{y})$ denote the probability that an individual's genotype g is j given phenotypic value $\mathbf{y}$. In the current situation, the proportional odds model models the cumulative probabilities $\pi_0(\mathbf{y})$ and $\pi_0(\mathbf{y}) + \pi_1(\mathbf{y})$ jointly as follows:

$$\log\left(\frac{\pi_0(\mathbf{y})}{1-\pi_0(\mathbf{y})}\right) = \alpha_1 + \boldsymbol{\beta}^t\mathbf{y}, \tag{1}$$

$$\log\left(\frac{\pi_0(\mathbf{y}) + \pi_1(\mathbf{y})}{\pi_2(\mathbf{y})}\right) = \alpha_2 + \boldsymbol{\beta}^t\mathbf{y}. \tag{2}$$

Here α_1 and α_2 are intercepts and $\boldsymbol{\beta}$ is a vector of coefficients. This model implies $\alpha_2 \geq \alpha_1$ because $\pi_0(\mathbf{y}) + \pi_1(\mathbf{y}) \geq \pi_0(\mathbf{y})$. Since $\pi_0(\mathbf{y}) + \pi_1(\mathbf{y}) = 1 - \pi_2(\mathbf{y})$, an alternative form of equation (2) is

$$\log\left(\frac{1-\pi_2(\mathbf{y})}{\pi_2(\mathbf{y})}\right) = \alpha_2 + \boldsymbol{\beta}^t\mathbf{y}.$$

Equation (1) models the effect of phenotype y on the odds of genotype AA versus Aa or aa while equation (2) models the effect of phenotype y on the odds of genotype aa versus Aa or AA. We have

$$\pi_0(\mathbf{y}) = \frac{\exp(\alpha_1 + \boldsymbol{\beta}^t\mathbf{y})}{1 + \exp(\alpha_1 + \boldsymbol{\beta}^t\mathbf{y})},$$

$$\pi_2(\mathbf{y}) = \frac{1}{1 + \exp(\alpha_2 + \boldsymbol{\beta}^t\mathbf{y})},$$

and $\pi_1(\mathbf{y})$ is determined by $\pi_1(\mathbf{y}) = 1 - \pi_0(\mathbf{y}) - \pi_2(\mathbf{y})$. This model assumes that the difference of the left hand side of (1) or (2) for two phenotype vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ depends only on $\boldsymbol{\beta}^t(\mathbf{y}_1 - \mathbf{y}_2)$ and is independent of genotype aa or AA:

$$\log\left(\frac{\pi_0(\mathbf{y}_1)}{1-\pi_0(\mathbf{y}_1)}\right) - \log\left(\frac{\pi_0(\mathbf{y}_2)}{1-\pi_0(\mathbf{y}_2)}\right) = \boldsymbol{\beta}^t(\mathbf{y}_1 - \mathbf{y}_2), \tag{3}$$

$$= \log\left(\frac{1-\pi_2(\mathbf{y}_1)}{\pi_2(\mathbf{y}_1)}\right) - \log\left(\frac{1-\pi_2(\mathbf{y}_2)}{\pi_2(\mathbf{y}_2)}\right). \tag{4}$$

That is,

$$\frac{\pi_0(\mathbf{y}_1)\pi_2(\mathbf{y}_1)}{(1-\pi_0(\mathbf{y}_1))(1-\pi_2(\mathbf{y}_1))} = \frac{\pi_0(\mathbf{y}_2)\pi_2(\mathbf{y}_2)}{(1-\pi_0(\mathbf{y}_2))(1-\pi_2(\mathbf{y}_2))}$$
$$= \exp(\alpha_1 - \alpha_2)$$

does not depend on the value of $\mathbf{y}$.

Let i be the index for the ith individual in a sample of size n, the log-likelihood function is

$$l(\alpha_1, \alpha_2, \boldsymbol{\beta}; \{\mathbf{y}_i\}) = \sum_{j=0,1,2} \sum_{i:g_i=j} \log(\pi_j(\mathbf{y}_i)).$$

The hypotheses of interest are

$$H_0 : \boldsymbol{\beta} = 0, \;\; H_1 : \boldsymbol{\beta} \neq 0. \tag{5}$$

These hypotheses can be tested by the likelihood ratio statistic. To introduce the score statistic, define

$$\mathbf{w} = \left(\sum_{i:g_i=0} \mathbf{y}_i + \bar{\pi}_2 \sum_{i:g_i \neq 0} \mathbf{y}_i\right) - \left(\sum_{i:g_i=2} \mathbf{y}_i + \bar{\pi}_0 \sum_{i:g_i \neq 2} \mathbf{y}_i\right),$$

