OPEN ACCESS Freely available online

PLOS ONE

# A Robust Model-free Approach for Rare Variants Association Studies Incorporating Gene-Gene and Gene-Environmental Interactions

**Ruixue Fan, Shaw-Hwa Lo***

Department of Statistics, Columbia University, New York, New York, United States of America

## Abstract

Recently more and more evidence suggest that rare variants with much lower minor allele frequencies play significant roles in disease etiology. Advances in next-generation sequencing technologies will lead to many more rare variants association studies. Several statistical methods have been proposed to assess the effect of rare variants by aggregating information from multiple loci across a genetic region and testing the association between the phenotype and aggregated genotype. One limitation of existing methods is that they only look into the marginal effects of rare variants but do not systematically take into account effects due to interactions among rare variants and between rare variants and environmental factors. In this article, we propose the *s*ummation of *p*artition *a*pproach (SPA), a robust model-free method that is designed specifically for detecting both marginal effects and effects due to gene-gene (G×G) and gene-environmental (G×E) interactions for rare variants association studies. SPA has three advantages. First, it accounts for the interaction information and gains considerable power in the presence of unknown and complicated G×G or G×E interactions. Secondly, it does not sacrifice the marginal detection power; in the situation when rare variants only have marginal effects it is comparable with the most competitive method in current literature. Thirdly, it is easy to extend and can incorporate more complex interactions; other practitioners and scientists can tailor the procedure to fit their own study friendly. Our simulation studies show that SPA is considerably more powerful than many existing methods in the presence of G×G and G×E interactions.

**Citation:** Fan R, Lo S-H (2013) A Robust Model-free Approach for Rare Variants Association Studies Incorporating Gene-Gene and Gene-Environmental Interactions. PLoS ONE 8(12): e83057. doi:10.1371/journal.pone.0083057

**Editor:** Zhi Wei, New Jersey Institute of Technology, United States of America

**Received** May 24, 2013; **Accepted** October 30, 2013; **Published** December 17, 2013



**Funding:** The authors have no support or funding to report.

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

* E-mail: slo@stat.columbia.edu(SHL)

## Introduction

Despite of the success of large scale biological studies such as GWAS in discovering many disease variants, most of which are common variants with minor allele frequency (MAF) greater than 0.05, for diabetes, heart disease, Alzheimer disease, etc., the variants identified thus far confer relatively small risk, explain a small fraction of familial clustering, and add little practical value in disease prediction. The issue of so-called ''missing heritability'' has been a serious concern that has attracted considerable attention and discussion recently. [1,2,3,4,5] A number of explanations have been suggested for this phenomenon including: (1) an as-yet undiscovered larger set of variants of smaller effects, (2) rare variants with larger effects that may be eluding the current GWAS, (3) unaccounted effects, due to gene-gene (G×G) and gene-environment (G×E) interactions, (4) undetected structure effects including copy number variations (CNVs), and (5) over-estimated heritability.[6,7,8,9,10,11] This article presents a simple yet easy-to-extend method to address issues (2) and (3).

In genetic association studies, the 'common-disease common-variants' (CDCV) model states that common diseases are caused by common variants with MAFs greater than 5% or 1%. [12] However, recently more and more evidence support the alternative 'common-disease rare-variants' (CDRV) hypothesis which claims that complex disorders are caused by multiple rare variants with MAF$<$1%. [13,14] Unlike common variants that do not affect protein function directly, most rare variants are missense mutations in promoter region or protein coding regions and they are capable of altering gene expression level, changing amino acids sequence and affecting protein-protein interactions. [15,16] Furthermore, rare variants may have higher odds ratios (above 2), compared with small odds ratios (1.1∼1.5) of common variants. [17] Therefore, the investigation of rare variants will help researchers further understand the disease etiology and may provide new insights into medical treatments. With the development and commercialization of next generation sequencing technologies, large number of SNPs with low frequencies can be detected in a relatively short time and at relatively low cost. [5] In the near future, whole-genome sequencing will become possible for large numbers of individuals, and, as a result, large amounts of sequence data with rare variants will be generated. Methods that are capable of detecting these casual variants are very much in need.

Due to the low frequencies and large number of rare variants, traditional single-marker association tests that have worked well for common variants will in general lack power for rare variants. [18] In recent years, several statistical methods have been developed based on collapsing rare variants in a specific region of interest, e.g. a gene or genes from a specific pathway, followed by performing a region-based test rather than individual tests for

each variants. The Combined Multivariate and Collapsing (CMC) method proposed by Li and Leal [19] tests whether the proportions of rare variants carriers in cases and controls are significantly different. The weighted sum (WS) method by Madsen and Browning [20] is designed to weight variants according to their estimated frequencies in controls, so that less frequent variants receive higher weights compared with more common variants. Instead of using the conventional cutoff values 0.05 or 0.01 to define rare variants, Price et al. [21] proposed to choose a variable threshold (VT) that gives an optimal testing power. Ionita-Laza et al. [22] developed a replication-based (RB) approach, also based on a weighted-sum statistic, that can be more powerful in the presence of both risk and protective variants. Wu et al. proposed a sequence kernel association test (SKAT) that is a score-based variance component test. [23] SKAT uses a linear weighted kernel $K(G_i,G_{i'})=\sum_{j=1}^{K} w_j G_{ij} G_{i'j}$ to measure the similarity between individuals $i$ and $i'$ ($K$ is the number of markers and $w_j$ is the weight of SNP $j$). A weighted quadratic kernel $K(G_i,G_{i'})=\left(1+\sum_{j=1}^{K} w_j G_{ij} G_{i'j}\right)^2$ was also proposed in [23] to account for both main effects and genetic interaction effects but it was not systematically studied. Many alternative methods that have also been proposed can be considered variations of these approaches. [24,25,26]

### Why another approach?

The aforesaid methods have been shown to work well in different simulated models (mostly with marginal effects only). However, all these tests only consider marginal effects from rare variants and they do not systemically address the issue of interactions among rare variants (G×G), or between rare variants and covariates, such as environmental factors (G×E). Therefore, additional statistical methods are needed to generate scientific knowledge on the etiology of complex diseases where interactions among genetic, biological and environmental variables work together to produce a phenotype. In this article, we propose the *s*ummation of *p*artition *a*pproach (SPA), a robust model-free method that not only tests the marginal effects of rare SNPs but also naturally incorporates G×G and G×E interactions. As with existing methods, SPA is based on aggregating information across rare variants in a region of interest. We shall demonstrate the power of SPA and compare with existing methods for both dichotomous and quantitative phenotypes. Simulation studies show that in disease models without interactions, the performance of SPA is comparable to or even better than the most competitive existing method in current literature, and in the presence of G×G interactions, SPA substantially outperforms all the other methods. Another advantage of our procedure is its simplicity and extensibility. We also demonstrate in this article how to incorporate an environmental factor in the proposed framework and show that the augmented test score is powerful in detecting G×E interactions. Similar approaches can be taken to account for interactions with common variants or other covariates. In addition, we compare the proposed method with several existing tests on the dataset provide by Genetic Analysis Workshop 17 (GAW17) and find that SPA is robust for detecting different genes. When large volumes of datasets with rare variants become available in the near future, the proposed procedure will become a powerful tool to detect complicated interaction effects in various genetic regions and it will help us to better understand the mechanisms of complex human diseases.

## Materials and Methods

To better understand the motivation and rational behind SPA, we briefly review a general framework that has been adopted for detecting common variants with interactions. A core element in this framework is the influence score $I$ derived from what we now know as the Partition Retention (PR) method. [27] Several forms and variations were associated with the PR method before it was finally coined this name in 2009.

### A General Framework Used for Detecting Common Variants

We demonstrate a basic tool adopted by our method. Suppose there are $n$ subjects with a response variable $Y$ and K discrete explanatory variables $\{X_1,\ldots, X_K\}$. If each $X_i$ can take three discrete values, we generate a partition $\Pi$ with $3^K$ non-overlapping partition elements. Let $n_i$ be the number of subjects in partition $i$, $\overline{Y}_i$ the average response for subjects in partition $i$, and $\overline{Y}$ the average response from all subjects. An influence measure between the response and the predictors is defined as:

$$I(X_1,\ldots,X_K)=\sum_{i\in\Pi} n_i^2\left(\overline{Y}_i-\overline{Y}\right)^2$$

It has been shown that under the null hypothesis that none of the predictors has influence on $Y$, the normalized $I$, $I/(n\sigma^2)$($\sigma^2$ denotes the variance of $Y$) is asymptotically distributed as a weighted sum of $\chi^2$ random variables of 1 degree of freedom each such that the total weight is less than 1. [27] The main structure of this measure is the partition formed by the $K$ discrete variables with $3^K$ partition elements each containing non-overlapping observations. This influence measure captures any discrepancy between the conditional mean and the grand mean of $Y$ and thus is able to detect $X$-$Y$ association regardless of the structure of dependence. It can be easily generalized to any discrete random variables with finite number of outcomes.

