From Interaction to Co-Association —A Fisher $r$-To-$z$ Transformation-Based Simple Statistic for Real World Genome-Wide Association Study

Zhongshang Yuan¹, Hong Liu², Xiaoshuai Zhang¹, Fangyu Li¹, Jinghua Zhao³, Furen Zhang², Fuzhong Xue¹*

¹ Department of Epidemiology and Health Statistics, School of Public Health, Shandong University, Jinan, China, ² Shandong Provincial Institute of Dermatology and Venereology, Shandong Academy of Medical Science, Jinan, China, ³ MRC Epidemiology Unit and Institute of Metabolic Science, Cambridge, United Kingdom

Abstract

Currently, the genetic variants identified by genome wide association study (GWAS) generally only account for a small proportion of the total heritability for complex disease. One crucial reason is the underutilization of gene-gene joint effects commonly encountered in GWAS, which includes their main effects and co-association. However, gene-gene co-association is often customarily put into the framework of gene-gene interaction vaguely. From the causal graph perspective, we elucidate in detail the concept and rationality of gene-gene co-association as well as its relationship with traditional gene-gene interaction, and propose two Fisher $r$-to-$z$ transformation-based simple statistics to detect it. Three series of simulations further highlight that gene-gene co-association refers to the extent to which the joint effects of two genes differs from the main effects, not only due to the traditional interaction under the nearly independent condition but the correlation between two genes. The proposed statistics are more powerful than logistic regression under various situations, cannot be affected by linkage disequilibrium and can have acceptable false positive rate as long as strictly following the reasonable GWAS data analysis roadmap. Furthermore, an application to gene pathway analysis associated with leprosy confirms in practice that our proposed gene-gene co-association concepts as well as the correspondingly proposed statistics are strongly in line with reality.

Introduction

Since the first successful genome-wide association study (GWAS) for age-related macular degeneration published in 2005 [1], numerous loci associated with complex human disease or traits have been identified. Despite high expectations, the genetic variants identified by GWAS, though providing valuable insights into genetic architecture, generally only account for a small proportion of the total heritability for complex disease [2,3]. Potential explanations may include underestimation of the effects of alleles identified, the existence of gene-gene joint effects, the contribution of rare variation, the possibility that inherited epigenetic factors lead to resemblance between relatives, and possible overestimation of heritability of the complex traits [2,3,4,5]. Moreover, recent technological advances in high-throughput sequencing platforms enables the acquisition of genomic data at unprecedented speed and amounts, in fact, the capacity to generate the data greatly outpaces our ability to analyze and interpret. It is, therefore, quite desirable to further develop more efficient data mining strategy to extract more information from huge GWAS data, rather than put them aside.

Among the data analysis demand, one major issue refers to the joint effects of multiple genes contributing to the interested disease or trait. The joint effect of two genes included their main effects and co-association. We have proposed the concept of gene-gene co-association in previous studies [6,7], which refers to the extent to which the joint effect of two genes on disease (or trait) differs from the main effects of each gene. Traditional methods customarily put gene-gene co-association into the framework of gene-gene interaction. To determine the presence of interactions between two genes, regression-based approaches are still regarded as the most natural first-line approach, though some alternative methods have been developed [8,9,10,11,12,13,14,15]. A product term is usually added to the logistic regression model (LRT) $Logit(P) = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 A \times B$ for the popular case-control design in GWAS, which implies a nearly independence assumption, at least not much correlation, between gene A and gene B for inferring the interaction ($\beta_3$). Nevertheless, one common sense is that the development of most common diseases is attributed to complex gene network system. Genes (or SNPs) are often correlated with each other in the following situations: 1) genes (or SNPs) within pathways or networks contributing to a disease; 2) SNPs with linkage disequilibrium (LD) located in two or more linked genes within one chromosome; 3) SNPs with LD in one gene. Hence the above assumption is rarely satisfied. It will be inevitable to lose efficiency using LRT blindly when high
correlation existed between SNPs. Actually, the genetic pathway or network, even SNPs co-association within one high LD genome region, can be deemed as a graph and studies should be conducted under graphical framework [16]. Specifically, taking 2 SNPs for simplicity, from a causal diagram perspective (Fig. 1 in Methods), suppose the main effects for SNP1 and SNP2 are $\beta_1$ and $\beta_2$ respectively, and the correlation between them is $r$, which is usually far away from zero (e.g., SNPs with LD or two SNPs within one pathway). Then the total effects for SNP1 and SNP2 will be $\beta_1 + \beta_2 + r(\beta_1 + \beta_2) + \beta_3$, and the term ($r(\beta_1 + \beta_2)$) is obviously attributed to the correlation between the two SNPs, which would not be detected efficiently by LRT.

