RESEARCH ARTICLE Open Access

A powerful hybrid approach to select top single-nucleotide polymorphisms for genome-wide association study

Jian Wang, Sanjay Shete*

Abstract

Background: Genome-wide association (GWA) study has recently become a powerful approach for detecting genetic variants for common diseases without prior knowledge of the variant's location or function. Generally, in GWA studies, the most significant single-nucleotide polymorphisms (SNPs) associated with top-ranked p values are selected in stage one, with follow-up in stage two. The value of selecting SNPs based on statistically significant p values is obvious. However, when minor allele frequencies (MAFs) are relatively low, less-significant p values can still correspond to higher odds ratios (ORs), which might be more useful for prediction of disease status. Therefore, if SNPs are selected using an approach based only on significant p values, some important genetic variants might be missed. We proposed a hybrid approach for selecting candidate SNPs from the discovery stage of GWA study, based on both p values and ORs, and conducted a simulation study to demonstrate the performance of our approach.

Results: The simulation results showed that our hybrid ranking approach was more powerful than the existing ranked p value approach for identifying relatively less-common SNPs. Meanwhile, the type I error probabilities of the hybrid approach is well-controlled at the end of the second stage of the two-stage GWA study.

Conclusions: In GWA studies, SNPs should be considered for inclusion based not only on ranked p values but also on ranked ORs.

Background

Genome-wide association (GWA) study has recently become a powerful approach for detecting genetic variants for common diseases without prior knowledge of the variant's location or function [1-4]. Currently, almost all the GWA studies are conducted in two stages: certain numbers (10-50) of the most significant single-nucleotide polymorphisms (SNPs) associated with top-ranked p values are selected in stage one, and follow-up is performed in stage two.

This two-stage approach has been widely used and has successfully identified SNPs with novel susceptibility for different complex diseases such as lung [5-7], prostate [8-11], and breast cancers [12-15]; glioma [16,17]; and type 2 diabetes [18-24]. The published GWA studies showed that in different studies, different numbers of the most significant SNPs were selected at stage one for follow-up. For example, in the GWA study of lung cancer, 10 top-ranked SNPs were selected in stage one [5], whereas in the GWA study of type 2 diabetes, 59 SNPs were selected in stage one [21]. While the p value can indicate when the association of an SNP with a disease is statistically significant, it does not consider the associated odds ratio (OR). The rationale for using OR as a selection criterion is that when minor allele frequencies (MAFs) of causal SNPs are relatively low, much less-significant p values would be observed, even if they could correspond to higher ORs, which might be more useful for prediction of disease status; nevertheless, the less significant p value may limit the inclusion of these significant variants in a replication study. Therefore, selecting a certain number of significant SNPs only based on p values might overlook some important genetic variants that could have an even greater effect on the disease causation.

* Correspondence: sshete@mdanderson.org
Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA

In this paper, we proposed a hybrid approach for selecting candidate SNPs from stage one of a GWA study that is based on both ranked p values and ranked ORs. Based on simulation studies, the power comparison results showed that our hybrid ranking approach was more powerful than the existing ranked p value approach for identifying relatively less-common SNPs. We performed additional simulation studies to investigate the type I error probabilities of the proposed hybrid approach, and found that the type I errors are well-controlled at the end of second stage of the two-stage GWA study.

Methods

For the hybrid approach proposed in this paper, ORs are considered in addition to p values when selecting SNPs in stage one. For stage one, we selected a set of SNPs: half were selected based on ranked p values, and the other half were selected based on ranked ORs.

To demonstrate the increased power of our proposed hybrid approach, we performed a simulation study. We assumed two independent disease-causing loci: D_1 and D_2. At locus D_1, we set the MAF at 10%, and at locus D_2, the MAF was set at 40%. Furthermore, we assumed that there were two marker loci, M_1 and M_2, which were associated with two disease-causing loci, D_1 and D_2, respectively. MAFs of markers M_1 and M_2 were also set at 10% and 40%, respectively. The disease loci and their corresponding marker loci were assumed to be in high linkage disequilibrium (LD, r^2 = 0.8). For all the SNPs, we used (0, 1, 2) to denote the three genotypes, where the values corresponded to the number of copies of the deleterious allele. This coding assumed an additive model, but different coding for representing a dominant or recessive model could have also been used. Our proposed hybrid approach was not restricted to an additive model. We defined a categorical random variable, Y = (0, 1), to indicate the case-control status: 0 represented individuals in the controls and 1 represented individuals in the cases. We used the logistic model defined below to simulate data.