In case-control studies, the influence measure can be rewritten as:

$$I=\sum_{i\in\Pi} n_i^2\left(\hat{p}_i^D-\frac{N_A}{N_A+N_U}\right)^2$$

where $N_A$ is the number of affected individuals, $N_U$ is the number of unaffected individuals, and $\hat{p}_i^D$ is the proportion of cases in partition $i$. Several variations of this partition-based method have been successful at identifying influential common variants and their interactions in human diseases, such as Rheumatoid Arthritis [28,29,30] and breast cancer [31,32]. Its success in detecting common variants relies on the essence that many partition cells contain more than singleton subjects, however, this property will diminish for rare variants due to their extremely low frequencies. To effectively deal with rare variants, we need to modify the partition procedure properly to accommodate for the sparseness, which can be achieved by the proposed *s*ummation of *p*artition *a*pproach (SPA). We introduce below several test statistics of SPA, including the marginal test score $I_1$, G×G interaction score $I_2$, and G×E interaction scores $I_2^*$.

### Rare Variants Marginal Association Score $I_1$

The general framework mentioned above can be extended to rare variants association analysis for both dichotomous and continuous phenotypes.

In population-based case-control studies, suppose there are $N$ unrelated individuals, among which $N_A$ are cases and $N_U = N - N_A$ are controls. The region of interest $G$ contains $K$ rare variants and the genotype of the $j^{th}$ individual is denoted $(X_1^{(j)},...,X_K^{(j)})$. Each $X_i^{(j)} (i = 1,...,K)$ can take values 0, 1 or 2, indicating the number of rare variants at this position. The SPA test score $I_1$ that accounts for all marginal information contributed by these $K$ rare SNPs is defined as:

$$I_1 = \sum_{i=1}^{K} n_i^2 \left( \hat{p}_i^D - \frac{N_A}{N_A + N_U} \right)^2$$

where $\hat{p}_i^D$, for the $i^{th}$ SNP, is the fraction of all observed rare variants that are from cases, and $n_i$ is the total number of $i^{th}$rare variant observed in all subjects.

For continuous traits, $I_1$ is defined as:

$$I_1 = \sum_{i=1}^{K} n_i^2 \left( \overline{Y}_i - \overline{Y} \right)^2$$

where $\overline{Y}_i$, for the $i^{th}$ SNP, is the averaged response for subjects bearing at least one rare variant, $\overline{Y}$ is the averaged response from all subjects and $n_i$ is defined as above. Different from the original influence measure, $I_1$ recognizes the partition elements formed by individual SNP and hence the partitions from different SNPs are not non-overlapping any more; therefore, $I_1$ does not suffer from the sparseness of rare variants. Under the null hypothesis of no influence, the differences between $\hat{p}_i^D$ and $\frac{N_A}{N_A + N_U}$ for dichotomous traits (or between $\overline{Y}_i$ and $\overline{Y}$ for continuous traits) for all $i$ are small, so a large $I_1$ value indicates that some rare variants in the region might be associated with the disease phenotype. Additionally, since each term of $I_1$ is the squared difference between the conditional average and the grand average, it can detect both directions of departure from the expected difference zero, which renders $I_1$ the ability to capture the association even in a region with both risk and protective rare variants. Unlike PR's influence measure $I$, the statistical property of $I_1$ is more complicated to obtain since the dependence between partition cells created by different SNPs will not asymptotically disappear even under the null hypothesis of no influence. Therefore, in our analyses we will rely on the method of permutation to assess its statistical significance.

## Rare Variants G×G Interaction Association Score $I_2$

In order to increase the power of detecting the genotype-phenotype associations as well as to elucidate the biological pathways that underpin disease, more and more attentions have been given to the identification of interactions between SNP loci. [33,34,35] A limitation of $I_1$ is that it considers little interactions among rare SNPs. From the general framework, we propose a second SPA test score $I_2$ that evaluates G×G interactions among rare variants.

As the genotype at each SNP position can take 3 values, in theory we are facing a maximum of $3^K$ partition elements for all levels of interactions. However, due to the low frequencies of rare variants, the higher order (>2) interaction information among rare SNPs in current sample size will be small. For example, if the sample size is 1,000 and the SNP frequency is 0.01, the expected number of observing one specific rare variants triplet is $1{,}000 \times 0.01^3 = 10^{-3}$. If a region contains 20 independent rare SNPs, the expected total number of rare variants triplets would be $\binom{20}{3} \times 0.001 = 1.14$, which provides very low signal for 3-way interaction detection. Therefore, for current sample size, we only consider an influence measure that takes into account 2-way interactions among rare variants. For a pair of rare SNPs $i$ and $j$, we consider three aggregated cells: individuals with rare variants only on SNP $i$ (denoted $mM$), individuals with rare variants only on SNP $j$ (denoted $Mm$) and individuals with rare variants on both SNPs (denoted $mm$). Note that we do not consider the cell $MM$ where individuals have no rare variant at either position. For dichotomous trait, the SPA test score $I_2$ for G×G interaction is defined as:

$$I_2 = \sum_{i \geq 1, j > i}^{K} n_{ij}^2 \left[ \left( \hat{p}_{ij,mM}^D - \frac{N_A}{N_A + N_U} \right)^2 + \left( \hat{p}_{ij,Mm}^D - \frac{N_A}{N_A + N_U} \right)^2 + \left( \hat{p}_{ij,mm}^D - \frac{N_A}{N_A + N_U} \right)^2 \right]$$

where $n_{ij}$ is the number of subjects who have at least one rare variant in either SNP ($i$ or $j$), $\hat{p}_{ij,mM}^D$ is the fraction of subjects that are cases in partition $mM$, $\hat{p}_{ij,Mm}^D$ is that fraction in partition $Mm$, and $\hat{p}_{ij,mm}^D$ in partition $mm$. For quantitative trait, $I_2$ is defined as:

$$I_2 = \sum_{i \geq 1, j > i}^{K} n_{ij}^2 \left[ \left( \bar{Y}_{ij,mM} - \bar{Y} \right)^2 + \left( \bar{Y}_{ij,Mm} - \bar{Y} \right)^2 + \left( \bar{Y}_{ij,mm} - \bar{Y} \right)^2 \right]$$

where $\bar{Y}_{ij,mM}$ is the average response for individuals in partition $mM$, $\bar{Y}_{ij,Mm}$ in partition $Mm$, and $\bar{Y}_{ij,mm}$ in partition $mm$. If two rare variants have interactions, the difference between the conditional average and the unconditional average will be large, leading to a large $I_2$ value. Again, permutation is used to evaluate the significance of the test statistic $I_2$. Even though $I_2$ only considers 2-way interaction, it can be easily extended to include higher-order (≥3) interactions by generating partitions based on $m$-tuples ($m \geq 3$) of rare SNPs.

## Adaptive Test Score $p^*$

When we are unclear whether G×G interaction is involved in the onset of disease, we propose an adaptive score $p^*$ that is a compromise between $I_1$ and $I_2$. We first evaluate the significance of $I_1$ and $I_2$. Then the adaptive test score is defined as: $p* = min(p(I_1), p(I_2))$ where $p(I_1)$ and $p(I_2)$ are the p-values of $I_1$ and $I_2$ separately. We evaluate the significance of $p^*$ by permutation.

## Rare Variants G×E Interaction Association Score $I_2^*$

Increasing evidence have shown that gene and environmental (G×E) interactions are widely involved in the etiology of complex diseases, including diabetes, cancer and psychiatric disorders [36,37,38,39,40]. Conventional methods to detect G×E interactions are mostly based on regression models, which will lose power for rare variants. SPA can be easily extended to incorporate covariates, such as environmental factors in the testing procedure, considering both the environmental marginal effect and the G×E interaction information. Here we focus on case-control study design. Suppose an environmental factor $E$ has $J$ levels. The SPA test score for detecting the effect of the environmental factor is expressed as:

**Table 1.** Type I error estimates of $I_1$, $I_2$and $p*$.

| **Dichotomous Trait** | | | | | | |
|---|---|---|---|---|---|---|
| | **α = 0.05** | | | **α = 0.01** | | |
| Sample Size | $I_1$ | $I_2$ | $p*$ | $I_1$ | $I_2$ | $p*$ |
| 600 | 0.052 | 0.055 | 0.054 | 0.009 | 0.012 | 0.009 |
| 1000 | 0.053 | 0.055 | 0.053 | 0.010 | 0.010 | 0.010 |
| 1500 | 0.048 | 0.049 | 0.049 | 0.007 | 0.007 | 0.007 |
| 2000 | 0.053 | 0.057 | 0.053 | 0.010 | 0.013 | 0.010 |
| **Continuous Trait** | | | | | | |
| | **α = 0.05** | | | **α = 0.01** | | |
| Sample Size | $I_1$ | $I_2$ | $p*$ | $I_1$ | $I_2$ | $p*$ |
| 600 | 0.055 | 0.055 | 0.058 | 0.013 | 0.011 | 0.013 |
| 1000 | 0.05 | 0.046 | 0.044 | 0.009 | 0.005 | 0.010 |
| 1500 | 0.061 | 0.048 | 0.061 | 0.015 | 0.011 | 0.010 |
| 2000 | 0.045 | 0.046 | 0.043 | 0.013 | 0.009 | 0.009 |

doi:10.1371/journal.pone.0083057.t001

$$I_2^* = \sum_{j=1}^{J}\sum_{i=1}^{K} n_i^2, j\left(\hat{p}_{i,j}^D - \frac{N_A}{N_A+N_U}\right)^2$$

where $\hat{p}_{i,j}^D$ is the fraction of rare variants at position $i$ on level $j$ that are from cases, and $n_{i,j}$is the total number of $i^{th}$rare variants observed at level $j$. $I_2^*$ is a modification of $I_1$ by building additional overlapping rare variants partition cells to $J$ non-overlapping partitions created by the environmental factor. The significance of $I_2^*$is evaluated by permutation. We propose two permutation strategies: (1) global permutation that permutes the phenotype among all individuals; and (2) local permutation that permutes the phenotype within each stratum of the environmental factor. Both permutation strategies are investigated in our study.