The argument is that, for two genes, the extent to which the joint effects differs from the main effects (i.e. gene-gene co-association), are not only refer to the traditional interaction ($\beta_3$) but the correlation between two genes ($r(\beta_1 + \beta_2)$). While traditional LRT only provide one way to identify the part under the nearly independent condition, with less power for the left attributed to the correlation. To solve this problem, in the context of a standard case-control design, three gene-based statistics, CCU [6], KCCU [7] and PLSPM-based statistic [17], have been developed in our former work based on the difference of correlation of two genes between cases and controls. Actually, similar idea has already been employed to develop new statistics recently [9,11,12]. However, these statistics do not yet jump out from the scope of gene-gene interaction. Particularly, the statistics recently proposed by Rajapakse [11] will be invalid when heavy multicollinearity (strong LD) between SNPs existed as the algorithm needs the computation of the inverse of covariance matrix. For GWAS in practice, at least five aspects for gene-gene co-association should be considered: 1) the theory basis and rationality; 2) efficient and robust statistics to detect it; 3) simple and universally accessible statistics, just like Armitage trend test [18]; 4) the acceptable false positive rate in real world GWAS; 5) the feasibility for computation challenge.

Although various novel statistics for gene-gene interaction or gene-gene co-association have been proposed few of them are successfully used in real world full GWAS data analysis. This is not only due to their elusive statistical model for general geneticist and epidemiologist, but their unrealistic computation burden attributed to some non-parametric methods (e.g., bootstrap or permutation). In this paper, the concept and rationality of gene-gene co-association is elucidated by a simple causal graph. Based on the difference of correlation of two genes between cases and controls, two simple statistics ($U_W$ and $U_F$) for detecting gene-gene co-association are proposed using Fisher $r$-to-$z$ transformation [19,20]. The former was constructed according to the asymptotic distribution theory of the empirical product-moment correlation coefficient for counting variables [21], while the latter is developed by empirically calibrating Fisher $r$-to-$z$ transformation-based simple statistic [19,20]. Various simulation studies are firstly conducted to assess the type I error rate and power, and to clarify the relationship between gene-gene co-association and gene-gene interaction. And then simulations are carried out to evaluate whether the proposed statistics can be affected by strong LD between SNPs. Furthermore, based on the experimental strategy of gain-of-function in functional genomics [22,23], simulations are performed by mimicking real world GWAS roadmap to assess their false positives. Finally, we analyze a GWAS real data from a plausible biologic network underlying susceptibility to leprosy [24], and the computation time is also reported.

**Methods**

**Fisher r-to-z transformation-based statistics**

For GWAS in case-control design, SNP1 and SNP2 denote the two markers. In the framework of causal graph (Fig. 1), no matter whether they are independent or correlated (within same pathway or with LD between them), the total effects for SNP1 and SNP2 can be illustrated by $\beta_1 + \beta_2 + r(\beta_1 + \beta_2) + \beta_3$, the co-association between them can be defined as $r(\beta_1 + \beta_2)$ under independence condition and often be detected by LRT. Let $r_D$ denote the sample correlation coefficient between SNP1 and SNP2 among cases, and $r_C$ between them among controls. We use $r_D - r_C$ to measure the co-association between the two SNPs contributing to the disease. A Fisher $r$-to-$z$ transformation was proposed earlier for testing the difference between two correlation coefficients [19,20]. This transformation was done to $r_D$ and $r_C$, i.e. $z_D = \frac{1}{2}(\log(1+r_D) - \log(1-r_D))$ and $z_C = \frac{1}{2}(\log(1+r_C) - \log(1-r_C))$, furthermore, Wellek and Ziegler [21] have derived the asymptotic distribution of the empirical correlation for counting variables, our proposed statistic $U_W$ for detecting gene-gene co-association was defined as

