

# Association Testing Strategy for Data from Dense Marker Panels

Donghyung Lee\*, Silviu-Alin Bacanu

Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, United States of America

## Abstract

Genome wide association studies have been usually analyzed in a univariate manner. The commonly used univariate tests have one degree of freedom and assume an additive mode of inheritance. The experiment-wise significance of these univariate statistics is obtained by adjusting for multiple testing. Next generation sequencing studies, which assay 10-20 million variants, are beginning to come online. For these studies, the strategy of additive univariate testing and multiple testing adjustment is likely to result in a loss of power due to (1) the substantial multiple testing burden and (2) the possibility of a non-additive causal mode of inheritance. To reduce the power loss we propose: a new method (1) to summarize in a single statistic the strength of the association signals coming from all not-very-rare variants in a linkage disequilibrium block and (2) to incorporate, in any linkage disequilibrium block statistic, the strength of the association signals under multiple modes of inheritance. The proposed linkage disequilibrium block test consists of the sum of squares of nominally significant univariate statistics. We compare the performance of this method to the performance of existing linkage disequilibrium block/gene-based methods. Simulations show that (1) extending methods to combine testing for multiple modes of inheritance leads to substantial power gains, especially for a recessive mode of inheritance, and (2) the proposed method has a good overall performance. Based on simulation results, we provide practical advice on choosing suitable methods for applied analyses.

**Citation:** Lee D, Bacanu S-A (2013) Association Testing Strategy for Data from Dense Marker Panels. PLoS ONE 8(11): e80540. doi:10.1371/journal.pone.0080540

**Editor:** Dan E. Arking, Johns Hopkins University, United States of America

**Received:** May 27, 2013; **Accepted:** October 14, 2013; **Published:** November 12, 2013

**Copyright:** © 2013 Lee, Bacanu. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** Funding provided by a VCU (Virginia Commonwealth University) start-up fund for SAB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

\*E-mail: dlee4@vcu.edu

## Introduction

Genome-wide association studies (GWASs) have been broadly used to test for association between genetic variants and various phenotypes. These studies have been quite successful in identifying numerous single nucleotide polymorphisms (SNPs) associated with a variety of human traits and diseases [1]. So far, GWASs have been commonly analyzed univariately using one degree of freedom (df) tests assuming an additive mode of inheritance [2–4]. The experiment-wise significance of the univariate statistics was assessed using a Bonferroni adjustment [5,6] or a permutation procedure [7–9]. While this approach was reasonably successful for GWAS, the field is moving away from this paradigm towards whole genome sequencing. When compared to GWAS, variant panels for sequencing studies (1) are substantially denser and (2) have different patterns of linkage disequilibrium (LD). Consequently, for these studies it is not clear if the most desirable approach is still a univariate testing

for an additive mode of inheritance followed by the adjustment of the statistics for the large number of tests.

Intuitively, a test for association between phenotype and genotype achieves optimal power when the assumed mode of inheritance matches the underlying one. However, the underlying mode of inheritance is usually unknown in practice and an incorrect choice for it may cause a substantial power loss. Various strategies to minimize the effect of the possible model misspecification have been studied and developed [10–17]. Among these, one simple strategy is to test other modes of inheritance, e.g. dominant and recessive, in addition to the commonly used additive mode of inheritance [16,17]. This approach involves testing for three different modes of inheritance, i.e. additive (A), dominant (D) and recessive (R), and adjusting the lowest p-values for multiple testing. For brevity, the combination of testing for the three modes of inheritance will be henceforth denoted as ADR. Because the ADR paradigm is not in widespread use yet, it is of interest to estimate the performance improvement when applied to

methods which were initially developed assuming an additive mode of inheritance.

To control the rate of false positives in GWAS analyses, the statistical significance of univariate p-values is adjusted for around a million univariate tests. With the advent of next generation sequencing, for univariate analyses, the number of tests will increase dramatically when compared to GWAS. Summary data from 1000 Genomes Project suggests that sequencing studies consisting of subjects from any main ethnic regions, i.e. Europe, East Asia, South Asia, Africa and the Americas, will result in the typing of at least 5 million SNPs having a minor allele frequency (MAF) above 5% [18]. For larger sequencing studies, if we assume that all SNPs with  $MAF > 0.5\%$  are analyzed individually, the number of tests increases to more than 10 million for Caucasian and Asian cohorts and approaches 20 million for African cohorts. Consequently, univariate approaches entail an even more substantial multiple testing adjustment burden for sequencing studies. It is conceivable that using multilocus approaches, e.g. by summarizing the association of multiple SNPs simultaneously, opens the possibility of decreasing the multiple testing burden and, thus, increasing the power of detection for association signals.

To minimize the power penalty due to multiple testing adjustment, researchers proposed to analyze simultaneously all SNPs in a biological functional block of interest, e.g. a gene [19]. However, this approach might yield low power due to the large number of df. Subsequently, researchers proposed methods to decrease the number of df, and thus increase power, by (1) summarizing the LD information mainly from low-frequency SNP variation in an LD block [20], (2) using data-adaptive sum of squared scores (aSUM) [21], (3) summarizing the LD information by the first few principal components (PC) [22,23], (4) combining the Simes p-value of all univariate p-values in a gene with the p-value associated with the first few principal components of tests in the gene (S-PC) [24] and (5) combining individual SNP-based variance component score statistics (SKAT) [25].

Recently, two fast non-regression based multilocus methods were proposed for gene-based analysis. The first method, denoted as VEGAS, summarizes the association signals in a gene using the sum of squares (SS) or the minimum p-value (minP) of univariate statistics in the gene [26]. Instead of performing permutations, VEGAS simulates multivariate normal variables for a rapid assessment of the asymptotic null distribution of summary statistics. The second method is an improved Simes procedure for association studies, denoted as Gene-based Association Test using Extended Simes (GATES) [27]. GATES is an extension to a technique proposed initially by Cheverud [28] to determine the effective number of independent tests in a region of interest. (For more background regarding this type of methods, see 28–32.) Because VEGAS-SS (V-SS), VEGAS-minP (V-minP) and GATES do not use permutations, they are faster than permutation based multilocus methods.