## Simulation Scheme

We simulated several scenarios for the purpose of evaluation and comparison of our test scores with several existing rare variants association methods. The genotype consists of 20 independent rare variants in each scenario. Scenario 'Null-1' is a 'null model' where none of the 20 variants affects the phenotype. For dichotomous traits, the phenotypes are determined by the baseline penetrance only. This is the null setting for $I_1$, $I_2$, $p$* and $I_2^*$ with global permutation. In scenario 'Null-2', the dichotomous outcomes are affected by the environmental factor. 'Null-2' is the null setting for $I_2^*$with local permutation.

For empirical power comparisons, we generate three different sets of simulations. The first set of simulations are marginal effect models, in which the MAF of all SNPs are uniformly distributed between 0.0001 and 0.01. In scenario 1, 5 out of the 20 rare SNPs are risk SNPs and the effect size is constant. Scenario 2 is similar to scenario 1 except that the risk effect is negatively correlated with MAF. Scenario 3 has 5 protective variants and 5 deleterious variants with effect size negatively correlated with MAF. The second set of simulations contains 2-way G×G interaction between rare variants, with MAF 0.01 for all 20 SNPs. In scenario 4, 50% of the SNPs (10 out of 20 SNPs) have interaction effects. Scenario 5 is similar to scenario 4 but 75% of the SNPs are involved in G×G interactions. Both main effect and G×G interaction effect exist in scenario 6. The third set of simulation models involves G×E interaction effects with a binary environmental factor. Scenario 7 has positive G×E interaction effects and environmental marginal effect; scenario 8 has both positive and negative G×E interaction effects. Logistic regressions or linear regression was used to generate dichotomous or quantitative phenotypes. 1,000 repetitions were simulated for each scenario with four different sample sizes, each having equal number of cases and controls. Detailed simulation models are provided in Table S1 in file S1.

# Results

We compared the power of SPA test scores $I_1$, $I_2$ and $p$* with existing methods: CMC, WS, VT, RB SKAT (with the weighted linear kernel) and SKATint (a modified SKAT score with the weighted quadratic kernel) in a series of simulation scenarios, including marginal effect models and G×G interaction effect models for both dichotomous traits and continuous traits. RB only deals with binary outcomes, so it is not included in our analysis for continuous traits. We also evaluated the power of $I_2^*$ in G×E interaction effect models for dichotomous traits. (See *Material and Methods* for details of simulation models; numerical results from our simulation studies are presented in Table S2−S7 in file S1.)

## Type I Error of $I_1$, $I_2$ and $p$*

The empirical type I error rates for $I_1$, $I_2$ and $p$* are presented in Table 1 for nominal levels $\alpha = 0.05$ and $\alpha = 0.01$ with four different sample sizes: 600, 1000, 1500 and 2000. The results show that $I_1$, $I_2$ and $p$* are well controlled at both significance levels for either dichotomous or continuous trait, even when the sample size is small, indicating that the proposed tests are valid methods. Additional results of type I error for competing methods are presented in Fig. S1 in file S1.

## Power Comparison in Marginal Effect Models for both Dichotomous and Continuous Traits

We compare the power of $I_1$, $I_2$ and $p$* with competing methods in three marginal effect models when (1) only risk variants exist and the effect size is constant, (2) only risk variants exist and the

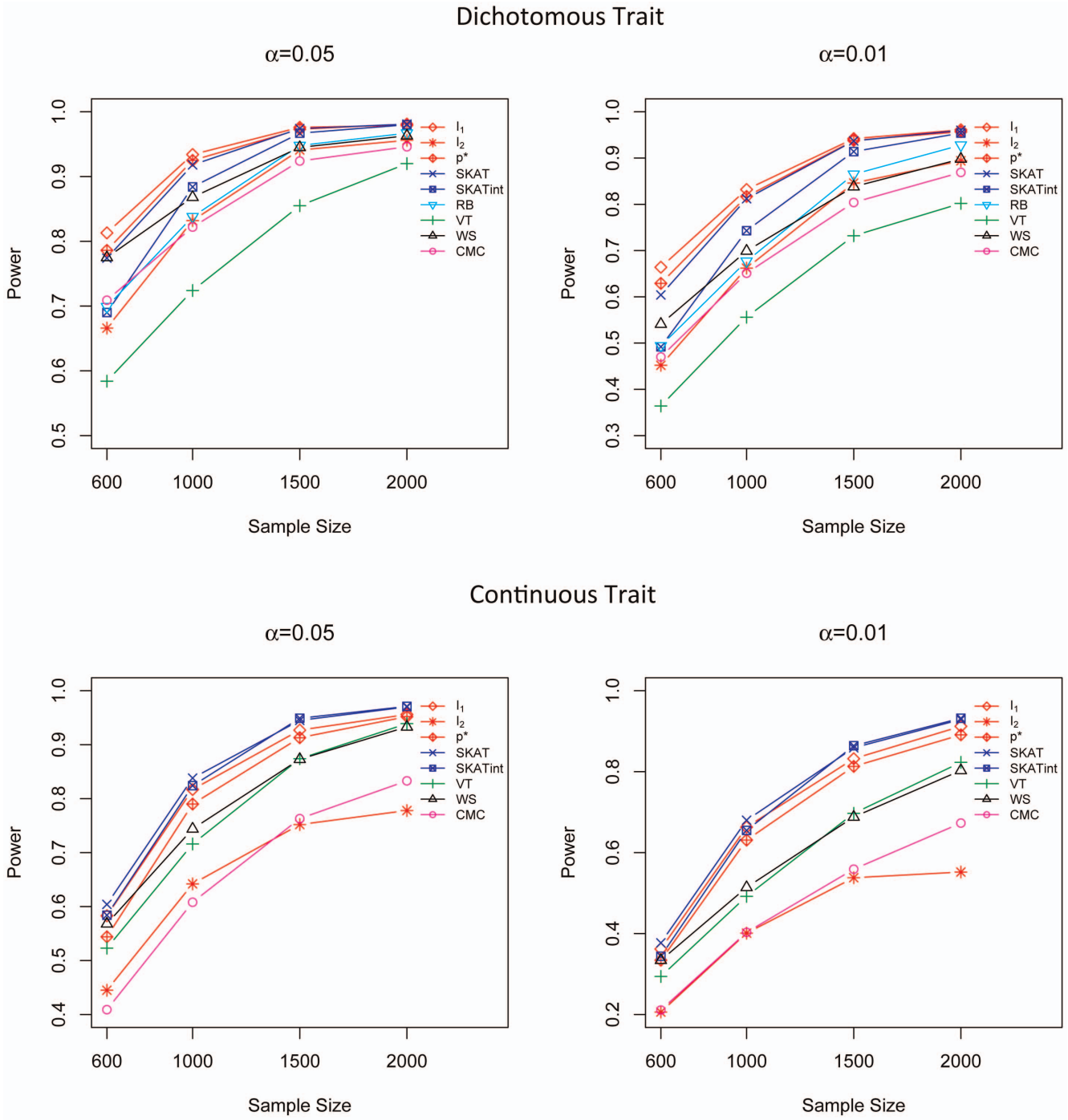


**Figure 1. Power comparison in the marginal effect model when the effect sizes are constant.** Powers were calculated for nominal α levels 0.05 (left) and 0.01(right) and for dichotomous traits (upper) and continuous traits (lower). Powers were evaluated for $I_1$, $I_2$, $p^*$, SKAT, SKATint, VT, RB, WS and CMC. Scenarios with different sample sizes were considered. P-values were estimated using 10,000 permutations and power was evaluated using 1,000 replicates.
doi:10.1371/journal.pone.0083057.g001

effect size is negatively correlated with MAF, or (3) a mixture of risk and protective rare variants exists.