$$\text{Logit}\,(P(Y=1)) = \beta_0 + \beta_1 D_1 + \beta_2 D_2.$$

For the purpose of our study, we chose different ORs for different loci. At locus D_1, the OR was set as 1.8 (β_1 = 0.5878), and at locus D_2, the OR was set as 1.5 (β_2 = 0.4055). β_0 was fixed as a constant 2. Since the markers M_1 and M_2 were not directly associated with the disease of interest, they were not included in the logistic model for simulation. The specific parameters for simulation studies are listed in Table 1.

In stage one, we first simulated genotypes for the disease-causing loci D_1 and D_2 for each individual. Given the dataset of realizations of SNPs D_1 and D_2, we randomly generated a disease status for each individual using the logistic model above. Then, conditioned on data of D_1, genotypes for marker locus M_1 were simulated using the r^2 value of 0.8. Similarly, genotypes for marker M_2 were simulated conditioned on data of D_2 with the same r^2 value. Because the markers M_1 and M_2 were in high LD with the disease loci D_1 and D_2, respectively, they were also associated with the simulated disease. In this way, we simulated a large amount of data for the population of interest and then randomly sampled 1,000 disease-related cases along with 1,000 normal controls from this population. In this study, unless otherwise specified, we employed logistic regression to obtain ORs and used Wald's test to assess significance.

Table 1 Parameters for Simulation under Alternative Hypothesis

	Disease Locus 1 (D_1)	Marker 1 (M_1)	Disease Locus 2 (D_2)	Marker 2 (M_2)
OR	1.8		1.5	
MAF	10%	10%	40%	40%

OR = odds ratio, MAF = minor allele frequency, LD = linkage disequilibrium. The LD between the disease loci and the corresponding marker loci is 0.8, and the two disease loci are in linkage equilibrium.

To investigate the power for each SNP to be selected in stage one and followed up in stage two, we generated 1,000,000 replicates of a single SNP under the null hypothesis of no association between the SNP and the disease, each replicate with 1,000 cases and 1,000 controls. To mimic realistic patterns of LD, we applied a forward-time population simulation software program (genomeSIMLA) to generate the 1,000,000 unassociated SNPs [25]. We used information about markers on human chromosome 2, including rs numbers, allele frequencies, recombination fractions, and positions, to seed the initial population. The initial population was then advanced through 1,000 generations of mating to create a pool of individuals. The case and control status for individuals was assigned based on a penetrance function assuming one disease-causal SNP. We then randomly permuted the case-control statuses of individuals to break the association between the SNP and disease of interest. Therefore, all the SNPs simulated were unassociated with the disease of interest. The p values and ORs of all simulated markers were evaluated using PLINK [26]. For the purpose of our simulation, we considered markers with a range of MAFs (10%~40%) that covered the MAFs defined for the disease loci. The p values and ORs obtained under the null hypothesis were used to determine the thresholds for selecting SNPs in stage one. We studied different selections of 10, 20, 30, and 40 SNPs in stage one to follow up on in stage two. Basically, we ranked the 1,000,000 SNPs with respect to p values as well as OR values. If investigators decided to select the top 20 SNPs in stage one using the existing ranked p-value approach, the top 20th-ranked p value out of the 1,000,000 p values would be the threshold for selection, and the SNPs in stage one with more significant p values than the threshold would be forwarded to stage two. Using our proposed hybrid approach, two thresholds for selection were considered. One threshold was the top 10th-ranked p value, and the other was the top 10th-ranked OR value. In order to take into account potential overlapping of the sets of top 10 p-value-based SNPs and top 10 OR-based SNPs, we would first pick SNPs based on p value threshold and then pick additional SNPs based on the OR threshold. The SNPs we selected in stage one either had more significant p values than the p value threshold or had larger OR values than the OR threshold. The reason for selecting 10 OR-based SNPs and 10 p-value-based SNPs was to have the same number (20) of total top SNPs carried over to stage two.