where $\bar{\pi}_0$ and $\bar{\pi}_2$ are the sample proportions of the genotypes for which $g = 0$ and 2, respectively. $\mathbf{w}$ is the difference of two weighted summations of $\mathbf{y}_i$. The summation in the first pair of parentheses weights $\mathbf{y}_i$ with $g_i = 0$ more than other $\mathbf{y}_i$s (i.e., 1 versus $\bar{\pi}_2$) while the summation in the second pair of parentheses weights $\mathbf{y}_i$ with $g_i = 2$ more (i.e., 1 versus $\bar{\pi}_0$). Let $\bar{\pi}_1 = 1 - \bar{\pi}_0 - \bar{\pi}_2$. It is shown in the Methods section that a score statistic for testing hypotheses (5) is

$$S = \frac{1}{n(1-\bar{\pi}_0)(1-\bar{\pi}_1)(1-\bar{\pi}_2)} \mathbf{w}^t\mathbf{V}^{-1}\mathbf{w},$$

where

$$\mathbf{V} = n^{-1}\sum_{i=1}^{n} \mathbf{y}_i^t\mathbf{y}_i - n^{-1}\sum_{i=1}^{n} \mathbf{y}_i^t \cdot n^{-1}\sum_{i=1}^{n} \mathbf{y}_i$$

is the sample variance matrix of $\mathbf{y}_i$, $i = 1, \ldots, n$. The non-centrality parameter of S is

$$NCP = \frac{E(\mathbf{w})^t\mathbf{V}^{-1}E(\mathbf{w})}{n(1-\pi_0)(1-\pi_1)(1-\pi_2)},$$

where the expectation in $E(\mathbf{w})$ is taken under the alternative. This NCP can be used to compute power at significance level α in the following way:

$$\Pr(X > \chi^2_{1-\alpha,k})$$

where X follows a chi-square distribution with $df = k$ and non-centrality parameter NCP and $\chi^2_{1-\alpha,k}$ is the critical value from a chi-square distribution with $df = k$ and non-centrality parameter 0.

Simulation Study

The simulation study consists of two parts. In the first part, multivariate phenotypes were simulated the same way as in [7]. Genotype data at a single SNP were simulated with minor allele frequency q under the assumption Hardy-Weinberg equilibrium. Three phenotypes, denoted by y_1,y_2, and y_3, were simulated using the following relationship:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} =$$
$$MVN\left(\begin{bmatrix} rG \cdot a_1 \\ a_2 \\ a_3 \end{bmatrix}, \begin{bmatrix} 1-h_1^2 & rE_{12}\sqrt{1-h_1^2}\sqrt{1-h_2^2} & rE_{13}\sqrt{1-h_1^2}\sqrt{1-h_2^2} \\ & 1-h_2^2 & rE_{23}\sqrt{1-h_1^2}\sqrt{1-h_2^2} \\ Symmetry & & 1-h_3^2 \end{bmatrix}\right),$$

where rE_{12}, rE_{13}, and rE_{23} are pre-spedfied residual correlations between phenotypes excluding the QTL effect, $rG = 1$ or -1 controls the effect direction of y_1 (those of y_2 and y_3 are fixed at 1) and $a_j = \sqrt{h_j^2/2(1-q)q}$. Three scenarios were considered with only 1, 2, or all phenotypes were associated with the SNP. See [7] for more details. For each simulated data set, the LRT *p*-value is obtained from the R package MultiPhen and the score *p*-value is obtained from the R package iGasso. Empirical rejection rates over 1,000 replicates are reported in Table 1. The performance of the score test is very similar to that of LRT. The empirical power are very close to the power of LRT reported in [7].

Is the proportional odds model always more powerful than the usual tests of association? To address this question, simulation were conducted on univariate phenotypes. The description of the simulation studies is provided in the Methods section. In addition to the proposed score statistic, the other test statistics used in the simulation include the Pearson's chi-square test, the Cochran-Armitage trend test, and the likelihood ratio test for the proportional odds model. The number of simulation replicates is fixed at 10,000. The sample size is 2,000. Half of the subjects are cases and half are controls. The simulated type I error rate for all these statistics is reported in Table 2. The empirical rejection rates are very close to their nominal levels, which are 0.1, 0.01, and 0.005. The simulated power is presented in Figures 1, 2, and 3. It is striking to see that for recessive models the proportional odds model is the least powerful. For instance, when prevalence $K=0.1$ and minor allele frequency $p=0.3$, the power are 0.486 for Chi-Square test, 0.353 for Trend test, and 0.19 for both LRT and Score test. For other two models, there are situations it is more powerful than other methods. The simulated power for the S statistic is in line with the calculated power reported in Table 3.