In all three marginal effect scenarios, the performance of $I_1$ and SKAT are comparable and they are both superior to the other tests (Fig. 1, 2 and 3). For dichotomous traits, $I_1$ is the most powerful method, followed by SKAT and $p^*$. For continuous traits, SKAT and $I_1$ are most competitive; both of them are more powerful than the other methods. The power of the adaptive score $p^*$ is very close to $I_1$,; $p^*$ is much more powerful than CMC, WS, VT and RB. In addition, $I_1$ and $p^*$ are quite robust to different simulation scenarios, even in the presence of a mixture of risk and protective variants, while CMC, WS and VT suffer substantial power loss when causal rare variants have opposite effects (Fig. 3). It is worth noting that although $I_1$ does not intentionally highlight

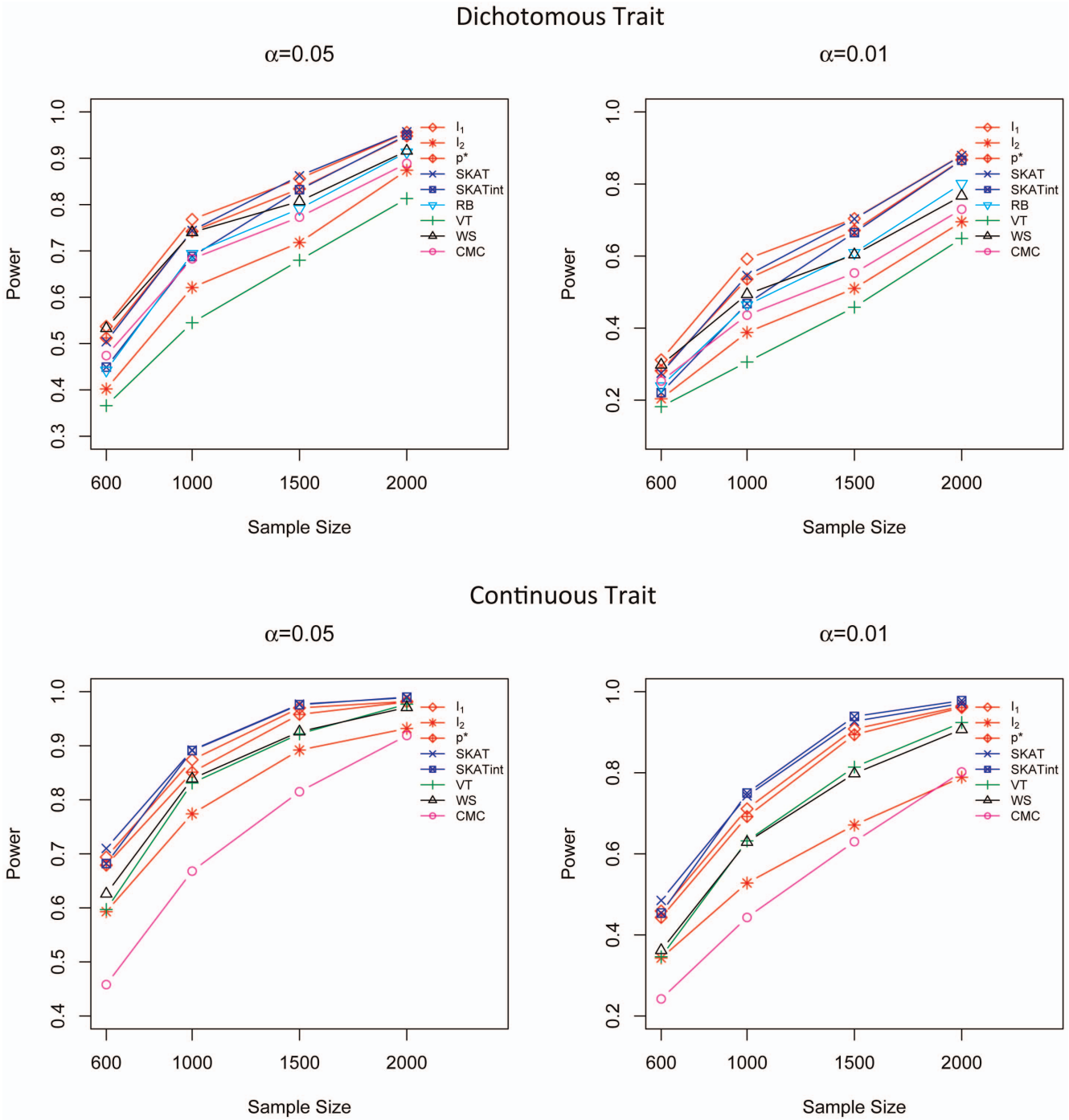


**Figure 2. Power comparison in the marginal effect model when the effect sizes of causal variants are negatively correlated with MAFs.** Powers were calculated for nominal α levels 0.05 (left) and 0.01(right) and for dichotomous traits (upper) and continuous traits (lower). Powers were evaluated for $I_1$, $I_2$, $p^*$, SKAT, SKATint, VT, RB, WS and CMC. Scenarios with different sample sizes were considered. P-values were estimated using 10,000 permutations and power was evaluated using 1,000 replicates.
doi:10.1371/journal.pone.0083057.g002

less frequent variants by giving them higher weights, it is still the most powerful (for dichotomous trait) or the second most powerful (for quantitative traits) method even in scenarios where the effect size is negatively correlated with MAF, showing that its good performance is intrinsic and is not driven by a specific weighting scheme. The test score $I_2$ does not show a high power in these marginal effect models as it is designed to detect G×G interaction effects but not the marginal effect.

## Power Comparison for G×G Interaction Effect Models for both Dichotomous and Continuous Traits

We evaluated the power of different methods in two G×G interaction effect models (scenarios 4 and 5). The advantage of the

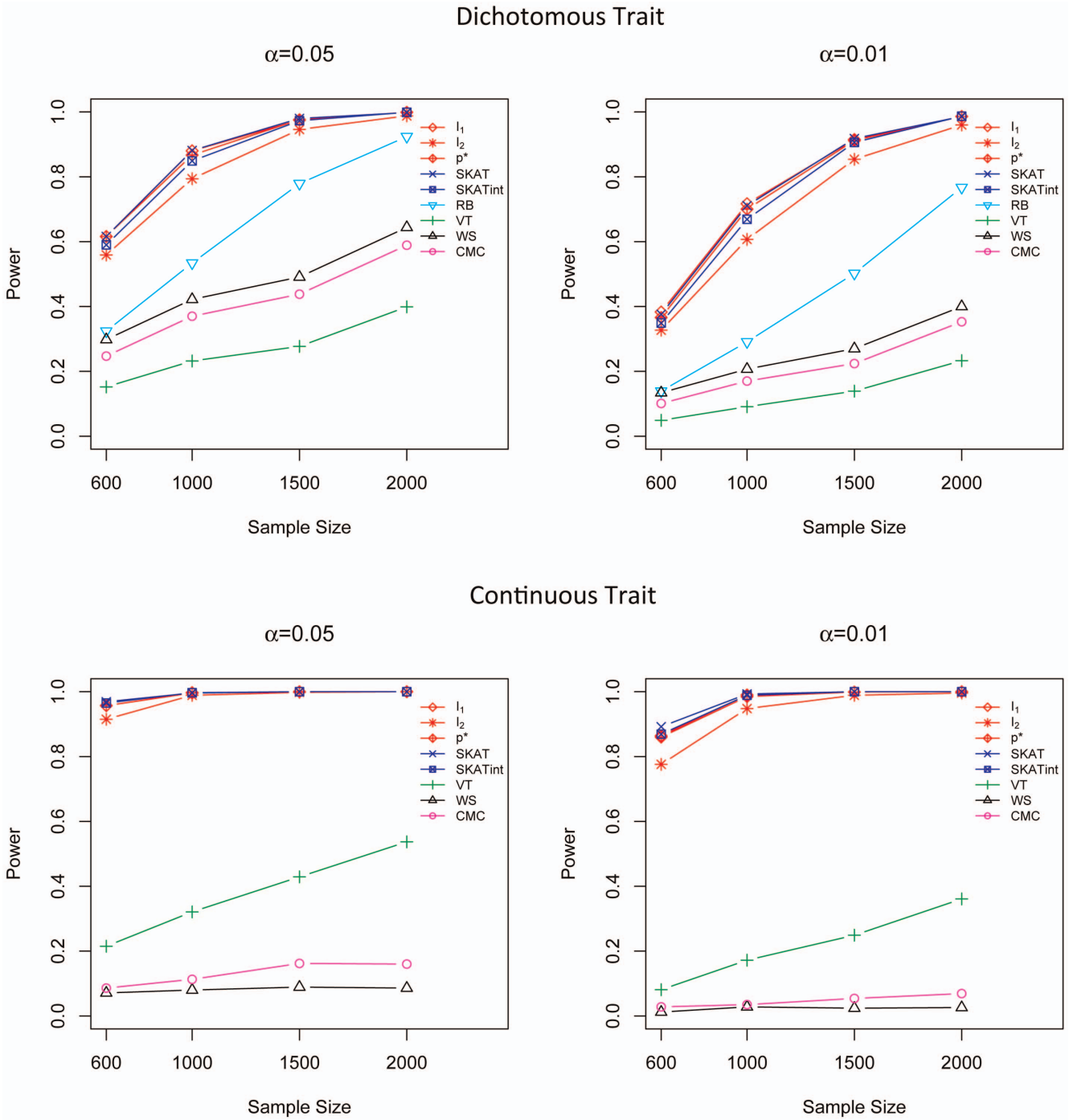


**Figure 3. Power comparison in the marginal effect model with a mixture of protective and risk rare variants.** Powers were calculated for nominal α levels 0.05 (left) and 0.01(right) and for dichotomous traits (upper) and continuous traits (lower). Powers were evaluated for $I_1$, $I_2$, $p^*$, SKAT, SKATint, VT, RB, WS and CMC. Scenarios with different sample sizes were considered. P-values were estimated using 10,000 permutations and power was evaluated using 1,000 replicates.
doi:10.1371/journal.pone.0083057.g003

G×G interaction association score $I_2$ over all the other methods is apparent for both dichotomous and continuous traits (Fig. 4 and Fig. 5). For dichotomous traits, when the sample size is large, the power of $I_2$ is substantially higher than all the other methods. For continuous traits, $I_2$ is uniformly the most powerful method for all sample sizes; for example, when the sample size is 2000, $I_2$ is 38% more powerful than SKATint at $\alpha = 0.01$. Moreover, the adaptive score $p^*$ has a power that is just slightly less than $I_2$, and $p^*$ is substantially more powerful than the rest. On the other hand, VT, WS and CMC suffer from significant loss of power in the presence of complicated G×G interaction effects.