Further, we investigated what percentage of the SNPs could reach a genome-wide threshold for declaring significance. Therefore, for each replicate selected in stage one, we simulated independently another 1,000 cases and 1,000 controls using the same simulation approach and the same parameters. We then employed a joint analysis using fixed indicators for stages one and two, as joint analysis of two stages is efficient and always results in increased power to detect genetic variants [27]. For each pair of corresponding replicates from stages one and two, data from both stages were pooled into one data set. We assumed that the data from different stages were from possibly different sites. Therefore, in order to control for the possible confounding effects of sites, we used (1, 2) to denote the indicator for each stage, with 1 representing stage one and 2 representing stage two. Multivariable logistic regression analysis was applied to the SNP and the stage indicator, and the significance was estimated using Wald's test. All results were based on 1,000 replicates.

We performed additional simulations to examine the type I error probabilities of the proposed hybrid approach under the null hypothesis of no association between SNPs and disease. We assumed that there were two marker loci, M_1 and M_2, that were not associated with the disease of interest, with MAFs similar to that used in the power studies. We applied the same software, genomeSIMLA [25], to generate the unassociated SNPs. We initialized the population with small ranges of MAFs. We simulated 10,000,000 replicates each with 1,000 cases and 1,000 controls. The case-control status was assigned at random, independent of the markers, so that markers were unassociated with the disease status. As in the power comparison studies, the same sets of p values and ORs of 1,000,000 replicates of a single SNP under the null hypothesis were employed to determine the thresholds for selecting SNPs in stage one to follow up. Similarly, for each replicate selected in stage one, we independently simulated another replicate using the same simulation approach and same parameters and performed a joint analysis using data from both stages.

Results

Table 2 lists the resulting medians of negative logarithm to base 10 of the p values and ORs based on 1,000 replicates generated in stage one. The medians of OR estimators for the disease-causing loci D_1 and D_2 were 1.8016 and 1.4973, respectively, which had good agreement with the ORs used to generate the data sets (1.8 and 1.5). The results for D_1 and D_2 were only reported for comparison purposes. In real GWA studies, the causal SNPs might not be genotyped. Most of the real markers studied are probably in LD with the causal SNPs and, therefore, are associated with the disease of interest, such as the markers M_1 and M_2 in our simulation study. In stage two, we assumed that D_1 and D_2 were not observed, and therefore, we focused only on the study of two markers. The associated markers generally show lower ORs than those of the actual causal SNPs because of the imperfect LD (r^2 = 0.8). Therefore, it is not surprising to see that the medians of the OR estimators of markers M_1 and M_2 were smaller than the medians of the corresponding disease loci. As discussed previously, when MAFs of disease-causing SNPs are relatively low, p values would be less significant, even if the corresponding ORs are much higher. This situation applies to the following case. Marker M_1 (MAF = 10%) was associated with a higher OR (OR = 1.6777) but with a less-significant p value (-Log10 of p value = 4.9597); whereas M_2 (MAF = 40%) had a lower OR (OR = 1.4323) but a much more significant p value (-Log10 of p value = 7.1542). Therefore, ranking SNPs based only on p values will have less power for marker M_1 to be included in stage two, even though it has a higher OR and is potentially more useful for making predictions of disease status.

Table 2 Medians of Significance and Odds Ratios in Stage One Based on 1,000 Replicates

	Stage One			
Median	**D_1**	**M_1**	**D_2**	**M_2**
-Log10 of p value	6.0613	4.9597	8.7299	7.1542
OR	1.8016	1.6777	1.4973	1.4323

OR = odds ratio.

Power comparisons

We reported the powers for each SNP to be selected in stage one in Table 3. We considered ranges of SNPs selected in stage one (10 to 40). The results in Table 3 are grouped into two panels with respect to two different ranking approaches. For example, if an investigator decided to select 20 SNPs (2nd row in Table 3) in stage one using the existing ranked p-value approach, marker M_1 would be selected for follow-up in 579 out of 1,000 replicates, and marker M_2 would be selected in 909 out of 1,000 replicates; power to select marker M_1 would be 57.9%, and power to select marker M_2 would be 90.9%. However, using our hybrid approach of selecting top SNPs (10 SNPs based on ranked p values and 10 based on ranked ORs) resulted in powers of 89.9% and 87.0% for selecting markers M_1 and M_2, respectively. Compared with the existing ranked p-value approach, our proposed hybrid approach was more powerful for selecting marker M_1, without losing much power for selecting M_2, in stage one. Considering the above example, power gain for selecting marker M_1 was 32.0%, whereas power loss for selecting marker M_2 was only 3.9%.