While the gene is commonly considered as the biological functional unit, it might not be the best unit for statistical analysis. First, since not all parts of a gene are equally

important functionally, it would be more powerful to analyze separately SNPs in regions with important functions such as gene promoter regions, which are known to be involved in initiating and regulating the transcription process [33,34]. Second, the association signal from a causal SNP, e.g. from the promoter region, is diffused only among SNPs in the same LD block [20]. Thus, using as the unit of analysis a gene containing multiple LD blocks might increase the noise and reduce overall statistical power to detect an association signal. An alternative unit of statistical analysis might be a LD block (e.g. SNPs with  $D'$  near 1).

Unlike GWAS, which are based on a tag SNP approach, whole genome sequencing studies assay most genetic variants in the genome. Consequently, the typed variation in these studies will likely form a large number of LD blocks, each block having a large number of SNPs. Given the number and size of these blocks, it is conceivable that approaches analyzing simultaneously all SNPs in a LD block might be of great importance for a successful investigation of sequencing studies. Although the partition of the genome into LD block requires a substantial computing time, once they are computed for the main ethnic groups, these blocks can be reused for future analyses. While the LD block analysis approach was investigated before [17], these findings need to be updated for the present-day technological and methodological environment, i.e. increased variant density and the new developments in gene based methods.

In this paper, we attempt to develop/identify the most powerful methods to detect association signals in a LD block. To achieve this goal, we (1) propose a simple statistic consisting of the sum of the squares of significant univariate tests in a LD block, (2) use two simulation experiments to assess the performance of the proposed method and competing methods a) with or b) without an ADR extension and (3) provide practical recommendations based on simulation results.

## Materials and Methods

### Methods

Assume that we are interested in assessing the association between a binary phenotype,  $Y$ , and  $m$  SNPs in a LD block using a case-control cohort consisting of  $n$  cases and  $n$  controls. For the  $i^{\text{th}}$  subject, let  $Y_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ subject is a case} \\ 0 & \text{if } i^{\text{th}} \text{ subject is a control} \end{cases}$ ,  $i = 1, \dots, 2n$ , be the phenotype and  $\mathbf{G}_{A,i} = (G_{A,i1}, \dots, G_{A,im})$  be the additively coded genotypes for the  $m$  SNPs in the LD block, i.e. the number of reference alleles. To test for a dominant mode of inheritance, we denote the indicators for the heterozygote and the reference allele homozygote as  $\mathbf{G}_{D,i} = (G_{D,i1}, \dots, G_{D,im})$ . To test for a recessive mode of inheritance, let the indicators of reference allele homozygote be  $\mathbf{G}_{R,i} = (G_{R,i1}, \dots, G_{R,im})$ . For each mode of inheritance, we can obtain the vector of normally distributed test statistics by regressing the phenotype on the relevant genotype vector. Assume that the vectors of test statistics corresponding to the additive, dominant and recessive modes of inheritance are  $\mathbf{Z}_A = (Z_{A,1}, \dots, Z_{A,m})$ ,  $\mathbf{Z}_D = (Z_{D,1}, \dots, Z_{D,m})$  and

$Z_R=(Z_{R,1},\dots,Z_{R,m})$ , respectively. Based on these statistics, we can compute p-value vectors for the three modes of inheritance as  $p_A=(p_{A,1},\dots,p_{A,m})$ ,  $p_D=(p_{D,1},\dots,p_{D,m})$  and  $p_R=(p_{R,1},\dots,p_{R,m})$ .

Certain gene based approaches, such as V-SS, sum the squares of all univariate statistics in a LD block. However, this approach might lose power by including SNPs with weak association signals, which only add noise to the test statistic. V-minP also might suffer some statistical power loss by using only the most significant statistic. To avoid such a loss of power we propose a new statistical test consisting of the sum of squared statistics exceeding a threshold, i.e.  $\sum_{j=1}^m Z_{A,j}^2 I(Z_{A,j}^2 \geq t)$ , where  $t>0$  is a reasonably high threshold. We denote this test statistic as the sum of square above a threshold (SS-T). The statistical significance of SS-T can be assessed in a manner similar to V-SS and V-minP, i.e. via multivariate normal simulations based on the LD pattern in the block.

We compare the performance of the proposed test (SS-T), to the performance of various association tests developed mainly for an additive mode of inheritance and, where possible, their ADR extensions. As association tests, we include in our simulation studies Bonferroni, Simes [35], GATES [27], V-SS/V-minP [26], PC [22,23], S-PC [24], SKAT [25] and aSUM [21]. In our comparisons we also include the ADR extensions for SS-T (ADR SS-T), Bonferroni (ADR Bonferroni), Simes (ADR Simes), GATES (ADR GATES), V-SS (ADR V-SS), V-minP (ADR V-minP), PC (ADR PC) and S-PC (ADR S-PC). ADR extensions for SKAT and aSUM are not attempted because the additive coding for genotype is implicit in the software implementing these methods.