We also examine the scenario in which the phenotypes are influenced by both genetic marginal and G×G interaction effects (scenario 6). Here the marginal effect is set to be small so that it

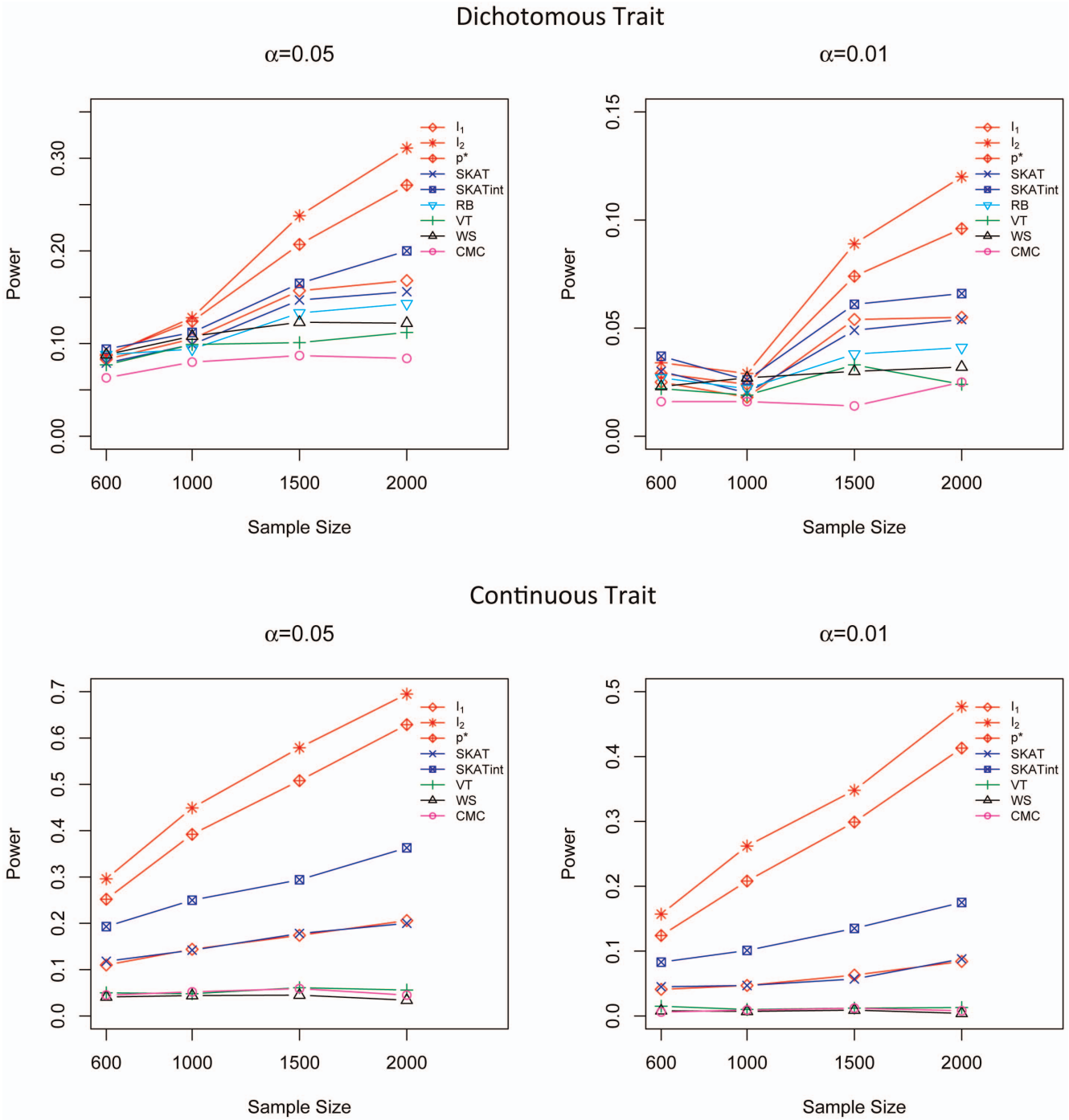


**Figure 4. Power comparison in G×G interaction effect model when 50% of rare variants participate in the interaction effect.** Powers are calculated for nominal α levels 0.05 (left) and 0.01(right) and for dichotomous traits (upper) and continuous trait (lower). Power was evaluated for $I_1$, $I_2$, $p^*$, SKAT, SKATint, VT, RB, WS and CMC. Scenarios with different sample sizes were considered. P-values were estimated using 10,000 permutations and power was evaluated using 1,000 replicates.
doi:10.1371/journal.pone.0083057.g004

will not mask the interaction effect. $I_2$ is still consistently the most powerful test and $p^*$ is the second best, followed by SKATint (Fig. 6). For continuous traits with sample size 2000, $I_2$ is 29% more powerful than SKATint, and $p^*$ is 28% more powerful than SKATint at $\alpha = 0.01$.

## Type I Error and Power of $I_2^*$ for Dichotomous Trait

For the G×E interaction score $I_2^*$, we investigated its type I error and power for dichotomous trait using two permutation strategies – global permutation and local permutation (see *Materials and Methods*), denoted by $I_2^*$-*Global* and $I_2^*$-*Local* respectively. As $I_2^*$ considers both the genetic and environmental marginal effects as

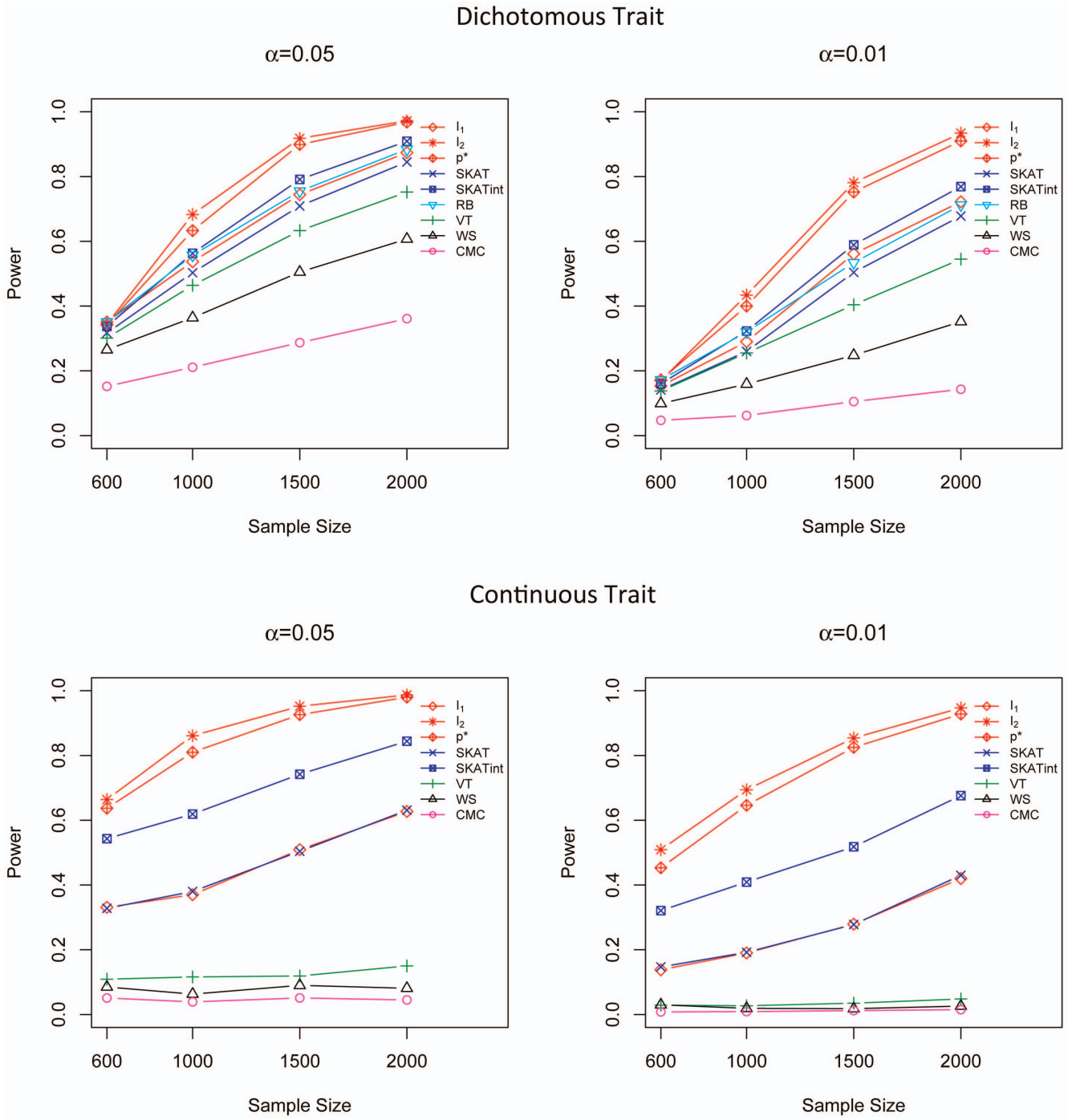


**Figure 5. Power comparison in G×G interaction effect model when 75% of rare variants participate in the interaction effect.** Powers are calculated for nominal α levels 0.05 (left) and 0.01(right) and for dichotomous traits (upper) and continuous trait (lower). Power was evaluated for $I_1$, $I_2$, $p^*$, SKAT, SKATint, VT, RB, WS and CMC. Scenarios with different sample sizes were considered. P-values were estimated using 10,000 permutations and power was evaluated using 1,000 replicates.
doi:10.1371/journal.pone.0083057.g005

well as G×E interaction effect, $I_2^*$-*Global* is appropriate for testing the null hypothesis of no association at all (no G marginal, E marginal or G×E interaction effects), and $I_2^*$-*Local* is appropriate for testing the null hypothesis of no E marginal effect.