Even though our approach increased power to select marker M_1 for follow-up in stage two, we wanted to test what percentage of these SNPs could reach a genome-wide threshold for declaring significance. Therefore, for each replicate selected in stage one, we simulated independently another 1,000 cases and 1,000 controls using the same simulation approach and the same parameters. We then employed a joint analysis using fixed indicators for stages one and two (see Methods section). The power comparisons and medians of ORs for the joint analysis are shown in Table 4. The results are arranged into two panels according to use of the existing p-value-based ranking approach and use of our proposed hybrid ranking approach. The p values and ORs were estimated based on the joint analysis from both stages.

As expected, all of the p values from the joint analysis were much more significant than those from stage one. We used a genome-wide threshold p value of 1.7×10^{-7} [27] for declaring the significance of the SNPs. Our proposed hybrid approach for GWA study gained considerable power for marker M_1 over that of the existing ranked p-value approach. For instance, when the number of selections in stage one was 20, the observed powers were 56.3% and 90.9% for selecting markers M_1 and M_2, respectively, using the existing ranked p-value approach; when our proposed hybrid approach was used, however, the observed powers were 80.0% and 87.0% for selecting M_1 and M_2, respectively. Even when 40 SNPs were selected for follow-up, we still observed a 20.6% increase in power for selecting marker M_1 with loss of about 2% power for selecting marker M_2.

Table 3 Power to Select Top Single-Nucleotide Polymorphisms in Stage One for Replication Based on 1,000 Replicates

Number of Selections	p-Value-Based Ranking Approach		Hybrid Ranking Approach	
	M_1	M_2	M_1	M_2
10	52.2%	87.0%	88.9%	84.2%
20	57.9%	90.9%	89.9%	87.0%
30	61.1%	92.4%	90.4%	90.4%
40	62.6%	93.0%	91.2%	90.9%

OR = odds ratio.

MAF was set at 10% for M_1 and 40% for M_2.

In addition, we also investigated power of selecting both markers loci (M_1 and M_2) simultaneously. The power results are reported in Table 5, which are organized into two panels with respect to stage one and joint analysis. In the joint analysis, the powers for selecting both markers were decided by the genome-wide significance as before. The results in Table 5 showed that the hybrid ranking approach has higher power to select both markers M_1 and M_2. For example, if 20 SNPs were selected in stage one, the observed powers for selecting both markers were 52.3% and 78.2%, respectively, for using the p-value-based approach and the hybrid approach. In joint analysis, the observed powers for selecting both markers were 50.5% and 70.4% for the two different ranking approaches, respectively, and the proposed hybrid approach gained 19.9% increase in power.

Type I error estimate

Based on 10,000,000 replicates for each marker locus, we first estimated the percentages of replicates of unassociated SNPs selected for follow-up in stage two. If 10 SNPs were selected in stage one, using the hybrid approach, marker M_1 would be selected for follow-up in 4,365 of 10,000,000 replicates, and marker M_2 would be selected in 58 of 10,000,000 replicates. When the standard ranked p-value approach was applied, marker M_1 would be selected for follow-up in 143 of 10,000,000 replicates and marker M_2 would be selected in 111 of 10,000,000 replicates. If 20, 30, and 40 SNPs were selected in stage one, using the hybrid approach, marker M_1 would be selected for follow-up in 5,823, 6,736, and 7,803 replicates, respectively, and marker M_2 would be selected in 111, 227 and 250 replicates, respectively. Using the standard ranked p-value-based approach, marker M_1 would be selected for follow-up in 292, 401 and 511 replicates, respectively, and marker M_2 would be selected in 250, 362 and 468 replicates, respectively. However, importantly, after performing the joint analysis, we found that none of the replicates that have been moved forward to the second stage could satisfy the genome-wide threshold p value of 1.7×10^{-7} for

Table 4 Median Odds Ratios and Powers in Joint Analysis (Stage One and Stage Two Combined)

Number of Selections		Joint Analysis			
		p-Value-Based Ranking Approach		Hybrid Ranking Approach	
		M_1	M_2	M_1	M_2
10	OR	1.7387	1.4467	1.6866	1.4447
	Power at 1.7 × 10^{-7}	51.2%	87.0%	79.4%	84.2%
20	OR	1.7342	1.4418	1.6787	1.4467
	Power at 1.7 × 10^{-7}	56.3%	90.9%	80.0%	87.0%
30	OR	1.7343	1.4363	1.6801	1.4398
	Power at 1.7 × 10^{-7}	58.9%	92.3%	79.7%	90.4%
40	OR	1.7224	1.4379	1.6761	1.4418
	Power at 1.7 × 10^{-7}	60.6%	93.0%	81.2%	90.9%