Bonferroni procedure summarizes the  $m$  univariate p-values in a LD-block p-value defined as  $\min_j(m p_{A,j})$ ,  $j=1,\dots,m$ . Simes method alleviates the conservativeness of Bonferroni adjustment by using  $\min_j\left(\frac{m p_{A,j}}{j}\right)$ ,  $j=1,\dots,m$ , as the block p-value [35]. GATES method enhances the power of Simes by using an effective number of tests approach. The p-value of GATES method is obtained as  $\min_j\left(\frac{m_{eff} p_{A,j}}{m_{eff(j)}}\right)$ ,  $j=1,\dots,m$ , where  $m_{eff}$  is the effective number of tests of all p-values ( $p_{A,1},\dots,p_{A,m}$ ) and  $m_{eff(j)}$  is the effective number of tests computed from the  $j$  smallest p-values ( $p_{A,(1)},\dots,p_{A,(j)}$ ) [27]. V-SS and V-minP compute the sum of squares,  $\sum_{j=1}^m Z_{A,j}^2$ , and minimum p-value (which is equivalent to  $\max_j(Z_{A,j}^2)$ ) of the univariate statistics and assess their significance by simulating their null distributions based on the multivariate normal distribution [26]. The test statistic for the PC method,  $Q_k$ , is defined as the sum of squares of the first  $k$  PC statistics, denoted as  $U_1,\dots,U_k$ , of the genotype correlation matrix. The  $j^{th}$  PC statistic is defined as  $U_j=v_j \cdot Z_A / \sqrt{\lambda_j}$ , where  $v_j$  and  $\lambda_j$  are the  $j^{th}$  eigenvector and eigenvalue of the correlation matrix of genotype data  $G_A$ . [22,23]. Under the null hypothesis, of no association between genotype and phenotype,  $Q_k$  is distributed as a  $\chi^2$  with  $k$  df. In our comparison studies, we consider the first three PC statistics, i.e.  $k=3$ . The p-value for S-PC is obtained by performing a Simes correction on the p-values generated from (i) a Simes procedure applied on  $p_A$  and (ii) the above PC method [24].

**Table 1.** Number of cases/controls ( $n$ ), relative risk of heterozygote to non-risk allele homozygote ( $R_1$ ) and relative risk of risk allele homozygote to non-risk allele homozygote ( $R_2$ ) used at each simulation setting under the single causal variant scenario of Experiment II.

Number of causal variants ( $k$ )	Genetic Model		
	Additive	Dominant	Recessive
1	1,000 (1.3, 1.6)	1,000 (1.5, 1.5)	1,000 (1, 3)

Within each cell, the settings are presented as  $n (R_1, R_2)$ .  
doi: 10.1371/journal.pone.0080540.t001

**Table 2.** Number of cases/controls ( $n$ ) and effect size ( $\delta$ ) of any causal allele used at each simulation setting under the two and five non-interacting causal variant scenario of Experiment II.

Number of causal variants ( $k$ )	Genetic Model		
	Additive	Dominant	Recessive
2	1,000 (0.005)	1,000 (0.008)	1,000 (0.03)
5	1,000 (0.002)	1,000 (0.003)	1,000 (0.01)

Within each cell, the settings are presented as  $n (\delta)$ .  
doi: 10.1371/journal.pone.0080540.t002

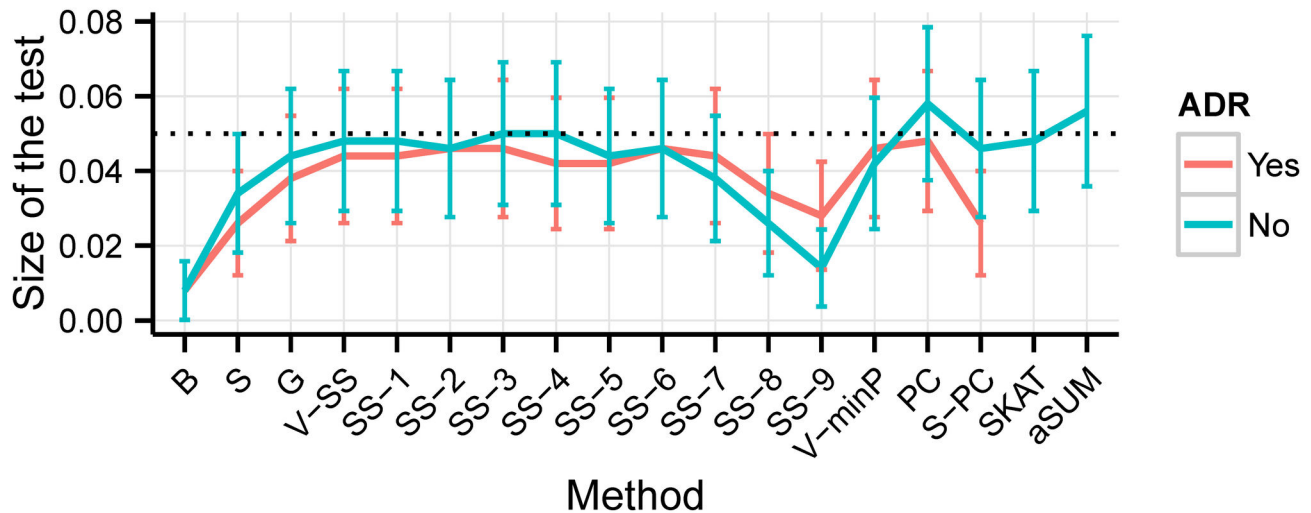
The ADR extension for methods in the previous paragraph can be achieved similarly by substituting: (i)  $3m$  for  $m$  as the number of tests in the LD block, (ii) ADR p-values ( $p_{ADR}=(p_A,p_D,p_R)$ ) for p-values assuming an additive mode of inheritance ( $p_A$ ) and (iii) ADR univariate statistics ( $Z_{ADR}=(Z_A,Z_D,Z_R)$ ) for univariate statistics assuming an additive mode of inheritance ( $Z_A$ ). While the three tests (A, D and R) are not independent, both within and between SNPs, given that the power loss is small [17], we conservatively use  $3m$  as the number of tests for the ADR extension for Bonferroni and related methods.

We implemented in R all methods tested in this paper, with the exception of SKAT and aSUM. SKAT statistics were obtained using SKAT 0.63 R package with a linear kernel and the default options. For aSUM, we used its implementation from Assotester 0.1-1 R package (<http://www.gastonsanchez.com/assotester>) with the default options.