$$U_W = \frac{z_D - z_C}{\sqrt{\text{var}(z_D) + \text{var}(z_C)}}$$

Where var($z_D$) and var($z_C$) denote the variance estimator from their work. Although $U_W$ is theoretically accurate, the variance formula in the denominator cannot be obtained quickly since it needs the estimation of population frequencies for various combinations of two specific loci, and we may have to compute it one SNP pairs one time, which is inadvisable for enormous real world GWAS data. Therefore, it is critical to develop a further simpler and more efficient statistic to improve the feasibility and practicability. Fisher has provided a well-known statistic for comparing the correlation coefficients from two samples for data from a bivariate normal distribution, which can be used to detect gene-gene co-association by

![Figure 1. A causal graph framework for two SNPs affected the disease, $\beta_1$ and $\beta_2$ represents main effects, $\beta_3$ denotes the traditional interaction, the nondirectional arc between SNP1 and SNP2 (correlation $r$) indicated that the two variables are associated for reasons other than affecting one another.](https://doi.org/10.1371/journal.pone.0070774.g001)
\[ U = \frac{z_d - z_c}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \]

where \(n_1\) and \(n_2\) represent the sample size for case and control respectively. Simulations (data not shown) show this simple statistic can have good performance in gene-gene co-association detection for SNPs without much correlation. Nevertheless, when the correlation between SNP pairs is high, the statistic is normal but with variance far from 1 under the null hypothesis. Therefore, an empirically calibrated statistic \(U_F\) was further proposed as

\[ U_F = \frac{z_d - z_c}{f(z) \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \]

where \(f(z) = I_{(|z|<0.5)} + I_{(|z|>0.5)} \times e^{-0.176 + 0.09 x + 0.315 x^2},\) and \(z\) is the corresponding Fisher transformation for pooled sample correlation between SNP1 and SNP2. \(U_F\) is obtained by empirical simulation based on the fitted functional relationship between correlation coefficients and their variance. Under the additive genetic model, using 88 SNP pairs with various correlation chosen correlation coefficients and their variance. For gene-gene co-association, it might give rise to an illusion that SNP pairs would be detected powerfully as long as they were highly correlated (e.g. high LD). Therefore, our second series of simulation studies were devoted to evaluate whether the proposed statistics can be affected by strong LD between SNPs. Two neighboring genes on chromosome 17q21 (\(ZPB2\) gene and \(GSDMB\) gene) (Fig. S2 for LD plot), which has been confirmed to be associated with asthma [26,27], were chosen to simulate genotypes similarly as the aforementioned design. First, the 7\(^{th}\) SNP (rs47795400, MAF = 46%) on \(GSDMB\) gene was defined as the causal SNP with various odds ratio (1.1 to 1.5), and the co-association between the 5\(^{th}\) SNP (rs2290400) and 9\(^{th}\) SNP (rs7216389) was detected. Second, the 9\(^{th}\) SNP (rs12150079, MAF = 26%) on \(ZPB2\) gene and the 5\(^{th}\) SNP (rs2290400, MAF = 40%) on \(GSDMB\) gene were specified as causal SNPs, and we fixed odds ratio 1.3 for the second causal SNP, while ranging the odds ratio of the first causal SNP from 1.1 to 1.5. The co-association between the 1\(^{th}\) SNP (rs11557466) on \(ZPB2\) gene and the 9\(^{th}\) SNP (rs7216389) on \(GSDMB\) gene was detected. We randomly sample 6000 individuals for each simulation (totally 3000 simulations).