Discussion

In this report, we introduced a score test statistic for the proportional odds model for testing the association between a SNP and multiple phenotypes and provided an implementation of this statistic. Same simulation studies as those reported in [7] were conducted to assess it performance. We also did simulation analyses to study the performance of proportional odds model for univariate phenotypes which is covered by [5]. Although appealing to studies on multiple phenotypes, this method may be less powerful for univariate traits than regular methods. For case-control data, our results suggest that the traditional Pearson's chi-square test and the Cochran-Armitage trend test are preferred when the disease allele frequency is less than 0.5 and the disease is recessive.

Nonetheless, the proportional odds model method provides a convenient way for analyzing multiple phenotypes, especially when these phenotypes are of different types [5]. If the proportional odds assumption is of concern, one may remove this assumption and adopt a multinomial logistic regression. For our simulation studies, the multinomial logistic regression would be equivalent to the Pearson's chi-square test statistic. There are quite few implementations of the multinomial logistic regression, for instance, the multinom function in R package nnet.

Methods

Derivation of the score statistic

The first-order derivatives of the log-likelihood function $l(\alpha_1,\alpha_2,\boldsymbol{\beta})$ are

$$\frac{\partial l}{\partial \alpha_1}=\sum_{i:g_i=0}(1-\pi_0(\mathbf{y}_i))-\sum_{i:g_i=1}\frac{\pi_0(\mathbf{y}_i)(1-\pi_0(\mathbf{y}_i))}{1-\pi_0(\mathbf{y}_i)-\pi_2(\mathbf{y}_i)},$$

$$\frac{\partial l}{\partial \alpha_2}=\sum_{i:g_i=1}\frac{\pi_2(\mathbf{y}_i)(1-\pi_2(\mathbf{y}_i))}{1-\pi_0(\mathbf{y}_i)-\pi_2(\mathbf{y}_i)}-\sum_{i:g_i=2}(1-\pi_2(\mathbf{y}_i)),$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}}=\sum_{i:g_i=0}(1-\pi_0(\mathbf{y}_i))\mathbf{y}_i+\sum_{i:g_i=1}(\pi_2(\mathbf{y}_i)-\pi_0(\mathbf{y}_i))\mathbf{y}_i-\sum_{i:g_i=2}(1-\pi_2(\mathbf{y}_i))\mathbf{y}_i,$$

and the second-order derivatives are

$$\frac{\partial^2 l}{\partial \alpha_1^2}=-\sum_{i:g_i=0}\pi_0(\mathbf{y}_i)(1-\pi_0(\mathbf{y}_i))-\sum_{i:g_i=1}\left[\frac{\pi_0(\mathbf{y}_i)(1-\pi_0(\mathbf{y}_i))(1-2\pi_0(\mathbf{y}_i))}{\pi_1(\mathbf{y}_i)}+\frac{\pi_0(\mathbf{y}_i)^2(1-\pi_0(\mathbf{y}_i))^2}{\pi_1(\mathbf{y}_i)^2}\right],$$

Table 1. Non-centrality value and the associated power (presented in parentheses) for models used in the simulation studies.

	Frequency of Allele *a*	Effect of Allele *a*		
K	*p*	**Recessive**	**Additive**	**Dominant**
0.01	0.1	—	4.6780 (0.2597)	14.4110 (0.8387)
	0.3	2.8697 (0.1329)	9.7282 (0.6225)	18.5977 (0.9339)
	0.4	6.9507 (0.4323)	—	—
0.1	0.1	—	5.7000 (0.3374)	17.6383 (0.9182)
	0.3	3.4771 (0.1730)	11.7847 (0.7343)	22.5123 (0.9737)
	0.4	8.4233 (0.5379)	—	—

The relative genotype risks are $f_1/f_0=1, f_2/f_0=1.5$ for recessive models; $f_1/f_0=1.5, f_2/f_0=1.5$ for dominant models; and $f_1/f_0=1.25, f_2/f_0=1.5$ for additive models.
doi:10.1371/journal.pone.0106918.t001

Table 2. Simulated type I error rate in the case of a univariate phenotype under various generating models.

