Testing Genetic Association by Regressing Genotype over Multiple Phenotypes

Kai Wang

University of Iowa
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.

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 $y$ 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 $y$ are continuous, one may use MANOVA given genotype $g$. When the components of $y$ 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 $y$ 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 $y$ [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 $y$ 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 $y$, the genotype $g$ follows a trinomial distribution. In most cases, the effect of an allele is...
monotonic. That is, as the number of 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 y using the proportional odds model [5].

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

\[
\log \left( \frac{\pi_0(y)}{1 - \pi_0(y)} \right) = z_1 + \beta'y, \tag{1}
\]

\[
\log \left( \frac{\pi_0(y) + \pi_1(y)}{\pi_2(y)} \right) = z_2 + \beta'y. \tag{2}
\]

Here \( z_1 \) and \( z_2 \) are intercepts and \( \beta \) is a vector of coefficients. This model implies \( z_2 \geq z_1 \) because \( \pi_0(y) + \pi_1(y) \geq \pi_0(y) \). Since \( \pi_0(y) + \pi_1(y) = 1 - \pi_2(y) \), an alternative form of equation (2) is

\[
\log \left( \frac{1 - \pi_2(y)}{\pi_2(y)} \right) = z_2 + \beta'y. \tag{3}
\]

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

\[
\pi_0(y) = \frac{\exp(z_1 + \beta'y)}{1 + \exp(z_1 + \beta'y)},
\]

\[
\pi_1(y) = \frac{1}{1 + \exp(z_2 + \beta'y)},
\]

and \( \pi_2(y) \) is determined by \( \pi_1(y) = 1 - \pi_0(y) - \pi_2(y). \) This model assumes that the difference of the left hand side of (1) or (2) for two phenotype vectors \( y_1 \) and \( y_2 \) depends only on \( \beta' (y_1 - y_2) \) and is independent of genotype aa or AA:

\[
\log \left( \frac{\pi_0(y_1)}{1 - \pi_0(y_1)} \right) - \log \left( \frac{\pi_0(y_2)}{1 - \pi_0(y_2)} \right) = \beta' (y_1 - y_2). \tag{4}
\]

That is,

\[
\frac{\pi_0(y_1) \pi_2(y_1)}{(1 - \pi_0(y_1))(1 - \pi_2(y_1))} = \frac{\pi_0(y_2) \pi_2(y_2)}{(1 - \pi_0(y_2))(1 - \pi_2(y_2))}
\]

\[
= \exp(z_1 - z_2)
\]

does not depend on the value of \( y \).

Let \( i \) be the index for the \( i \)th individual in a sample of size \( n \), the log-likelihood function is

\[
\ell(x_1, x_2, \beta; \{y_i\}) = \sum_{j=0,1,2} \sum_{i \ni y_i=j} \log(\pi_j(y_i)).
\]

The hypotheses of interest are

\[
H_0 : \beta = 0, \quad H_1 : \beta \neq 0. \tag{5}
\]

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

\[
w = \left( \sum_{i \ni y_i=0} y_i + \pi_2 \sum_{i \ni y_i=1} y_i \right) - \left( \sum_{i \ni y_i=2} y_i + \pi_0 \sum_{i \ni y_i=1} y_i \right),
\]

where \( \pi_0 \) and \( \pi_2 \) are the sample proportions of the genotypes for which \( g = 0 \) and 2, respectively, \( w \) is the difference of two weighted summations of \( y \). The summation in the first pair of parentheses weights \( y \) with \( g_i = 0 \) more than other \( y_k \) (i.e., in \( 1 \) versus \( \pi_2 \) while the summation in the second pair of parentheses weights \( y \) with \( g_i = 2 \) more (i.e., 1 versus \( \pi_0 \)). Let \( \pi_1 = 1 - \pi_0 - \pi_2 \). It is shown in the Methods section that a score statistic for testing hypotheses (5) is

\[
S = \frac{1}{n(1 - \pi_0)(1 - \pi_1)(1 - \pi_2)} w' V^{-1} w,
\]

where

\[
V = n^{-1} \sum_{i=1}^{n} y_i y_i' - n^{-1} \sum_{i=1}^{n} y_i' n^{-1} \sum_{i=1}^{n} y_i
\]

is the sample variance matrix of \( y_i \), \( i = 1, \ldots, n \). The non-centrality parameter of \( S \) is

\[
NCP = \frac{E(w)' V^{-1} E(w)}{n(1 - \pi_0)(1 - \pi_1)(1 - \pi_2)},
\]

where the expectation in \( E(w) \) is taken under the alternative. This NCP can be used to compute power at significance level \( \alpha \) 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:

\[
\begin{pmatrix}
    y_1 \\
    y_2 \\
    y_3
\end{pmatrix} =
\begin{pmatrix}
    r_{G, a_1} \\
    a_2 \\
    a_3
\end{pmatrix} \begin{pmatrix}
    1-h_i^2 & 0 & 0 \\
    0 & 1-h_i^2 & 0 \\
    0 & 0 & 1-h_i^2
\end{pmatrix} \begin{pmatrix}
    1-h_i^2 & r_{E_12} \sqrt{1-h_i^2} & r_{E_13} \sqrt{1-h_i^2} \sqrt{1-h_i^2} \\
    r_{E_12} \sqrt{1-h_i^2} & 1-h_i^2 & 0 \\
    r_{E_13} \sqrt{1-h_i^2} \sqrt{1-h_i^2} & 0 & 1-h_i^2
\end{pmatrix}.
\]

where \( r_{E_{12}}, r_{E_{13}}, \) and \( r_{E_{23}} \) are pre-specified residual correlations between phenotypes excluding the QTL effect, \( r_{G} = 1 \) 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, 0.337 for Trend test, and 0.250 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 its 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 $l(x_1, x_2, \beta)$ are

$$\frac{\partial l}{\partial x_1} = \sum_{i \in \Gamma_i=0} (1 - \pi_0(y_i)) - \sum_{i \in \Gamma_i=1} \pi_0(y_i)(1 - \pi_0(y_i)),$$

$$\frac{\partial l}{\partial x_2} = \sum_{i \in \Gamma_i=1} \pi_2(y_i)(1 - \pi_2(y_i)) - \sum_{i \in \Gamma_i=2} (1 - \pi_2(y_i)),$$

$$\frac{\partial l}{\partial \beta} = \sum_{i \in \Gamma_i=0} (1 - \pi_0(y_i))y_i + \sum_{i \in \Gamma_i=1} (\pi_2(y_i) - \pi_0(y_i))y_i - \sum_{i \in \Gamma_i=2} (1 - \pi_2(y_i))y_i,$$

and the second-order derivatives are

$$\frac{\partial^2 l}{\partial x_1^2} = -\sum_{i \in \Gamma_i=0} \pi_0(y_i)(1 - \pi_0(y_i))$$

$$- \sum_{i \in \Gamma_i=1} \left[ \frac{\pi_0(y_i)(1 - \pi_0(y_i))(1 - 2\pi_0(y_i))}{\pi_1(y_i)} + \frac{\pi_0(y_i)^2(1 - \pi_0(y_i))^2}{\pi_1(y_i)^2} \right],$$

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

<table>
<thead>
<tr>
<th>$K$</th>
<th>$p$</th>
<th>Frequency of Allele $a$</th>
<th>Effect of Allele $a$</th>
<th>Recessive</th>
<th>Additive</th>
<th>Dominant</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.1</td>
<td>—</td>
<td>2.8697 (0.1329)</td>
<td>9.7282</td>
<td>18.5977</td>
<td>22.5123</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>6.9507 (0.4323)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3</td>
<td>1.4771 (0.1730)</td>
<td>11.7847 (0.7343)</td>
<td>22.5123</td>
<td>0.9339</td>
<td>0.8387</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
<td>8.4233 (0.5379)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Penetrance (K)</th>
<th>Frequency of Allele (p)</th>
<th>Significance Level</th>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Trend</td>
</tr>
<tr>
<td>0.01</td>
<td>0.1</td>
<td>0.1</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.01</td>
<td>0.1</td>
<td>0.1</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.01</td>
<td>0.1</td>
<td>0.1</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
</tbody>
</table>

Under $H_0$ : $\beta = 0$, $\pi_j(y_i) = 0$, $1$, $2$ no longer depends on $y_i$. So their values are simply denoted by $\pi_0, \pi_1$, and $\pi_2$, respectively. Let
The expected Fisher information matrix evaluated at $H_0 : \beta = 0$ is

$$\mathbf{I} = \begin{pmatrix}
E(\hat{\gamma}_1^2/\hat{\gamma}_1) & E(\hat{\gamma}_1^2/\hat{\gamma}_2) & E(\hat{\gamma}_1^2/\hat{\gamma}_1\hat{\gamma}_2) \\
E(\hat{\gamma}_2^2/\hat{\gamma}_1) & E(\hat{\gamma}_2^2/\hat{\gamma}_2) & E(\hat{\gamma}_2^2/\hat{\gamma}_1\hat{\gamma}_2) \\
E(\hat{\gamma}_1\hat{\gamma}_2/\hat{\gamma}_1) & E(\hat{\gamma}_1\hat{\gamma}_2/\hat{\gamma}_2) & E(\hat{\gamma}_1\hat{\gamma}_2/\hat{\gamma}_1\hat{\gamma}_2)
\end{pmatrix}
$$