For gene-gene co-association, our third series of simulation studies by mimicking real world GWAS roadmap to assess their false positives. Similar as above, the first 1000 SNPs on chromosome 1 were chosen to simulate genotypes. Four situations were considered: two neighboring SNPs within one gene, two SNPs located in two linked exons within one gene, two SNPs located in two linked genes within one chromosome, and two SNPs located in two genes within one pathway from two different chromosomes. For the first situation, we chose the 9\(^{th}\) SNP on chromosome 11 as the causal SNPs with correlation 0.96, then we embedded the gene into the 1000 SNPs on chromosome 1 generated above to mimic 500000 SNP pairs. Three independent sample with sample size 3000, 3000 and 6000, were generated to mimic the GWAS roadmap and to see whether the three samples reported the same false positive SNP pairs under the following situations: co-association under nearly independent condition between gene A and gene B, co-association caused by correlation between gene A and gene B and co-association caused by both correlation and independent term A x B between gene A and gene B; abbreviate to Type I co-association, Type II co-association and Type III co-association; abbreviated to Type I co-association, Type II co-association and Type III co-association; Type I co-association, Type II co-association and Type III co-association. Three different scenarios were considered (Fig. 1). \(r = 0, \beta_1 \neq 0\) indicated the case that Type I co-association, \(\beta_1 = 0, r \neq 0\) for Type II co-association, and \(\beta_1 \neq 0, r = 0\) for Type III co-association. Three different main effects were set to make our simulations more practical, no marginal effects (\(\beta_1 = \beta_2 = 0\), one marginal effect (\(\beta_1 = 0, \beta_2 = \log(1.3)\)), and two marginal effects (\(\beta_1 = \log(1.5), \beta_2 = \log(1.3)\)). Different \(\beta_1\) and \(r\) are specified to evaluate the type I error (\(\hat{\beta}_1 = r = 0, \hat{\beta}_1 = \beta_2 = \beta_1 = 0, r \neq 0\)) and power. A total of 3000 simulations were repeated for each scenario, and we randomly sample 3000 individuals from the whole 100000 population for each simulation.

Application

Based on the GWAS of leprosy [24], using Ingenuity Pathways Analysis knowledge database (Ingenuity Systems), a plausible biologic network underlying susceptibility to leprosy was created for depicting the functional relationship between the identified five susceptibility genes (together with five other genes). To further confirm the relationship between the genes in the network, we attempt to detect the co-association between SNP pairs within 9
susceptibility genes (2257 SNPs) by the proposed statistics $U_F$ and $U_W$, using the initial GWAS data with 706 cases and 514 controls. These 9 genes locate on different chromosomes and totally contained 2257 SNPs (Table S1). Meanwhile, to compute the simulation time of all three statistics ($U_F$, $U_W$ and LRT), a desktop computer [Intel Core i3-2100 with 3.10 GHz CPU using 4 GB of RAM] was used to do calculations by R 2.14.0.

Results

Simulation studies

Shown in Table 1 were the estimated type I error rates of LRT and the two proposed statistics under $\beta_1 = r = 0$. It revealed that all type I error rates were close to nominal level 0.05 as a function of sample sizes. When $\beta_1 = \beta_2 = \beta_3 = 0, r \neq 0$, though stable for LRT and $U_W$, the type I error was slightly higher for $U_F$ under correlation 0.4 or 0.6 (Table 2). This might be due to $U_F$ essentially do modifications for high correlation and kept the same as Fisher test when correlation was relatively small.