Penetrance (*K*)	Frequency of Allele *a* (*p*)	Significance Level	Test Statistic			
			Trend	**Chi-Square**	**LRT**	**Score**
0.01	0.1	0.1	0.085	0.095	0.091	0.091
		0.01	0.007	0.007	0.006	0.007
		0.005	0.005	0.004	0.004	0.004
	0.3	0.1	0.077	0.091	0.082	0.082
		0.01	0.009	0.008	0.010	0.010
		0.005	0.006	0.002	0.005	0.005
0.1	0.1	0.1	0.096	0.099	0.101	0.102
		0.01	0.011	0.009	0.010	0.010
		0.005	0.006	0.006	0.007	0.007
	0.3	0.1	0.116	0.117	0.109	0.108
		0.01	0.009	0.011	0.010	0.010
		0.005	0.006	0.007	0.003	0.003
0.01	0.1	0.1	0.085	0.095	0.091	0.091
		0.01	0.007	0.007	0.006	0.007
		0.005	0.005	0.004	0.004	0.004
	0.3	0.1	0.077	0.091	0.082	0.082
		0.01	0.009	0.008	0.010	0.010
		0.005	0.006	0.002	0.005	0.005
0.1	0.1	0.1	0.096	0.099	0.101	0.102
		0.01	0.011	0.009	0.010	0.010
		0.005	0.006	0.006	0.007	0.007
	0.3	0.1	0.116	0.117	0.109	0.108
		0.01	0.009	0.011	0.010	0.010
		0.005	0.006	0.007	0.003	0.003
0.01	0.1	0.1	0.085	0.095	0.091	0.091
		0.01	0.007	0.007	0.006	0.007
		0.005	0.005	0.004	0.004	0.004
	0.3	0.1	0.077	0.091	0.082	0.082
		0.01	0.009	0.008	0.010	0.010
		0.005	0.006	0.002	0.005	0.005
0.1	0.1	0.1	0.096	0.099	0.101	0.102
		0.01	0.011	0.009	0.010	0.010
		0.005	0.006	0.006	0.007	0.007
	0.3	0.1	0.116	0.117	0.109	0.108
		0.01	0.009	0.011	0.010	0.010
		0.005	0.006	0.007	0.003	0.003

The test statistics are: Trend — Cochran-Armitage trend test; Chi-Square — Pearson's chi-square test; LRT — the likelihood ratio test for the proportional odds model computed by using the polr function in the R package MASS; Score — the proposed score statistic computed by using the SNPass.test function in the R package iGasso.

doi:10.1371/journal.pone.0106918.t002

$$\frac{\partial^2 l}{\partial\alpha_1\partial\alpha_2} = \sum_{i:g_i=1}\frac{\pi_0(\mathbf{y}_i)(1-\pi_0(\mathbf{y}_i))\pi_2(\mathbf{y}_i)(1-\pi_2(\mathbf{y}_i))}{\pi_1(\mathbf{y}_i)^2},$$

$$\frac{\partial^2 l}{\partial\alpha_1\partial\boldsymbol{\beta}^t} = -\sum_{i:g_i\neq 2}\pi_0(\mathbf{y}_i)(1-\pi_0(\mathbf{y}_i))\mathbf{y}_i^t,$$

$$\frac{\partial^2 l}{\partial\alpha_2^2} = -\sum_{i:g_i=1}\left[\frac{\pi_2(\mathbf{y}_i)(1-\pi_2(\mathbf{y}_i))(1-2\pi_2(\mathbf{y}_i))}{\pi_1(\mathbf{y}_i)} + \frac{\pi_2(\mathbf{y}_i)^2(1-\pi_2(\mathbf{y}_i))^2}{\pi_1(\mathbf{y}_i)^2}\right] - \sum_{i:g_i=2}\pi_2(\mathbf{y}_i)(1-\pi_2(\mathbf{y}_i)),$$

$$\frac{\partial^2 l}{\partial\alpha_2\partial\boldsymbol{\beta}^t} = -\sum_{i:g_i\neq 0}\pi_2(\mathbf{y}_i)(1-\pi_2(\mathbf{y}_i))\mathbf{y}_i^t,$$

$$\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t} = -\sum_{i:g_i\neq 2}\pi_0(\mathbf{y}_i)(1-\pi_0(\mathbf{y}_i))\mathbf{y}_i\mathbf{y}_i^t - \sum_{i:g_i\neq 0}\pi_2(\mathbf{y}_i)(1-\pi_2(\mathbf{y}_i))\mathbf{y}_i\mathbf{y}_i^t.$$

Under $H_0:\boldsymbol{\beta}=\mathbf{0}$, $\pi_j(\mathbf{y}_i)$,$j=0$, 1, 2 no longer depends on $\mathbf{y}_i$. So their values are simply denoted by π_0,π_1, and π_2, respectively. Let