**Simulations**

We employ two extensive simulation experiments to generate genotype-phenotype data sets that we subsequently use to assess the performance of relevant methods and their ADR-extensions. The first simulation experiment generates artificial LD-patterns (see Tables S1-S3) and the second experiment is based on LD-patterns from a real data set (see Tables 1 and 2).

To efficiently simulate a large number of correlated SNPs in a LD block, the first simulation experiment (Experiment I) employs the polychoric correlation (PCC) as a measure of LD [17,36]. The advantage of PCC over other LD measures is that it is directly applicable to simulating a large number of markers



**Figure 1. Empirical size of the test under Experiment II as a function of the method type and the ADR adjustment status.** The nominal type I error rate is  $\alpha=0.05$ . The bars represent the 95% confidence interval for the size of test. Abbreviations for methods are as follows: B - Bonferroni, S - Simes, G - GATES, V-SS - VEGAS-SS, SS-x - SS-T with  $x=1, \dots, 9$ , V-minP - VEGAS-minP, PC - principal component method, S-PC - Simes adjustment of Simes and PC methods, SKAT - sequence kernel association test, aSUM - data-adaptive sum test.

doi: 10.1371/journal.pone.0080540.g001

in LD regardless of their reference allele frequencies (RAFs) [17]. The simulation of a study cohort is achieved using a four-step process: i) simulate latent multivariate normal variables, ii) discretize the latent variables to obtain genotypes, iii) simulate the phenotype, i.e. case or control, for the simulated genotype vector and iv) accept cases and controls until achieving the required sample size. In this experiment, we investigate three settings for the number of causal variants ( $k$ ): i) null hypothesis, i.e. no causal variant ( $k=0$ ), ii) a single causal variant ( $k=1$ ) and iii) two non-interacting causal variants ( $k=2$ ). We present the simulation process in more detail under Supporting Information (see Methods S1).

The second simulation experiment (Experiment II) is based on 200 randomly chosen 250Kb genetic regions from UK10K (www.uk10k.org) reference data set. LD blocks are inferred by performing a hierarchical clustering analysis with an average link on SNP genotypes in each selected region. To match one of the settings from the first experiment, the LD blocks are defined using a PCC<sup>2</sup> threshold of 0.64. The causal LD block and the causal variant(s) within it are chosen randomly and, thus, their RAFs are not fixed. When compared to Experiment I, to increase robustness of the findings, we add to this simulation experiment a five non-interacting causal variant scenario ( $k=5$ ). Otherwise, the simulations for this experiment follow the conceptual flow of the first experiment.

To examine the effect of the causal model on the power of various methods, we simulated case-control cohorts under three underlying modes of inheritance, i.e. additive, dominant and recessive (see Tables S2 and S3 for Experiment I and Tables 1 and 2 for Experiment II). Under  $k$  causal variant scenario ( $k>1$ ), the underlying genetic modes of inheritance of all  $k$  causal SNPs are assumed to be identical (for more details

see Methods S1). Each case-control sample consists of 1,000 cases and 1,000 controls for most settings (see Tables S2 and S3 for Experiment I and Tables 1 and 2 for Experiment II). To attain the 50% average power target, for Experiment I the genotype penetrances and the sample sizes vary with the mode of inheritance and the causal allele frequency (see Tables S2 and S3).

For both experiments, we assume a binary trait of prevalence  $K=0.05$  and assess the empirical size of the test and power for tested methods at a type I error of 0.05. We obtain the size and power of each method by simulating 500 replicates at each setting. For Monte-Carlo simulation/permutation based methods, such as V-SS, V-minP, SS-T and aSUM, we performed 500 and 1,000 simulations under Experiment I and II, respectively.

## Results

For Experiment I, under the null hypothesis (of no association between SNP genotypes and trait) some tests (V-SS, V-minP, SS-T and aSUM) using Monte-Carlo simulation/permutation to assess the significance of the test statistic seem to have slightly (albeit not-statistically significant) inflated size of the test (Figure S1). However, when we increase to 1000 the number of replications for Experiment II, all tested methods control the size of the test at or below the nominal type I error (Figures 1). As a general rule, ADR extensions tend to make most methods more conservative. Simpler adjustments for multiple comparisons, e.g. Bonferroni and Simes, their ADR extensions and SS-T are the most conservative.

To find a reasonable threshold value for SS-T we use the results of Experiment I and II, which are summarized in Figures

S2-S7 and Figure 2, respectively. Under the first experiment, as the threshold value increases from 1 to 9, power tends to rise. This behavior is more apparent especially when 1) the mode of inheritance is recessive, 2) the tests are ADR-adjusted, 3) the size of the LD block is small and 4) the polychoric correlation between the genotypes in the LD block is high. Under Experiment II, where we use realistic LD blocks from UK10K data set, power decreases with an increase in threshold values. Though under the dominant and recessive modes of inheritance power tends to slightly increase as the threshold value increases from 4 to 6, the improvement in power is marginal compared to that observed under Experiment I. On the basis of the results from both comparisons, we deem SS-6 to be close to optimal power-wise and, henceforth we present only its performance.

Under Experiment I conditions, the single causal variant simulations show differences in power between the ADR adjusted methods and the non-ADR-adjusted methods (Figures S8-S10). When the underlying modes of inheritance are additive or dominant, the ADR-adjustment causes a small power loss averaging 4.9% (see "Additive" and "Dominant" panels in the above mentioned figures). However, for a recessive mode of inheritance, the adjustment considerably improves the power to detect the association signal for most settings. The average ADR adjustment power gain under a recessive mode of inheritance is around 34.3%, ranging from 6.9% for PC to 51.8% for V-minP. Under a dual causal variant scenario, the differences in power between methods and their ADR extensions are similar to the single causal variant scenario (Figures S11-S13). ADR adjustment results in a 5.3% decrease in average power for non-recessive models and a 30% increase in average power for a recessive mode of inheritance. The average power gain of the methods under a recessive mode of inheritance ranges from 8.9% for V-SS to 40.5% for V-minP. The power gain for both (single and dual) causal variant scenarios under a recessive mode of inheritance appears to increase with i) an increase in the polychoric correlation between SNPs in the LD block, ii) a decrease in the size of the LD block and iii) a decrease in causal allele frequency (CAF).