Fig. 2a showed the performance with $r=0$ and $\beta_2 \neq 0$ (Type I co-association) when $\beta_1$ was set to be 0.1, 0.2, 0.3, 0.4. It indicated that the power of all three methods increase monotonically with the interaction effect ($\beta_1$). Both $U_F$ and $U_W$ had almost comparable power with LRT, which was the gold standard in this case. Shown in Fig. 2b was the power for the situation of Type II co-association ($\beta_1 = 0$ and $r \neq 0$), we set $r$ to be 0.2, 0.4, 0.6, 0.8. The power for the two proposed statistics kept relatively high and increased slowly as the correlation coefficient ($r$) increases, while LRT had completely lost the power. Furthermore, the power for the situation with two main effects was higher than that for the situation with only one main effect. Fig. 2c: showed the power for the situation of Type III co-association with fixed $\beta_1 = 0.2$, the power for LRT was relatively lower and decreased as the correlation increases, this might be partly due to the high variance for $\beta_1$ attributed to the increasing correlation. The proposed two statistics had higher power. For the situation with one main effect, though $t(\beta_1 + \beta_2)$ increased, the power for co-association decreased as correlation increases from 0.2 to 0.8, which might be due to the power for detecting the interaction $\beta_1$ decreases as correlation increased. Fig. 2d presented the power for the situation of Type III co-association under fixed correlation 0.4. The power for all statistics increased as the interaction effect increase from 0.1 to 0.4, and the two proposed statistic always had higher power than LRT. All results illustrated that, under medium correlation, the power for $U_F$ seems a little higher than that for $U_W$.

**Table 1.** Type I error for three statistics without correlation and interaction.

| Sample size | LRT $\alpha$ | $U_W$ $\alpha$ | $U_F$ $\alpha$ |
|-------------|--------------|----------------|----------------|
| 1000        | 0.050        | 0.050          | 0.050          |
| 2000        | 0.046        | 0.049          | 0.047          |
| 3000        | 0.052        | 0.055          | 0.054          |
| 4000        | 0.050        | 0.049          | 0.049          |
| 5000        | 0.050        | 0.046          | 0.055          |

*For case with one main effects ($\beta_1 = 0, \beta_2 = \log(1.3)$),

Table 2. Type I error for three statistics without main effects under sample size 3000 and 5000.

| Correlation coefficient | LRT $\alpha$ | $U_W$ $\alpha$ | $U_F$ $\alpha$ |
|-------------------------|--------------|----------------|----------------|
| 0.2                     | 0.050        | 0.050          | 0.040          |
| 0.4                     | 0.046        | 0.049          | 0.045          |
| 0.6                     | 0.052        | 0.055          | 0.048          |
| 0.8                     | 0.050        | 0.049          | 0.058          |

Fig. 3b–3d and Table 3 showed the results of gene-gene co-association analysis for 2257 SNPs within 9 susceptibility genes belonging to the pathway associated with leprosy [24]. For ease of visualization, only SNPs within SNP pairs whose p-value less than $1 \times 10^{-7}$ in at least one of the three methods were presented. For SNP pairs in two different genes, all three statistics had similar results. The co-association between PARK2 and LRRK2 was detected at $z=1 \times 10^{-7}$ by $U_F$ and $U_W$, $1 \times 10^{-6}$ by LRT; the correlation coefficient between the two SNPs of PARK2 (rs904305) and LRRK2 (rs12814017) is 0.13. The co-association between NOD2 and IFNG, IFNG and PARK2 was also detected at $z=1 \times 10^{-4}$ by $U_F$ and $U_W$, $1 \times 10^{-4}$ by LRT; between IFNG and CARD6 at $z=1 \times 10^{-4}$ by $U_F$, $1 \times 10^{-2}$ by $U_W$ and LRT. The marked genes with self-regulation in the network [24] were also detected by both $U_F$ and $U_W$ at $1 \times 10^{-7}$ level, while nothing appeared by LRT due to the stronger LD between SNPs within one gene. This indicated that the results from $U_F$ and $U_W$ strongly agreed with that from the Ingenuity Pathways Analysis knowledge database, while not from LRT in the framework of traditional gene-gene interaction. In addition, all three statistics showed there was some co-association between IFNG and PARK2 ($z=1 \times 10^{-4}$), both $U_F$ and $U_W$ suggest that self-regulation within LRRK2, PARK2, TNSF15, and CARD6 may also exist ($z=1 \times 10^{-4}$), though these were not marked in the network. The computation time for $U_W$ takes nearly 25 hours, and about 28 hours for LRT, while only 3 minutes for statistic $U_F$ using the
Discussion