The type I error of $I_2^*$ are evaluated for two null hypotheses. The first null hypothesis (null-1) assumes the dichotomous traits are completely determined by the baseline penetrance. The second null hypothesis (null-2) assumes that the phenotypes are affected by environmental marginal (E marginal) effect. Table 2 presents the type I error of $I_2^*$-*Global* and $I_2^*$-*Local* in these two null settings. In null-1, both $I_2^*$-*Global* and $I_2^*$-*Local* are correctly controlled at levels $\alpha = 0.05$ and 0.01. In null-2, $I_2^*$-*Local* still hits the target level while $I_2^*$-*Global* has significant higher values. This is because $I_2^*$-*Global* is able to test any effect from genetic or environmental factors,

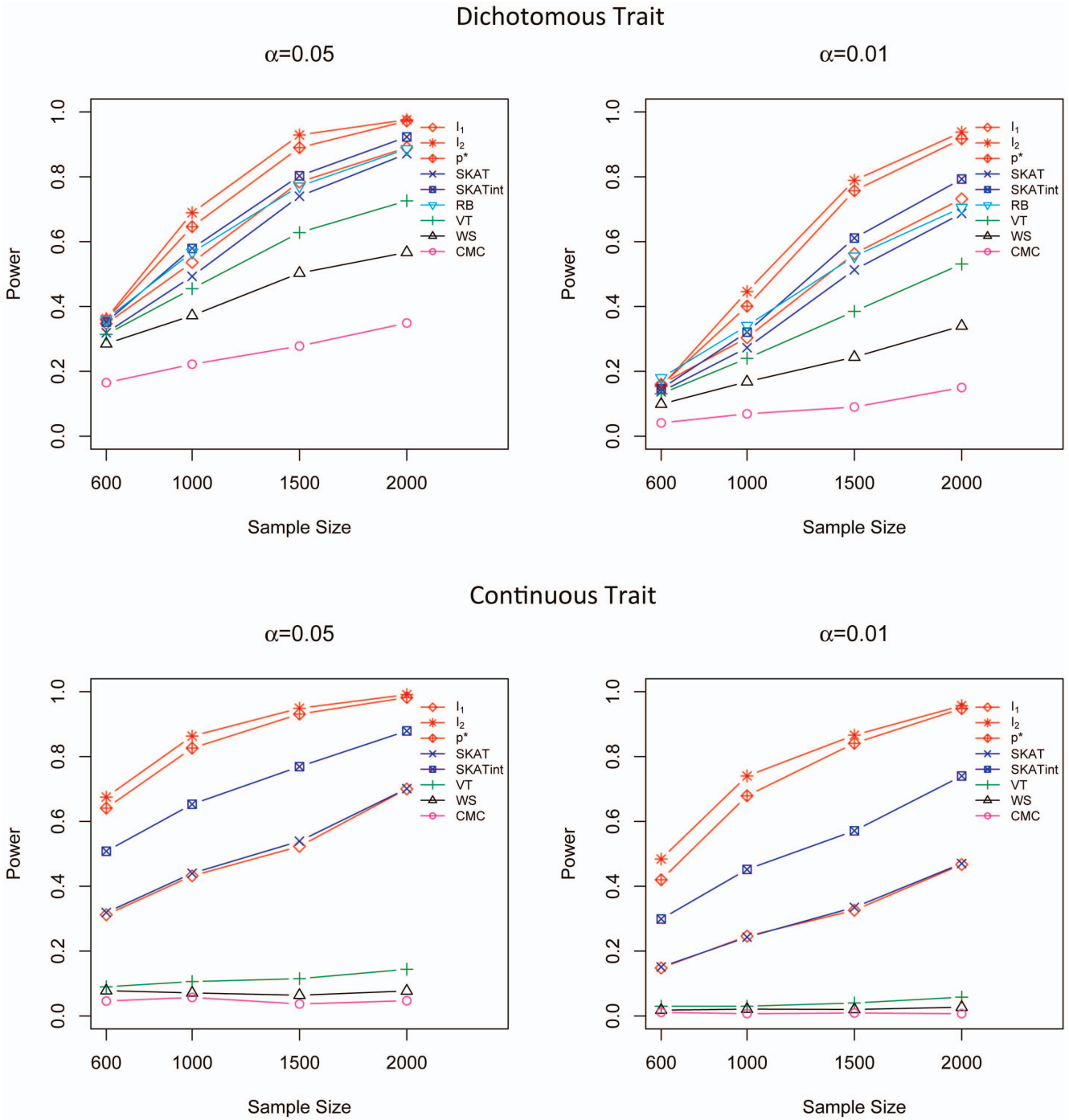


**Figure 6. Power comparison in the scenario with both main effect and G×G interaction effect.** Powers are calculated for nominal α levels 0.05 (left) and 0.01(right) and for dichotomous traits (upper) and continuous trait (lower). Power was evaluated for $I_1$, $I_2$, $p^*$, SKAT, SKATint, VT, RB, WS and CMC. Scenarios with different sample sizes were considered. P-values were estimated using 10,000 permutations and power was evaluated using 1,000 replicates.
doi:10.1371/journal.pone.0083057.g006

including the E marginal effect; hence the results of $I_2^*$-*Global* in null-2 are indeed the *power* of $I_2^*$-*Global* in the presence of E marginal effect. On the other hand, $I_2^*$-*Local* removes the E marginal effect, so it shows the correct type-I error in both null-1 and null-2.

Two scenarios are considered to compare the power of $I_2^*$-*Global*, $I_2^*$-*Local* and competing methods when (1) the phenotypes are affected by E marginal effect and positive G×E effect, (2) the phenotypes are affected by E marginal effect and both positive and negative G×E effects. In computation, SKAT and SKATint regress the phenotype on the environmental factor when calculating the test statistic [23]. $I_2^*$-*Global* and $I_2^*$-*Local* use the environmental factor as in their definition. All the other methods work on the phenotype and the genotype directly. The results

**Table 2.** Type I error estimates of $I_2^*$ in two different null settings.

| **Null-1: No G, E or G×E Effects** | | | | |
|---|---|---|---|---|
| | **α = 0.05** | | **α = 0.01** | |
| Sample Size | $_1I_2^*$-*Global* | $_1I_2^*$-*Local* | $_1I_2^*$-*Global* | $_1I_2^*$-*Local* |
| 600 | 0.053 | 0.050 | 0.007 | 0.007 |
| 1000 | 0.047 | 0.046 | 0.009 | 0.007 |
| 1500 | 0.045 | 0.048 | 0.008 | 0.009 |
| 2000 | 0.043 | 0.044 | 0.009 | 0.009 |
| **Null-2 : Marginal Environmental Effect only** | | | | |
| | **α = 0.05** | | **α = 0.01** | |
| Sample Size | $_1I_2^*$-*Global*[a] | $_1I_2^*$-*Local* | $_1I_2^*$-*Global*[a] | $_1I_2^*$-*Local* |
| 600 | 0.110 | 0.046 | 0.027 | 0.007 |
| 1000 | 0.169 | 0.050 | 0.058 | 0.012 |
| 1500 | 0.239 | 0.047 | 0.087 | 0.011 |
| 2000 | 0.282 | 0.046 | 0.114 | 0.017 |

[a]This is actually the **power** of $I_2^*$-*Global* in the presence of marginal environmental effect.

doi:10.1371/journal.pone.0083057.t002

show that $I_2^*$-*Global* has much higher power than all the other tests because it takes into account both E marginal and G×E interaction effects, and $I_2^*$-*Local* outperforms all the remaining methods that do not consider G×E interaction effects (Fig. 7).