Besides the difference in power between methods and their ADR extensions, it is of interest to establish which methods perform better under varying scenarios. For the first experiment, aSUM tends to have the highest power under additive/dominant modes of inheritance and it outperforms, sometimes considerably, the next best performing group (V-minP, GATES, SS-6 and their ADR extensions) (Figures S8-S13). Under additive/dominant modes of inheritance, as the CAF increases the difference in power between methods gradually lessens while the rank of each method tends to be maintained. For a recessive mode of inheritance, ADR V-minP and ADR GATES perform best overall and are followed by ADR Bonferroni, ADR Simes, ADR S-PC and ADR SS-6. These methods substantially outperform all non-ADR-adjusted methods, ADR V-SS and ADR PC. As CAF increases, under a recessive mode of inheritance, the performance of ADR SS-6 approaches that of the best performers (ADR V-minP and ADR GATES). For Experiment II, under the additive mode of

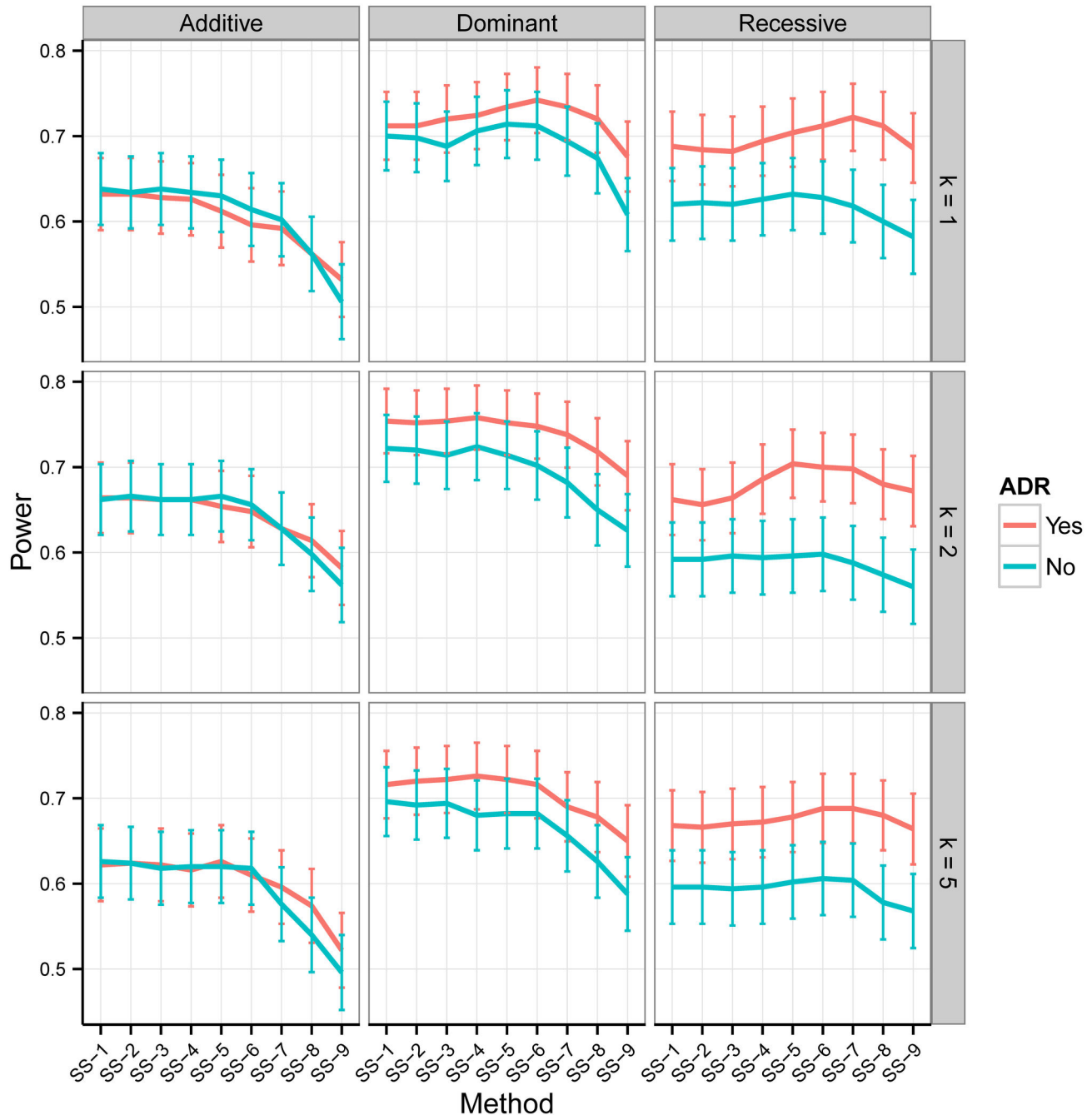
inheritance SKAT, V-SS, V-minP and SS-6 performed best (Figure 3). Under the non-additive mode of inheritance, ADR SS-6, ADR V-SS and ADR V-minP tend to have highest power. While, as we observed from Experiment I, under the additive model ADR adjustment results in slight power loss, under the dominant model it leads to slight power gain. Under the recessive model, ADR adjustment improves power significantly. When compared to Experiment I, SKAT performs much better under Experiment II conditions. On the other hand, aSUM performs rather poorly when compared to its excellent performance observed in the first simulation experiment. Even more, under a recessive mode of inheritance, aSUM had the lowest power in the second experiment.