The rationality and significance of gene-gene co-association

From the causal graph perspective [16], we elucidated the concept and rationality of gene-gene co-association, and clarify its relationship with the traditional gene-gene interaction. Simulation studies further confirm our viewpoint. Fig. 2a shows that the co-association is almost the same as the interaction in the situation co-association with standalone interaction. Fig. 2b demonstrates that the co-association still exists in the situation co-association without interaction. Fig. 2c and 2d illustrate the situation of Type III co-association, indicating that it will be lost some power when replacing co-association with interaction. Actually, these relationships have also been supported by a gene-gene interaction study [11], though it does not yet jump out from the scope of traditional gene-gene interaction. The statistic they proposed based on the difference of the covariance matrix between cases and controls, showing much power than the LRT, indeed measure the co-association between two genes essentially, rather than their interaction. Specifically, simulation indicated that their proposed statistic showed no power when the two genes only have marginal effect on the disease (case 1 in their work). This is actually the situation of Type II co-association (Fig. 2b in our simulation), the reason why no power emerge is that the two selected gene region (EXT2 and LRRC4CX2) in their simulation are far away from each other and can be considered to be independent. In summary, gene-gene co-association refers to the extent to which the joint effects of two genes differs from the main effects, not only due to the traditional interaction under the nearly independent condition but the correlation between two genes, while the part attributed to the correlation has usually been neglected in traditional interaction model using regression method. Genetically, most diseases are caused by multiple genes acting together through pathways or network that can lead to a common final disease or trait. In practice, when constructing a priori topological structure for establishing genetic networks that contribute to diseases of interest,
we often need to test whether significant relationships between any two nodes in such networks exist. It seems more reasonable to solve this by detection for gene-gene co-association rather than traditional interaction.

Fisher $r$-to-$z$ transformation-based simple statistics for gene-gene co-association

Wellek and Ziegler [21] derived the asymptotic distribution of the empirical product-moment correlation coefficient for counting variables. One statistic we here proposed, based on the strict theory from their work, is $U_W$. Alternatively, from the feasibility and practicability perspective, we empirically calibrate the traditional Fisher $r$-to-$z$ transformation-based simple statistic $U$ and proposed $U_F$, which is not only prone to easy understanding and universally accessible to everyone, but compute fast in practice. Simulation showed that the proposed two statistics are stable, though the type I error of $U_F$ slightly deviates from the nominal level due to the empirical approximation (Table 2). There seems a tradeoff between the accuracy and the computation burden, the theoretical statistic $U_W$ was accurate but with high computation burden, while $U_F$ could reduce the computation burden substantially but might lose some accuracy. Both $U_W$ and $U_F$ have comparable power with LRT under Type I co-association, while more powerful than LRT under Type II co-association and Type III co-association no matter what the correlation between the SNPs is (Fig. 2). This indicates that the two proposed statistics have good performance for detecting gene-gene co-association. Intuitively, it might give rise to an illusion that the co-association between SNP pairs would be detected powerfully as long as they were highly correlated (e.g. high LD). However, our results illustrate that both $U_W$ and $U_F$ cannot be affected by LD, and co-association indeed represents nothing but the effect contributing to the disease (Fig. 3a). It is important to guard against possible heterogeneity caused by some other
covariates (e.g., age, gender, smoking). One possible solution for this is Mantel-Haenszel method, which may suffer small sample size problem when the number of covariates is quite large. Another possible way is to calculate the partial correlation size problem when the number of covariates is quite large.