$$= \begin{pmatrix}
\pi_1 & \pi_2 & \pi_1\pi_2 \\
\pi_2 & \pi_1\pi_2 & \pi_1\pi_2 \\
\pi_1\pi_2 & \pi_1\pi_2 & \pi_1\pi_2
\end{pmatrix}
$$

$$= \pi_1 \begin{pmatrix}
(1-\pi_0)(1-\pi_2) \\
\pi_0(1-\pi_0) & -\pi_0\pi_2 & \pi_0\pi_1 \sum_{i=1}^{n} y_i^2 \\
-\pi_0\pi_2 & \pi_2(1-\pi_2) & \pi_1\pi_2 \sum_{i=1}^{n} y_i f_i \\
\pi_0\pi_1 \sum_{i=1}^{n} y_i f_i & \pi_1\pi_2 \sum_{i=1}^{n} y_i f_i & \pi_1(1-\pi_1) \sum_{i=1}^{n} y_i f_i
\end{pmatrix}
$$

$$= \begin{pmatrix}
\mathbf{I}_{xx} & \mathbf{I}_{x\theta} \\
\mathbf{I}_{\theta x} & \mathbf{I}_{\theta\theta}
\end{pmatrix}.$$
where the matrix partition is in an obvious manner. By standard asymptotic theory, the score statistic is

\[ S = w' \{I_\beta \beta - I_\beta \beta I_{ss}^{-1} I_\beta \beta \}^{-1} w \]

\[ = \frac{1}{n(1 - \pi_0)(1 - \pi_1)(1 - \pi_2)} w' V^{-1} w, \]

where \( w \) is \( \hat{c} \beta \hat{c} \beta \) evaluated at \( H_0 \):

\[ w = (1 - \pi_0) \sum_{i \in g_i = 0} y_i + (\pi_2 - \pi_0) \sum_{i \in g_i = 1} y_i - (1 - \pi_2) \sum_{i \in g_i = 2} y_i \]

\[ = \left( \sum_{i \in g_i = 0} y_i + \pi_2 \sum_{i \in g_i \neq 0} y_i \right) - \left( \sum_{i \in g_i = 2} y_i + \pi_0 \sum_{i \in g_i \neq 2} y_i \right). \]
The unknown values of $p_0$, $p_1$, and $p_2$ 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 $K = p_0f_0 + p_1f_1 + p_2f_2$. The genotype frequencies in cases are $p_1j = f_1f_j = f_0$, $j = 0, 1, 2$, and in controls are $p_0j = p_0(1-f_j)/(1-K), j = 0, 1, 2$. In this situation, the variance of $y$ is $\phi(1-\phi)$ where $\phi$ is the proportion of cases. The non-centrality parameter $\text{NCP}$ of test statistic $S$ is equal to

$$\text{NCP} = \frac{E(w)^2}{n(1-p_0)(1-p_1)(1-p_2)\phi(1-\phi)}$$

$$= \frac{n\phi(1-\phi)\left[p_2(K-f_2) - p_0(K-f_0) + p_0p_2(f_2-f_0)\right]^2}{(1-p_0)(1-p_1)(1-p_2)}.$$ 

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 $K$, $p$, $\gamma_1$, and $\gamma_2$. This is because $f_0 = K/(p_0 + \gamma_1p_1 + \gamma_2p_2)$, $f_1 = \gamma_1f_0$, and $f_2 = \gamma_2f_0$. Hence the genotype frequencies in cases and controls are determined and

![Figure 3. Simulated power for dominant model.](image-url)
Table 3. Simulated power under the same simulation scenarios of [7] over 1,000 replicates.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LRT</td>
<td>Score</td>
</tr>
<tr>
<td>rG = 1</td>
<td>0.0</td>
<td>0.125</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0.141</td>
<td>0.141</td>
</tr>
<tr>
<td>rG = 0.3</td>
<td>0.0</td>
<td>0.127</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0.124</td>
<td>0.123</td>
</tr>
<tr>
<td>rG = 0.7</td>
<td>0.0</td>
<td>0.221</td>
<td>0.221</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0.215</td>
<td>0.211</td>
</tr>
<tr>
<td>rG = -1</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rG = 0.3</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rG = 0.7</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In scenario I, only phenotype 1 is associated with the SNP. rG = 1 or rG = -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 ($c_1 = \gamma_2$), a recessive model ($c_1 = 1$), and an additive model ($c_1 = (1 + \gamma_2)/2$). 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.

References