## Application to the GAW17 Dataset

The genetic analysis workshop 17 (GAW17) provided genotypes of 3,205 autosomal genes on 697 individuals from the 1000 Genome Project. A dichotomous phenotype was simulated from a linear model using SNPs from 34 genes and most causal SNPs were rare variants. A total of 200 simulation replicates were carried out and the genotype was held fixed for all replicates. See [41] for more details of the simulation model. Here we chose to re-analyze two causal genes *FLT1* and *ANRT*. In the workshop, *FLT1* has been shown to exhibit a strong signal in many well-known methods while *ARNT* could not be identified by any existing approach. For both genes, an upper frequency of 0.05 was used as the MAF cutoff to define rare variants and only nonsynonymous SNPs were examined. We computed the power of our test scores and competing methods using all 200 replicates. Power was calculated as the proportion of replicates with p-value less than 0.05 out of the 200 simulations. As shown in Table 3, $I_1$ was fairly robust for detecting both genes. For *FLT1*, two count-based collapsing methods – CMC and WS are most powerful, followed by VT and $I_1$. For *ARNT*, $I_1$ is substantially more powerful than the other methods – its power is 47% higher than the second best method SKAT. Given that the simulated model is a simple additive linear model with genetic marginal effects only, methods considering G×G interactions, including $I_2$ and SKATint, do not have apparent advantages in power gain for detecting either *FLT1* or *ARNT*.

## Computation Time

The computation time of $I_1$, $I_2$ and $p^*$ depends on the sample size, the number of variants and the number of permutations. On a 2.66 GHz laptop with 4 GB memory, to reach a significance level of $10^{-4}$, the computation times to analyze a region with 20 SNPs for 600, 1000, 1500 and 2000 individuals are 3, 5, 7, 10 sec for $I_1$, and are about 1000, 1400, 1900, 2500 sec for $I_2$.

# Discussion

We propose here the summation of partition approach (SPA), a flexible robust model-free framework for rare variants association analysis that incorporates both G×G and G×E interactions. The proposed SPA test scores create partitions from individual SNP and combine the information across all rare variants in a region of interest. $I_1$ is designed to detect marginal effects of rare variants and $I_2$ is designed to capture the G×G interaction effects among rare variants. In various marginal effect models, $I_1$ is more powerful than most approaches examined in our study. Its performance is comparable to SKAT, which is regarded as the most competitive existing method. In G×G interaction models or in the scenario with both marginal and G×G interaction effects, $I_2$ is superior to all the other methods in terms of detection power. The adaptive score $p^*$ is a compromise between $I_1$ and $I_2$ and has the advantage of both test scores. Its performance is just a little shy of the better of the two scores $I_1$ and $I_2$, for both marginal effect models and interaction effect models. Therefore, $p^*$ is a self-tuning adaptive score that is able to gain power automatically regardless of the simulation scenario. In practice when we have no clue of how the genotype affects the phenotype, we suggest to use the adaptive score $p^*$. A significant p-value of $p^*$ indicates a potential true signal from either marginal or interaction effects of rare variants. In our study, we focus on the situation with 20 rare SNPs. If the SNP number changes to 30, the simulation results (Fig. S2 in file S1) are qualitatively similar in that $I_1$ is the most powerful in marginal effect models and $I_2$ is the most powerful in interaction effect models.

$I_2^*$ is an augmented score of $I_1$ that incorporates covariates. It can be used to test the hypothesis of no association at all (neither G marginal, E marginal nor G×E effect) using 'global permutation' or to test the hypothesis of no E marginal effect using 'local permutation'. By 'local permutation', $I_2^*$ removes the marginal effect of the environmental factor while still captures variations of the genetic effect at different levels of the environmental factor. In a similar fashion, covariates can be incorporated into $I_2$ and the resulting augmented score could be used to detect E×G×G 3-way interactions between an environmental factor and two rare variants.

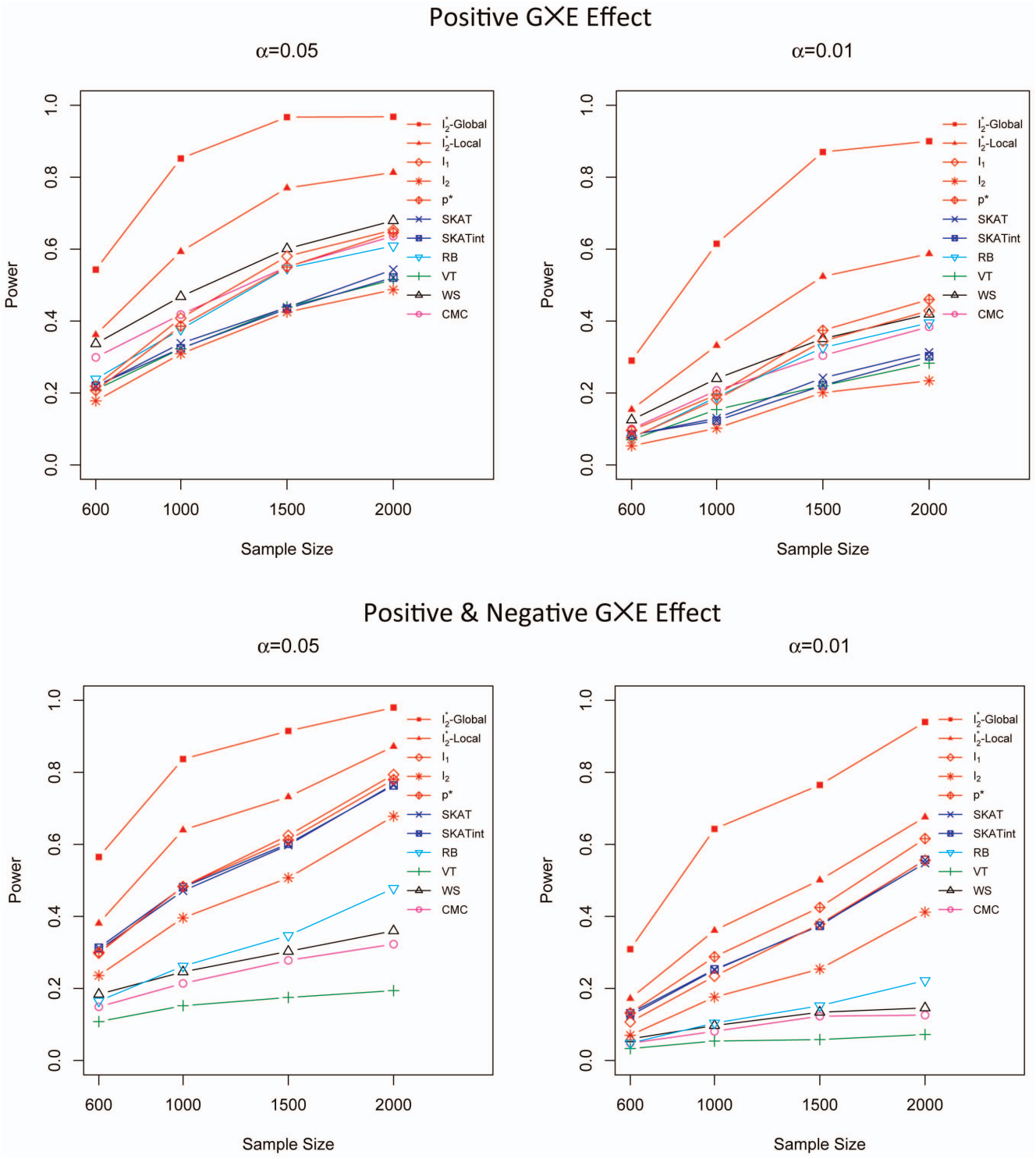


**Figure 7. Power comparison in two G×E interaction models for dichotomous trait.** Powers were calculated for nominal α levels 0.05 (left) and 0.01(right) when only positive G×E effects exist (upper) and when both positive and negative G×E effects exist (lower). Powers were evaluated for $_1I_2^*$(with both global and local permutations), $I_1$, $I_2$, $p^*$, SKAT, SKATint, VT, RB, WS and CMC. Scenarios with different sample sizes were considered. P-values were estimated using 10,000 permutations and power was evaluated using 1,000 replicates.
doi:10.1371/journal.pone.0083057.g007

$I_2^*$ can also be used to test the interaction effect between common and rare variants if one treats the common variant as an environmental factor. It can be further extended to detect 3-way interactions among the environmental factor, common and rare variants by building additional overlapping partitions based on rare variants on top of the non-overlapping partition cells generated by the environmental factor and the common variant. A global permutation can detect both main and interaction effects

**Table 3.** Power of two genes in GAW17 dataset.

| | FLT1 | ARNT |
|---|---|---|
| #Rare NS[b] SNPs | 19 (10 causal) | 9 (5 causal) |
| $I_1$ | 0.865 | 0.345 |
| $I_2$ | 0.505 | 0.05 |
| $p^*$ | 0.775 | 0.22 |
| SKAT | 0.82 | 0.235 |
| SKATint | 0.77 | 0.1 |
| RB | 0 | 0.005 |
| VT | 0.88 | 0.025 |
| WS | 0.95 | 0.075 |
| CMC | 0.95 | 0.055 |

[b]NS: nonsynonymous.

doi:10.1371/journal.pone.0083057.t003

of these factors, and a local permutation that permutes the phenotype within each non-overlapping partition cell will capture the E×common×rare 3-way interaction effect.