The advantages of statistics $U_W$ and $U_F$ in real world GWAS data analysis

For real world GWAS data analysis, one way to search for co-association (or interaction) is arguably by exhaustive search, which consider all possible pairs of loci and perform the desired co-association test for each pair (e.g., about 50,000 SNP pairs for 1000 loci). Therefore, whether one statistic can be used in real world GWAS data analysis depends on two key aspects at least, the acceptable false positive rate and computation burden. Simulation following up the GWAS data analysis roadmap indicates that the false positive rate of the proposed two statistics $(U_W$ and $U_F$) together with LRT are all at about $1 \times 10^{-7}$ order of magnitude. Also, it indicates that the false positives can be acceptable and control well as long as researchers strictly followed the reasonable GWAS data analysis roadmap. As an example, all three statistics were used to analyze 2257 SNPs (254596 SNP pairs) within 9 susceptibility genes belonging to the pathway associated with leprosy using a desktop computer (Intel Core i3-2100 with 3.10 GHz CPU using 4 GB of RAM), the computation time for leprosy using a desktop computer (Intel Core i3-2100 with 3.10 GHz CPU using 4 GB of RAM), the computation time for $U_F$ was nearly 25 hours, and about 28 hours for LRT, while only 3 minutes for statistic $U_F$, which may be currently the most realistic and feasible statistic.

Application to gene pathway analysis associated with leprosy

The GWAS for leprosy showed that variants of genes in the NOD2-mediated signaling pathway (which regulates the innate immune response) are associated with susceptibility to infection with *M. leprae*, and a further plausible biologic network was created for highlighting the functional relationship between the suscepti-

### Table 3. Gene-gene co-association for SNP pairs ($p < 1 \times 10^{-7}$) within 9 susceptibility genes belonging to the pathway associated with leprosy.

| SNP pairs Gene | SNP pairs Gene |
|----------------|----------------|
| rs16869977-rs10512739 CARD6 | rs39503-rs447618 RIPK2 |
| rs11744119-rs2271709 CARD6 | rs39503-rs411279 RIPK2 |
| rs4245977-rs10437328 CARD6 | rs447618-rs411279 RIPK2 |
| rs1815510-rs31772235 HLA-DRB1 | rs6470668-rs7812579 RIPK2 |
| rs1822520-rs1404269 PARK2 | rs6470668-rs7459577 RIPK2 |
| rs1822520-rs9365252 PARK2 | rs7855735-rs1125441 TNFSF15 |
| rs10945765-rs2281403 PARK2 | rs10880160-rs1390995 LRRK2 |
| rs11962721-rs10945770 PARK2 | rs776421-rs776207 IFNG |
| rs1789995-rs1789993 PARK2 | rs775450-rs775448 IFNG |
| rs1186144-rs8043960 NOD2 | rs7186144-rs8043960 NOD2 |

doi:10.1371/journal.pone.0070774.t003

Limitations

One has to realize that the implications of gene-gene interaction are scale-dependent, we here just illustrate gene-gene co-association by comparison with multiplicative interactions in LRT, where the final term expresses a departure from a simple additive model on the logit scale. As one reviewer suggested, we have also assessed the performance for rare variation (MAF<$0.05$) and found that the type I error is unstable, which suggested that the proposed methods was invalid for rare variation. In addition, although the proposed empirical statistic $U_F$ has nearly same performance with the theoretical statistic $U_W$, its type I error deviates slightly from the given nominal level (Table 2) after all, this may elevate a little false positive rate. We want to emphasize that different empirical fitness methods may generate different function $f$ in the denominator of $U_F$, the basic rule for $U_F$, we think, is to improve the computation efficiency and feasibility, meanwhile keep the performance nearly the same as $U_W$.

### Supporting Information

**Figure S1** The scatter plot for the correlation with corresponding empirical variance. (TIF)

**Figure S2** The LD plot for two neighbored genes on chromosome 17q21, with the first 8 SNPs belonging to ZPPB2 gene and the left belonging to GSDMB gene. (TIF)

**Table S1** The location and SNP number for 9 susceptibility genes belonging to the pathway associated with leprosy. (DOC)

### Acknowledgments

We would like to thank anonymous reviewers and the academic editor for providing us with constructive comments and suggestions. We also wish to acknowledge our colleagues for their invaluable work.