## Discussion

When compared to GWAS, whole genome sequencing studies assay tens of millions of genetic variants in human genome. Due to their large number, a univariate analysis of these variants involves a substantial multiple testing burden. To avoid such a burden, we propose a new method to summarize in a single statistic the association between phenotype and the genotypes of not-very-rare SNPs, i.e. those which can be reasonably analyzed individually, in a LD block. We use two simulation experiments to compare the performances of i) the proposed method, ii) competing methods, e.g. methods used for gene-based analyses and iii) the extensions of the previously mentioned methods which combine information from multiple modes of inheritance. The results of these simulations are helpful in identifying methods delivering close to optimal power when used to analyze data coming from sequencing studies.

One important conclusion of this paper is that, even for denser marker panels, when the true mode of inheritance is unknown, combining additive, dominant and recessive modes of inheritance together (i.e. ADR adjustment) is a suitable strategy to minimize the power loss caused by the model misspecification. The power gain of ADR adjustments over their non-ADR adjusted counterparts is due to the considerable improvement in the overall performance of the methods when the underlying mode of inheritance is recessive. The power gain of ADR extensions increases with an increase in LD between SNPs and a decrease in causal allele frequency. This behavior is consistent with findings for GWAS panels [16,17].

Under the theoretical experiment (Experiment I) with additive and dominant modes of inheritance, for the rarer causal allele frequencies range we assumed, aSUM performs best across all configurations closely followed by V-minP, GATES, SS-6 and their ADR extensions. Under a recessive mode of inheritance, the most powerful methods are ADR V-minP and ADR GATES, which are closely followed by ADR Bonferroni, ADR Simes, ADR S-PC and ADR SS-6. Probably because it assumes additively coded genotypes, aSUM yields low power under a recessive mode of inheritance. However, the realistic experiment (Experiment II) shows that SKAT and V-SS perform well. We believe that the relative difference in both experiments is due to the fact that the first experiment is biased towards



**Figure 2. Empirical power of SS-T methods under Experiment II as a function of the mode of inheritance, the number of causal variants in a LD block (*k*) and the ADR adjustment status.** The nominal type I error rate is  $\alpha=0.05$ . The bars represent the 95% confidence interval for the power of test. See Figure 1 for background and abbreviations.

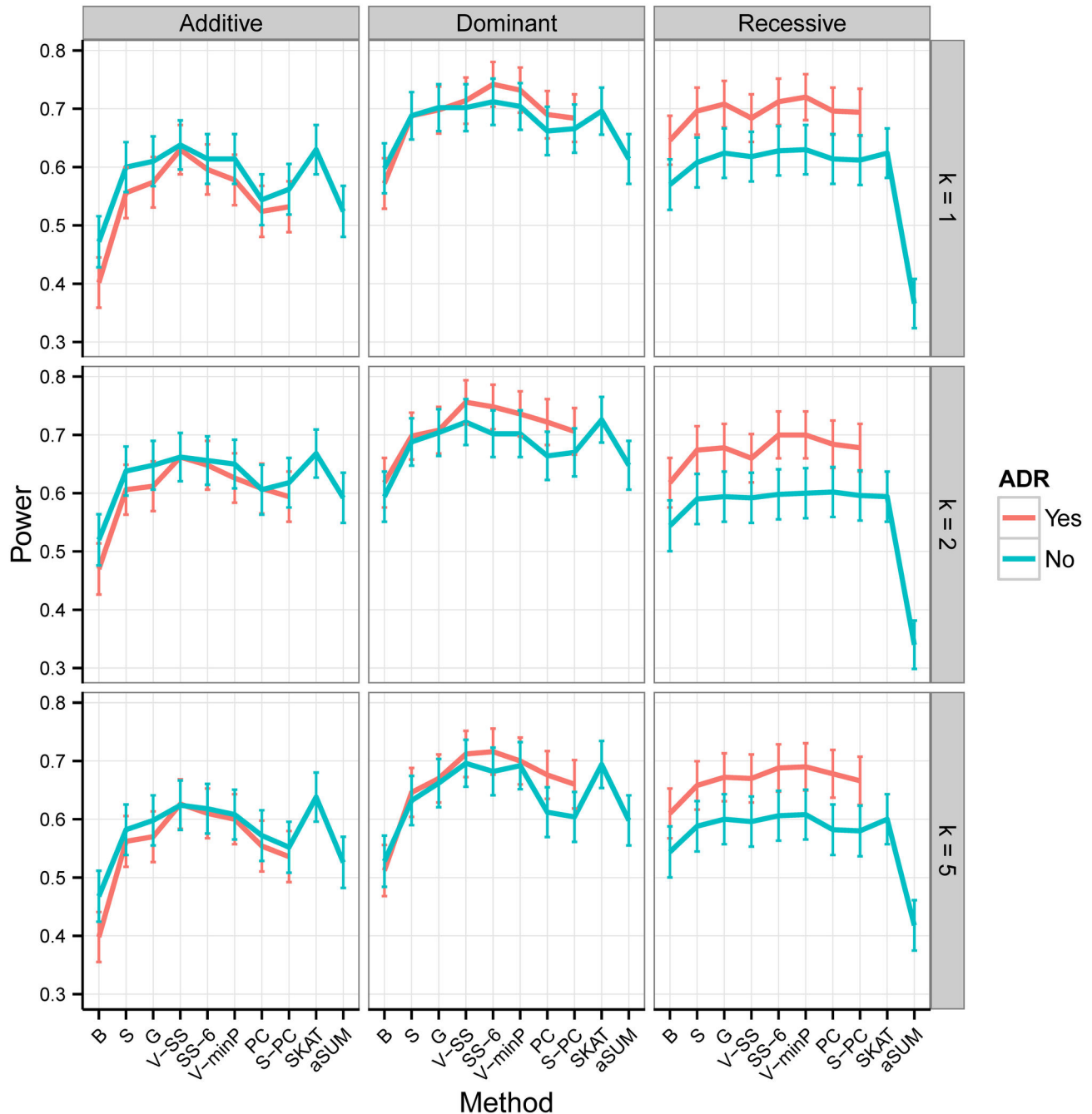
doi: 10.1371/journal.pone.0080540.g002

common tag SNP panels, whereas the second simulation experiment is geared towards denser, sequencing SNP panel.