$I_2^*$ deals with categorical covariates naturally. In order to handle continuous covariates, such as age, height and BMI, we suggest taking the discretization approach that divides continuous variables into distinct buckets. These 'pseudo-categorical' variables generated by discretization can be applied to $I_2^*$ directly. In practice, we usually set the number of buckets to be 2∼5 and the results are quite satisfactory. Moreover, a new influence measure dealing with continuous covariates directly is under preparation.

The insight of SPA is similar to the partition retention (PR) influence measure as in [27]. The PR method generates non-overlapping partition elements over the sample space and assigns each partition cell a weight that is proportional to the probability of falling into that cell. Its success in detecting influential variables relies on the essence that weights are not too small for all partition elements, especially for those cells that generate signals. Therefore, the PR method may lose power for rare variants association studies as the partition cells with true signals will have very low weights due to the extremely low frequencies of rare variants. SPA differs from the PR method by creating overlapping partition elements to avoid the sparseness and to boost the signal from rare variants.

The information measure $I_1$ can be viewed as a special case of $I_1 = \sum_{i=1}^{K} w_i n_i \left(\hat{p}_i^D - \frac{N_A}{N_A+N_U}\right)^2$ where $\{w_i\}$are weights that sum to 1. Weights can be defined in various ways. The inherent choice we take here is $w_i = n_i \Big/ \sum_{i=1}^{K} n_i$. If external information is available on possible effects of a rare variant to disease, it is straightforward to incorporate such information in our test approach by tuning the weight. Some commonly used weights are based on (1) MAF of the variant as in [20]; or (2) externally-defined weights such as predictions from SIFT and PolyPhen, as suggested by Price et al. [21]. In our study, even though we do not incorporate the weight information, SPA is still superior over the other methods. We believe that after tuning the weight, SPA will exhibit a better performance.

Population stratification has been shown to be an important problem for common variant association analysis. For rare variants, this problem is more likely to occur due to their low frequency and possible uneven distribution among populations. It is straightforward to control population stratification in our approach as we can consider population as an environmental factor and apply it to $I_2^*$. An alternative is to treat population with PCA and include the discretized eigenvalues in our analysis.

A major advantage of SPA is that it is highly extensible. The building blocks of SPA are the partitions formed by individual rare variant and it is easy to incorporate complex interactions. As demonstrated in the article, we are able to take into account interactions with environmental factors. Similar approaches can be applied when considering interactions with common variants or other covariates. It can also be generalized to other research areas to benefit the practitioners and scientists in various fields. We believe that the proposed framework of SPA will offer substantial opportunities in detecting potential complicated interactions. Once interaction effects indeed exist, our approach is capable of identifying these interactions and thus adding to the detection power.

This paper presents a simple novel (and easily implemented) tool SPA as an alternative to existing statistical methods for rare variants association studies, with a unique additional feature that SPA can easily incorporate various forms of interaction effects. This addition may add considerable power to disease-related detection in the future. From our studies, if the underlying model is a simple linear additive model with only marginal effects, the powers of SPA are comparable to several existing methods. However, if the model is more complex with interaction effects, the proposed approach provides a more powerful alternative in rare variants association analysis so that there is a better chance to find disease-associated factors. With the development of next-generation sequencing techniques, more and more data with a large amount of rare variants will be generated. It is highly unlikely that the disease phenotype is associated with genetic factors through a simple linear main effect model, so the proposed approach is going to be a powerful and rewarding tool to explore the complicated interaction effects revealed by larger datasets. It is worth noting that any interaction pattern, whether it is linear or nonlinear, can be detected by SPA, since it is model-free and is not subject to any distribution assumptions. Therefore, it is very robust and effective regardless of how the genotype affects the phenotype. The R code of the proposed test scores is available to download at http://www.columbia.edu/rf2283/Software.html

## Supporting Information

**File S1 The supporting information file for article "A Robust Model-free Approach for Rare Variants Association Studies Incorporating Gene-Gene and Gene-Environmental Interactions".** It contains the files: Table S1. Models to generate simulated phenotypes; Table S2. Power of different methods for dichotomous traits in scenarios 1∼6 ($\alpha = 0.05$); Table S3. Power of different methods for dichotomous traits in scenarios 1∼6 ($\alpha = 0.01$); Table S4. Power of different methods for continuous traits in scenarios 1∼6 ($\alpha = 0.05$); Table S5. Power of different methods for continuous traits in scenarios 1∼6 ($\alpha = 0.01$); Table S6. Power of different methods for dichotomous traits in G×E interaction effect models ($\alpha = 0.05$); Figure S1: Type I error for different methods in various sample sizes with nominal $\alpha$ levels 0.05 (left) and 0.01(right); Figure S2: Power comparison in scenarios 1∼6 for dichotomous traits with 500 cases and 500 controls when the SNP number is 30.
(DOC)

## Acknowledgments

We thank Prof. Tian Zheng and Dr. Chien-Hsun Huang for their valuable discussions in the preparation of this paper. We also appreciate the reviewers' invaluable comments, which are both insightful and constructive.

## Author Contributions

Conceived and designed the experiments: RF SHL. Performed the experiments: RF SHL. Analyzed the data: RF SHL. Contributed reagents/materials/analysis tools: RF SHL. Wrote the paper: RF SHL.

## References

1. 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.
2. Gibson G (2010) Hints of hidden heritability in GWAS. Nature genetics 42: 558–560.
3. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21.
4. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, et al. (2011) Beyond missing heritability: prediction of complex traits. PLoS genetics 7: e1002051.
5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
6. Clayton DG (2009) Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS genetics 5: e1000540.
7. de los Campos G, Gianola D, Allison DB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nature reviews Genetics 11: 880–886.
8. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS genetics 5: e1000337.
9. Janssens AC, van Duijn CM (2008) Genome-based prediction of common diseases: advances and prospects. Human molecular genetics 17: R166–173.
10. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews Genetics 9: 356–369.
11. Witte JS (2010) Genome-wide association studies and beyond. Annual review of public health 31: 9–20 24 p following 20.
12. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, et al. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42: 570–575.
13. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40: 695–701.
14. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? American journal of human genetics 69: 124–137.
15. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305: 869–872.
16. Frayling IM, Beck NE, Ilyas M, Dove-Edwin I, Goodman P, et al. (1998) The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. Proceedings of the National Academy of Sciences of the United States of America 95: 10722–10727.
17. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133–141.
18. Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34: 188–193.
19. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311–321.
20. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384.
21. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832–838.
22. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet 7: e1001289.
23. Wu MC, Lee S, Cai T, 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.
24. Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Human heredity 70: 42–54.
25. Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic epidemiology 34: 188–193.
26. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. (2011) Testing for an unusual distribution of rare variants. PLoS genetics 7: e1001322.
27. Chernoff H, Lo SH, Zheng T (2009) Discovering influential variables: a method of partitions. Annals of Applied Statistics 3: 1335–1369.
28. Qiao B, Huang CH, Cong L, Xie J, Lo SH, et al. (2009) Genome Wide Gene Based Analysis of Rheumatoid Arthritis Associated Interaction with PTPN22 and HLADRB1. BMC Proceedings (Suppl 7): S132.
29. Huang CH, Cong L, Xie J, QIao B, Lo SH, et al. (2009) Rheumatoid Arthritis-Associated Gene-Gene Interaction Network for Rheumatoid Arthritis Candidate Genes. BMC Proc (Suppl 7): S75.
30. Ding Y, Cong L, Ionita-Laza I, Lo SH, Zheng T (2007) Constructing gene association networks for rheumatoid arthritis using the backward genotype-trait association (BGTA) algorithm. BMC Proceedings 1 Suppl 1: S13.
31. Lo SH, Chernoff H, Cong L, Ding Y, Zheng T (2008) Discovering interactions among BRCA1 and other candidate genes associated with sporadic breast cancer. Proc Natl Acad Sci U S A 105: 12387–12392.
32. Zheng T, Wang S, Cong L, Ding Y, Ionita-Laza I, et al. (2007) Joint study of genetic regulators for expression traits related to breast cancer. BMC Proceedings 1 Suppl 1: S10.
33. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics 37: 413–417.
34. Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, et al. (2011) Characterizing genetic interactions in human disease association studies using statistical epistasis networks. BMC bioinformatics 12: 364.
35. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nature reviews Genetics 10: 392–404.
36. Hamza TH, Chen H, Hill-Burns EM, Rhodes SL, Montimurro J, et al. (2011) Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. PLoS genetics 7: e1002237.
37. Andreasen CH, Mogensen MS, Borch-Johnsen K, Sandbaek A, Lauritzen T, et al. (2008) Non-replication of genome-wide based associations between common variants in INSIG2 and PFKP and obesity in studies of 18,014 Danes. PloS one 3: e2872.
38. Wang H, Lo SH, Zheng T, Hu I (2012) Interaction-based feature selection and classification for high-dimensional biological data. Bioinformatics 28: 2834–2842.
39. Thomas D (2010) Gene–environment-wide association studies: emerging approaches. Nature reviews Genetics 11: 259–272.
40. Thomas D (2010) Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annual review of public health 31: 21–36.
41. Almasy L, Dyer TD, Peralta JM, Kent JW, Jr., Charlesworth JC, et al. (2011) Genetic Analysis Workshop 17 mini-exome simulation. BMC Proceedings 5 Suppl 9: S2.