### Author Contributions

Conceived and designed the experiments: ZY FX XZ FZ. Analyzed the data: ZY FX FL JZ. Wrote the paper: ZY FX XZ.

### References

1. Haines JL, Hauser MA, Schumack S, Scott WK, Olson LM, et al. (2005) Complement factor H variant increases the risk of age-related macular degeneration. Science 308: 419–421.

2. Gibson G (2010) Hints of hidden heritability in GWAS. Nat Genet 42: 558–560.

3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.

4. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics 11: 446–450.
5. Gibson G (2012) Rare and common variants: twenty arguments. Nature Reviews Genetics 13: 135–145.

6. Peng Q, Zhao J, Xue F (2009) A gene-based method for detecting gene–gene co-association in a case–control association study. European Journal of Human Genetics 18: 352–357.

7. Yuan Z, Gao Q, He Y, Zhang X, Li F, et al. (2012) Detection for gene–gene co-association via kernel canonical correlation analysis. BMC genetics 13: 83.

8. Cordell HJ (2009) Detecting gene–gene interactions that underlie human diseases. Nature Reviews Genetics 10: 392–404.

9. Kam-Thong T, Czamara D, Tsuda K, Borgwardt K, Lewis CM, et al. (2010) EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. European Journal of Human Genetics 19: 465–471.

10. Kang G, Yue W, Zhang J, Cui Y, Zuo Y, et al. (2008) An entropy-based approach for testing genetic epistasis underlying complex diseases. Journal of theoretical biology 250: 362–374.

11. Rajapakse I, Perlman MD, Martin PJ, Hansen JA, Kooperberg C (2012) Multivariate Detection of Gene-Gene Interactions. Genetic epidemiology 36: 622–630.

12. Ueki M, Cordell HJ (2012) Improved statistics for genome-wide interaction analysis. PLoS Genetics 8: e1002629.

13. Wai X, Yang C, Yang Q, Xue H, Fan X, et al. (2010) BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. American journal of human genetics 87: 325.

14. Wu X, Dong H, Liao L, Zhu Y, Peng G, et al. (2010) A novel statistic for genome-wide interaction analysis. PLoS genetics 6: e1001131.

15. Zhao J, Jin L, Xiong M (2006) Test for interaction between two unlinked loci. American journal of human genetics 79: 831.

16. Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. Epidemiology 10: 37–48.

17. Zhang X, Yang X, Yuan Z, Liu Y, Li F, et al. (2013) A PLSPM-Based Test Statistic for Detecting Gene-Gene Co-Association in Genome-Wide Association Study with Case-Control Design. PloS one 8: e62129.

18. Armitage P (1955) Tests for linear trends in proportions and frequencies. Biometrics 11: 375–386.

19. Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10: 507–521.

20. Fisher RA (1921) On the “Probable Error” of a Coefficient of Correlation Deduced from a Small Sample. Metron 1: 3–32.

21. Wellek S, Ziegler A (2008) A genotype-based approach to assessing the association between single nucleotide polymorphisms. Human Heredity 67: 120–139.

22. Gaiano N, Kohtz JD, Turnbull DH, Fuhell G (1999) A method for rapid gain-of-function studies in the mouse embryonic nervous system. Nature neuroscience 2: 812–819.

23. Winkhofer KF, Tatelle J, Haass C (2008) The two faces of protein misfolding: gain-and loss-of-function in neurodegenerative diseases. The EMBO journal 27: 336–349.

24. Zhang F, Huang W, Chen S, Sun L, Liu H, et al. (2009) Genomewide association study of leprosy. New England Journal of Medicine 361: 2609–2618.

25. Su Z, Marchini J, Donnelly P (2011) HAPGEN2: simulation of multiple disease SNPs. Bioinformatics 27: 2304–2305.

26. Beuzigio E, Corda E, Aschard H, Dizier M, Boland A, et al. (2008) Effect of 1q21 variants and smoking exposure in early-onset asthma. New England Journal of Medicine 359: 1983–1994.

27. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448: 470–473.