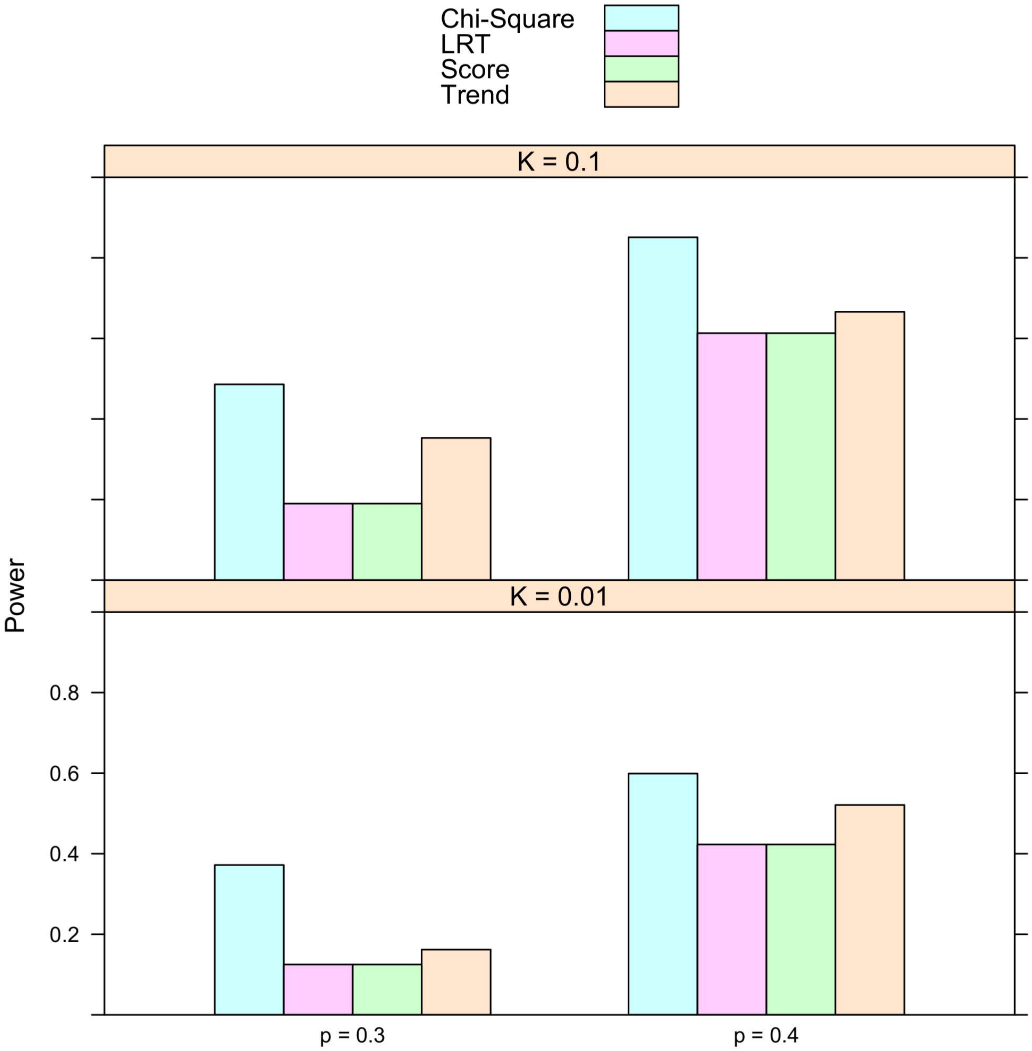

Figure 1. Simulated power for recessive model. The relative genotype risks are $f_1/f_0=1, f_2/f_0=1.5$. *K* represents disease prevalence and *p* is the frequency of allele *a*. The abbreviations for the test statistics are the same as in Table 2.
doi:10.1371/journal.pone.0106918.g001

$\boldsymbol{\alpha}=(\alpha_1,\alpha_2)^t$. The expected Fisher information matrix evaluated at $H_0:\boldsymbol{\beta}=0$ is

$$\mathbf{I}=-\begin{pmatrix} E(\partial^2 l/\partial\alpha_1^2) & E(\partial^2 l/\partial\alpha_1\partial\alpha_2) & E(\partial^2 l/\partial\alpha_1\partial\boldsymbol{\beta}^t) \\ E(\partial^2 l/\partial\alpha_1\partial\alpha_2) & E(\partial^2 l/\partial\alpha_2^2) & E(\partial^2 l/\partial\alpha_2\partial\boldsymbol{\beta}^t) \\ E(\partial^2 l/\partial\alpha_1\partial\boldsymbol{\beta}) & E(\partial^2 l/\partial\alpha_2\partial\boldsymbol{\beta}) & E(\partial^2 l/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t) \end{pmatrix}$$

$$=\frac{(1-\pi_0)(1-\pi_2)}{\pi_1}\begin{pmatrix} n\pi_0(1-\pi_0) & -n\pi_0\pi_2 & \pi_0\pi_1\sum_{i=1}^n \mathbf{y}_i^t \\ -n\pi_0\pi_2 & n\pi_2(1-\pi_2) & \pi_1\pi_2\sum_{i=1}^n \mathbf{y}_i^t \\ \pi_0\pi_1\sum_{i=1}^n \mathbf{y}_i & \pi_1\pi_2\sum_{i=1}^n \mathbf{y}_i & \pi_1(1-\pi_1)\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^t \end{pmatrix}$$

$$=\begin{pmatrix} \mathbf{I}_{\alpha\alpha} & \mathbf{I}_{\alpha\boldsymbol{\beta}} \\ \mathbf{I}_{\alpha\boldsymbol{\beta}}^t & \mathbf{I}_{\boldsymbol{\beta\beta}} \end{pmatrix},$$