Based on our simulation results, if the disease associated SNPs are highly likely to be acting additively (multiplicatively) or dominantly, for data coming from sequencing panels, we recommend the use of SKAT, ADR V-SS, ADR V-minP, and

ADR SS-T. However, researchers rarely know the mode of inheritance for a variant. When no prior information regarding the underlying mode of inheritance is available, we recommend the use of methods with good performance across all modes of inheritance and all simulation experiments, i.e. ADR V-SS, ADR V-minP, ADR GATES and ADR SS-T. However, we





**Figure 3. Empirical power of the main methods under Experiment II as a function of the mode of inheritance, the number of causal variants in a LD block ( $k$ ) and the ADR adjustment status.** The nominal type I error rate is  $\alpha=0.05$ . The bars represent the 95% confidence interval for the power of test. See Figure 1 for background and abbreviations.

doi: 10.1371/journal.pone.0080540.g003

believe that if the authors of SKAT and aSUM implement ADR adjustments in their software, these methods would become desirable tools for the analyses of SNPs in LD blocks regardless of the underlying mode of inheritance.

### Supporting Information

**Figure S1. The size of the test under Experiment I as a function of the method type, the number of SNPs ( $m$ ), the polychoric correlation ( $\rho$ ) between the genotypes of SNPs**

**in the LD block and the ADR adjustment status.** The nominal type I error rate is  $\alpha=0.05$ . Abbreviations for methods are as follows: B - Bonferroni, S - Simes, G - GATES, V-SS - VEGAS-SS, SS-x - SS-T with  $x=1, \dots, 9$ , V-minP - VEGAS-minP, PC - principal component method, S-PC - Simes adjustment of Simes and PC methods, SKAT - sequence kernel association test, aSUM - data-adaptive sum test.

(TIF)

**Figure S2. Empirical power of SS-T methods for the single causal variant scenario ( $k=1$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.01$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S3. Empirical power of SS-T methods for the single causal variant scenario ( $k=1$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.05$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S4. Empirical power of SS-T methods for the single causal variant scenario ( $k=1$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.10$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S5. Empirical power of SS-T methods for the dual non-interacting causal variant scenario ( $k=2$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.01$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S6. Empirical power of SS-T methods for the dual non-interacting causal variant scenario ( $k=2$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.05$  and the nominal type I error

rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S7. Empirical power of SS-T methods for the dual non-interacting causal variant scenario ( $k=2$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.10$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S8. Empirical power of the main methods for the single causal variant scenario ( $k=1$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.01$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S9. Empirical power of the main methods for the single causal variant scenario ( $k=1$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.05$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S10. Empirical power of the main methods for the single causal variant scenario ( $k=1$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.10$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S11. Empirical power of the main methods for the dual non-interacting causal variant scenario ( $k=2$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.01$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S12. Empirical power of the main methods for the dual non-interacting causal variant scenario ( $k=2$ ) under**



**Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.05$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**Figure S13. Empirical power of the main methods for the dual non-interacting causal variant scenario ( $k=2$ ) under Experiment I as a function of the mode of inheritance (panels), the number of SNPs in the LD block ( $m$ ), the polychoric correlation between the genotypes of SNPs in the LD block ( $\rho$ ) and the ADR adjustment status.** The causal allele frequency is  $p_a=0.10$  and the nominal type I error rate is  $\alpha=0.05$ . See Figure S1 for background and abbreviations.

(TIF)

**methods S1. Simulation of genotype-phenotype data.**

(DOC)