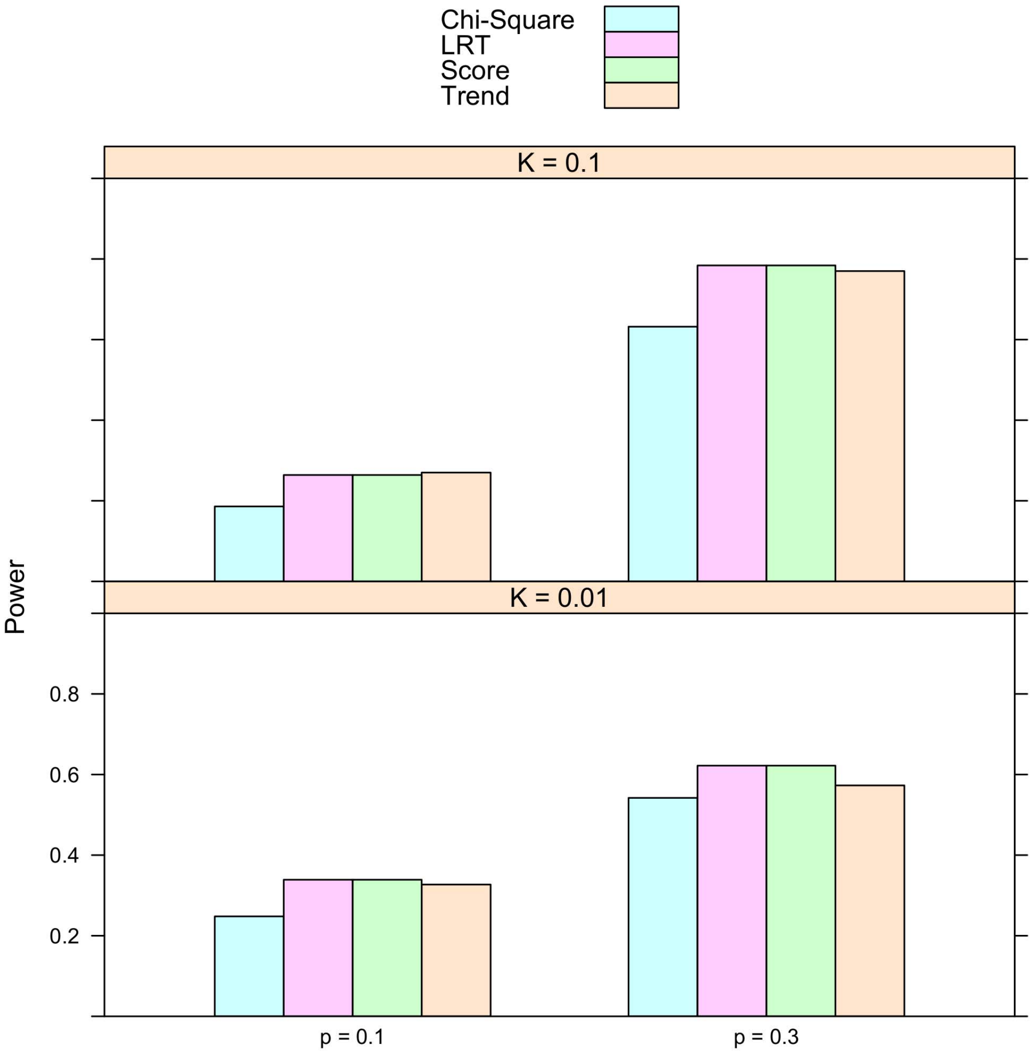

Figure 2. Simulated power for additive model. The relative genotype risks are $f_1/f_0 = 1.25, f_2/f_0 = 1.5$. *K* represents disease prevalence and *p* is the frequency of allele *a*. The abbreviations for the test statistics are the same as in Table 2.
doi:10.1371/journal.pone.0106918.g002

where the matrix partition is in an obvious manner. By standard asymptotic theory, the score statistic is

$$
\begin{aligned}
S &= \mathbf{w}^t[\mathbf{I}_{\beta\beta} - \mathbf{I}^t_{\alpha\beta}\mathbf{I}^{-1}_{\alpha\alpha}\mathbf{I}_{\alpha\beta}]^{-1}\mathbf{w} \\
&= \frac{1}{n(1-\pi_0)(1-\pi_1)(1-\pi_2)}\mathbf{w}^t\mathbf{V}^{-1}\mathbf{w},
\end{aligned}
$$

where $\mathbf{w}$ is $\partial l/\partial \boldsymbol{\beta}$ evaluated at H_0:

$$
\begin{aligned}
\mathbf{w} &= (1-\pi_0)\sum_{i:g_i=0}\mathbf{y}_i + (\pi_2-\pi_0)\sum_{i:g_i=1}\mathbf{y}_i - (1-\pi_2)\sum_{i:g_i=2}\mathbf{y}_i \\
&= \left(\sum_{i:g_i=0}\mathbf{y}_i + \pi_2\sum_{i:g_i\neq 0}\mathbf{y}_i\right) - \left(\sum_{i:g_i=2}\mathbf{y}_i + \pi_0\sum_{i:g_i\neq 2}\mathbf{y}_i\right).
\end{aligned}
$$

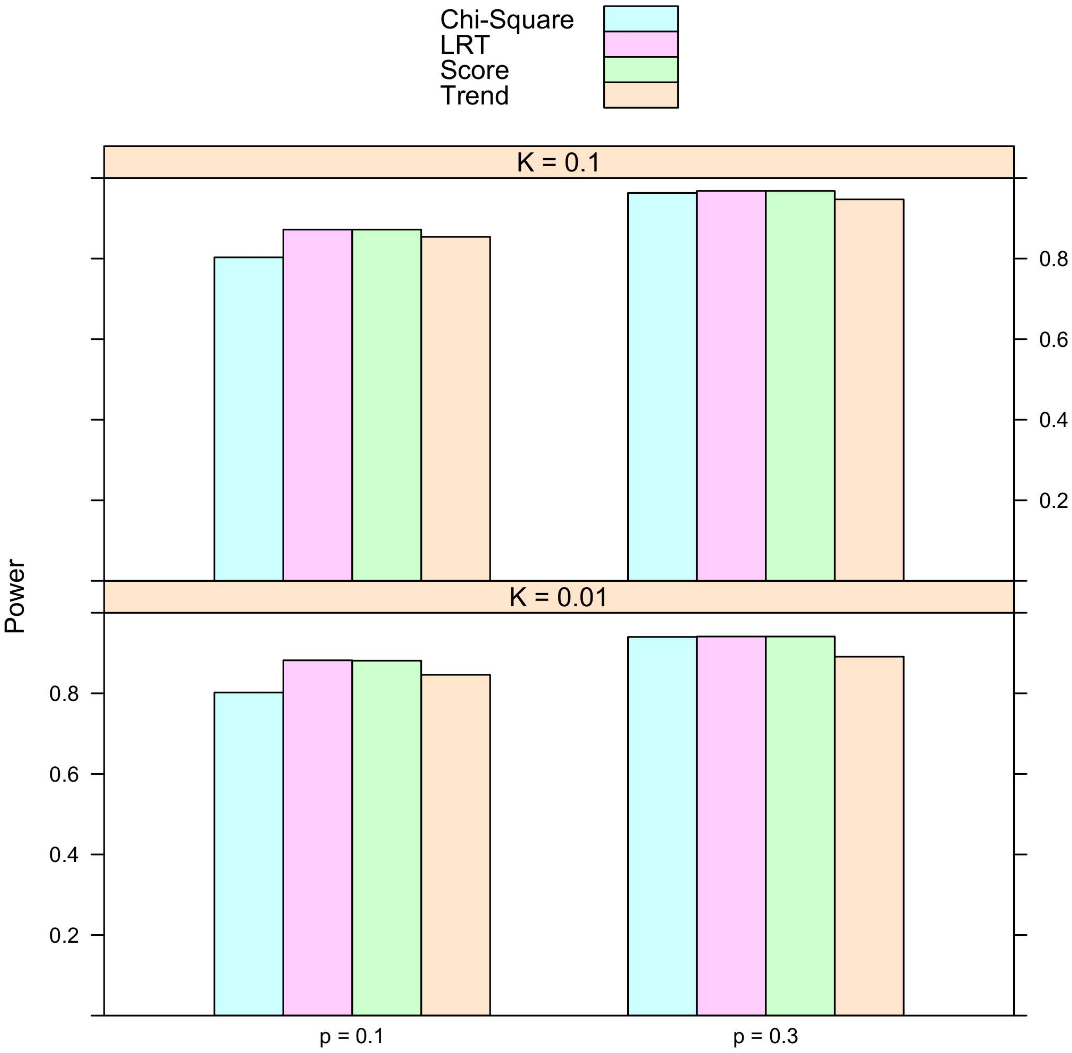

Figure 3. Simulated power for dominant model. The relative genotype risks are $f_1/f_0 = 1.5, f_2/f_0 = 1.5$. *K* represents disease prevalence and *p* is the frequency of allele *a*. The abbreviations for the test statistics are the same as in Table 2.
doi:10.1371/journal.pone.0106918.g003

The unknown values of π_0, π_1, and π_2 are estimated by their sample genotype proportions, respectively.