**Table S1. Settings used in Experiment I.**

## References

- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106: 9362-9367. doi:10.1073/pnas.0903103106. PubMed: 19474294.
- Roeder K, Bacanu SA, Sonpar V, Zhang XH, Devlin B (2005) Analysis of single-locus tests to detect gene/disease associations. *Genet Epidemiol* 28: 207-219. doi:10.1002/gepi.20050. PubMed: 15637715.
- Longmate JA (2001) Complexity and power in case-control association studies. *Am J Hum Genet* 68: 1229-1237. doi:10.1086/320106. PubMed: 11294658.
- Sasieni PD (1997) From genotypes to genes: Doubling the sample size. *Biometrics* 53: 1253-1261. doi:10.2307/2533494. PubMed: 9423247.
- Dewan A, Liu MG, Hartman S, Zhang SSM, Liu DTL et al. (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* 314: 989-992. doi:10.1126/science.1133807. PubMed: 17053108.
- Franke A, Balschun T, Karlsen TH, Sventoraityte J, Nikolaus S et al. (2008) Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat Genet* 40: 1319-1323. doi:10.1038/ng.221. PubMed: 18836448.
- Churchill GA, Doerge RW (1994) Empirical Threshold Values for Quantitative Trait Mapping. *Genetics* 138: 963-971. PubMed: 7851788.
- Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P et al. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 39: 645-649. doi:10.1038/ng.2022. PubMed: 17401363.
- Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 39: 870-874. doi:10.1038/ng.2075. PubMed: 17529973.
- Gastwirth JL (1985) The Use of Maximin Efficiency Robust-Tests in Combining Contingency-Tables and Survival Analysis. *J Am Stat Assoc* 80: 380-384. doi:10.1080/01621459.1985.10478127.
- Freidlin B, Zheng G, Li Z, Gastwirth JL (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered* 53: 146-152. doi:10.1159/000064976. PubMed: 12145550. 64976 . PII.
- Zheng G, Ng HK (2008) Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics* 9: 391-399. doi: 10.1093/biostatistics/kxm039. PubMed: 18003629. kxm039 . PII;Retrieved unpublished at whilst December year 1111 from . doi: 10.1093/biostatistics/kxm039
- Chen ZX, Ng HKT (2012) A Robust Method for Testing Association in Genome-Wide Association Studies. *Hum Hered* 73: 26-34. doi: 10.1159/000334719. PubMed: 22212363.
- Gastwirth JL (1966) On Robust Procedures. *J Am Stat Assoc* 61: 929-& doi:10.1080/01621459.1966.10482185.
- Wang K, Sheffield VC (2005) A constrained-likelihood approach to marker-trait association studies. *Am J Hum Genet* 77: 768-780. doi: 10.1086/497434. PubMed: 16252237.
- Lette G, Lange C, Hirschhorn JN (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol* 31: 358-362. doi:10.1002/gepi.20217. PubMed: 17352422.
- Bacanu SA, Nelson MR, Ehm MG (2008) Comparison of Association Methods for Dense Marker Data. *Genet Epidemiol* 32: 791-799. doi: 10.1002/gepi.20347. PubMed: 18551558.
- Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061-1073. doi:10.1038/nature09534. PubMed: 20981092. nature09534 . PII;Retrieved unpublished at whilst December year 1111 from . doi:10.1038/nature09534
- Neale BM, Sham PC (2004) The future of association studies: Gene-based analysis and replication. *Am J Hum Genet* 75: 353-362. doi: 10.1086/423901. PubMed: 15272419.
- Wang T, Elston RC (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet* 80: 353-360. S0002- doi:10.1086/511312. PubMed: 17236140. 9297(07)62693-7 . PII;Retrieved unpublished at whilst December year 1111 from . doi: 10.1086/511312
- Han F, Pan W (2010) A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. *Hum Hered* 70: 42-54. doi:10.1159/000288704. PubMed: 20413981.
- Gauderman WJ, Murcray C, Gilliland F, Conti DV (2007) Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol* 31: 383-395. doi:10.1002/gepi.20219. PubMed: 17410554.
- Wang K, Abbott D (2008) A principal components regression approach to multilocus genetic association studies. *Genet Epidemiol* 32: 108-118. doi:10.1002/gepi.20266. PubMed: 17849491.

(DOC)

**Table S2. Number of cases/controls ( $n$ ), relative risk of heterozygote to non-risk allele homozygote ( $R_1$ ) and relative risk of risk allele homozygote to non-risk allele homozygote ( $R_2$ ) used at each simulation setting under the single causal variant scenario ( $k=1$ ) of Experiment I.** Within each cell, the settings are presented as  $n$  ( $R_1$ ,  $R_2$ ).

(DOC)

**Table S3. Number of cases/controls ( $n$ ) and effect size ( $\delta$ ) of any causal allele used at each simulation setting under the two non-interacting causal variant scenario ( $k=2$ ) of Experiment I.** Within each cell, the settings are presented as  $n$  ( $\delta$ ).

(DOC)

## Author Contributions

Conceived and designed the experiments: DL SAB. Performed the experiments: DL SAB. Analyzed the data: DL SAB. Contributed reagents/materials/analysis tools: DL SAB. Wrote the manuscript: DL SAB.

24. Bacanu SA (2012) On optimal gene-based analysis of genome scans. *Genet Epidemiol* 36: 333-339. doi:10.1002/gepi.21625. PubMed: 22508187.
25. Wu MC, Lee S, Cai TX, Li Y, Boehnke M et al. (2011) Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am J Hum Genet* 89: 82-93. doi:10.1016/j.ajhg.2011.05.029. PubMed: 21737059.
26. Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR et al. (2010) A Versatile Gene-Based Test for Genome-wide Association Studies. *Am J Hum Genet* 87: 139-145. doi:10.1016/j.ajhg.2010.06.009. PubMed: 20598278.
27. Li MX, Gui HS, Kwan JSH, Sham PC (2011) GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *American J Hum Genet* 88: 283-293. doi:10.1016/j.ajhg.2011.01.019.
28. Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* 87: 52-58. doi:10.1046/j.1365-2540.2001.00901.x. PubMed: 11678987.
29. Galwey NW (2009) A New Measure of the Effective Number of Tests, a Practical Tool for Comparing Families of Non-Independent Significance Tests. *Genet Epidemiol* 33: 559-568. doi:10.1002/gepi.20408. PubMed: 19217024.
30. Gao XY, Starmer J, Martin ER (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* 32: 361-369. doi:10.1002/gepi.20310. PubMed: 18271029.
31. Moskvina V, Schmidt KM (2008) On multiple-testing correction in genome-wide association studies. *Genet Epidemiol* 32: 567-573. doi:10.1002/gepi.20331. PubMed: 18425821.
32. Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 74: 765-769. doi:10.1086/383251. PubMed: 14997420.
33. Guo YJ, Jamison DC (2005) The distribution of SNPs in human gene regulatory regions. *Bmc Genomics* 6: 140-. PubMed: 16209714.
34. Zalsman G, Huang YY, Oquendo MA, Burke AK, Hu XZ et al. (2006) Association of a triallelic serotonin transporter gene promoter region (5-HTTLPR) polymorphism with stressful life events and severity of depression. *Am J Psychiatry* 163: 1588-1593. doi:10.1176/appi.ajp.163.9.1588. PubMed: 16946185.
35. Simes RJ (1986) An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrika* 73: 751-754. doi:10.1093/biomet/73.3.751.
36. Olsson U (1979) Maximum Likelihood Estimation of the Polychloric Correlation-Coefficient. *Psychometrika* 44: 443-460. doi:10.1007/BF02296207.