Simulation Studies

Here is a description of the simulation procedure for the case of a dichotomous phenotype. Suppose the trait is Mendelian. Let p_0, p_1, and p_2 denote the frequencies of genotypes AA, Aa and aa in general population and f_0, f_1, and f_2 their penetrances, respectively. The prevalence of the disease would be $K = p_0 f_0 + p_1 f_1 + p_2 f_2$. The genotype frequencies in cases are $\pi_{1j} = p_j f_j / K, j = 0,1,2$, and in controls are $\pi_{0j} = p_j(1-f_j)/(1-K), j = 0,1,2$. In this situation, the variance of y is $\phi(1-\phi)$ where ϕ is the proportion of cases. The non-centrality parameter (NCP) of test statistic S is equal to

$$NCP = \frac{E(w)^2}{n(1-\pi_0)(1-\pi_1)(1-\pi_2)\cdot\phi(1-\phi)}$$
$$= \frac{n\phi(1-\phi)}{[K(1-K)]^2}\cdot\frac{[p_2(K-f_2)-p_0(K-f_0)+p_0p_2(f_2-f_0)]^2}{(1-\pi_0)(1-\pi_1)(1-\pi_2)}.$$

Let p be the population frequency of allele a. Assuming Hardy-Weinberg equilibrium in the population, the frequencies of genotypes AA, Aa, and aa are $p_0 = (1-p)^2, p_1 = 2p(1-p)$, and $p_2 = p^2$, respectively. Let $\gamma_i = f_i/f_0$, $i = 1,2$, be the relative risk of genotype i to genotype 0. A data generating model is completely determined by K, p, γ_1, and γ_2. This is because $f_0 = K/(p_0 + \gamma_1 p_1 + \gamma_2 p_2)$, $f_1 = \gamma_1 f_0$, and $f_2 = \gamma_2 f_0$. Hence the genotype frequencies in cases and controls are determined and

Table 3. Simulated power under the same simulation scenarios of [7] over 1,000 replicates.

			Scenario					
			I		II		III	
rG	*rE*	MAF *q*	LRT	Score	LRT	Score	LRT	Score
1	0.0	0.01	0.125	0.125	0.184	0.183	0.278	0.278
		0.40	0.141	0.141	0.198	0.197	0.257	0.257
	0.3	0.01	0.127	0.129	0.185	0.185	0.184	0.183
		0.40	0.124	0.123	0.177	0.178	0.187	0.185
	0.7	0.01	0.221	0.221	0.268	0.267	0.129	0.129
		0.40	0.215	0.211	0.292	0.292	0.138	0.136
−1	0.0	0.01	—	—	0.202	0.201	0.289	0.288
		0.40	—	—	0.194	0.192	0.269	0.262
	0.3	0.01	—	—	0.267	0.266	0.344	0.344
		0.40	—	—	0.277	0.273	0.351	0.354
	0.7	0.01	—	—	0.589	0.589	0.694	0.694
		0.40	—	—	0.548	0.542	0.717	0.713

In scenario I, only phenotype 1 is associated with the SNP. $rG = 1$ or -1 does not affect the simulation results. Only results with $rG = 1$ are shown.

doi:10.1371/journal.pone.0106918.t003

data can be simulated. We consider a dominance model ($\gamma_1 = \gamma_2$), a recessive model ($\gamma_1 = 1$), and an additive model ($\gamma_1 = (1+\gamma_2)/2$). The NCPs for the models used in simulation are reported in Table 3. So are the power associated with these NCPs.

R function SNPass.test

The R function SNPass.test in the package iGasso implements the proposed score statistic. R users can download and install iGasso from CRAN (http://cran.r-project.org/) or any CRAN mirror.

Acknowledgments

The author thanks an associate editor and four referees for their helpful comments.

Author Contributions

Conceived and designed the experiments: KW. Performed the experiments: KW. Analyzed the data: KW. Contributed reagents/materials/analysis tools: KW. Contributed to the writing of the manuscript: KW. Designed the software used in analysis: KW.

References

1. Zhu W, Zhang H (2009) Why do we test multiple traits in genetic association studies. Journal of the Korean Statistical Society 38: 1–10.
2. Xu X, Lu Tian L (2003) Combining dependent tests for linkage or association across multiple phenotypic traits. Biostatistics 4: 223–229.
3. Yang Q, Wu H, Guo CY, Fox CS (2010) Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genet Epidemiol 34: 444–454.
4. Wang K, Huang J (2011) Treating phenotype as given: A simple resampling method for genome-wide association studies. Genetic Analysis Workshop 17 5: S60.
5. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, et al. (2012) MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS. PLoS ONE 7: e34861.
6. Agresti A (2002) Categorical Data Analysis. John Wiley & Sons, Inc., 2nd edition.
7. Galesloot TE, van Steen K, Kiemeney LALM, Janss LL, Vermeulen SH (2014) A comparison of multivariate genome-wide association methods. PLoS ONE 9: e95923.