OR = odds ratio.

declaring significance, using either our approach or the standard approach. If we increased the threshold p value for declaring significance to 10^{-5}, one replicate of marker M_1 would be significant on the basis of both the hybrid and ranked p-value approaches. Therefore, we can conclude that the type I error rates of the proposed hybrid approach were well controlled at the end of the experiment (stage two).

Discussion

In this paper, we proposed a hybrid approach to rank and select SNPs in stage one of a GWA study – ORs are considered in addition to p values. The results from the simulation study show that the hybrid approach has an increased power for identifying the less-common genetic variants. Meanwhile, we concluded through simulation studies that the type I error rates of the hybrid approach are controlled at the end of the experiment.

In our study, we selected half of the candidate SNPs on the basis of ranked p values and the other half on the basis of ranked ORs. In reality, one could select the same or different numbers of candidate SNPs based on ranked p values and ranked ORs. For example, one could select the ranked-p-value-based SNPs using the common GWA study threshold (e.g., p value < 10^{-5}) but select the ranked-OR-based SNPs using a less significant threshold (e.g., p value < 10^{-3}) for follow-up in stage two. It should be noted that while ranking SNPs based on ORs, both risk and protective effect SNPs should be considered. In our study, we did not have knowledge about whether extremely large or extremely small effect size is more important; therefore, we reversed ORs corresponding to protective effect SNPs before ranking all the ORs (for ranking purposes only). In our simulation studies, we considered a range of MAFs of 10%~40%. A GWA study with 1,000 cases and 1,000 controls has adequate power to detect the SNPs with an MAF of 10% or higher. To detect variants with smaller MAFs, such as less than 5%, a larger sample size will be needed. The impact of the true OR, average LD and MAF on the power of a GWA study to detect susceptibility SNP markers is discussed in Park et al. [28]. Furthermore, these authors also provide sample sizes required for GWA studies to identify associations.

One may argue that using the ranked p values within the rare variants and ranked p values within the common variants to select candidate SNPs would lead to similar results. The obvious deficiency of this approach is the difficulty in defining a boundary for separating rare and common variants. Furthermore, even if one could define a threshold for rare and common variants (say 10%), among the set of rare (or common) variants, there will still be relatively rare variants and relatively

Table 5 Power to Select Both SNPs (M_1 and M_2) in Stage One and in Joint Analysis (Stage One and Stage Two Combined)

Number of Selection	Stage One		Joint Analysis (Power at 1.7 × 10^{-7})	
	p-Value-Based Ranking Approach	Hybrid Ranking Approach	p-Value-Based Ranking Approach	Hybrid Ranking Approach
10	45.0%	75.0%	43.6%	66.0%
20	52.3%	78.2%	50.5%	70.4%
30	56.0%	81.7%	54.0%	71.6%
40	58.1%	82.8%	55.9%	72.8%

common variants. Therefore, when only selecting top SNPs based on the ranked p values within rare variants, one may still be more likely to select the SNPs that are relatively common in the rare variants set, but miss the SNPs that are relatively rare in the rare variants set. For example, if defining SNPs with MAFs < 10% as rare variants and using the ranked p values within rare variants set, one might not be able to capture the SNP with MAF = 3% but OR = 2.5 in stage one (but select the SNP with MAF = 9% but OR = 1.5, for example), even though the OR value represents a very significant association between this SNP and the disease of interest. We had considered such an approach to our simulated data, where we defined SNPs with MAF = 10% as rare variants and SNPs with MAF = 40% as common variants. The powers for selecting both rare and common variants did not increase, however, and were almost identical to those obtained using the standard ranked-p-value approach (data not shown).

It is well known that large ORs usually correspond to rare variants, and significant p values correspond to common variants, regardless of whether or not the variants are associated with the disease of interest. In our simulation study, we actually found that this is the case, and this can further verify the rationale of our hybrid approach. Figure 1 shows the relationship between p values and ORs based on 1,500,000 null SNPs (no association with disease), where 500,000 SNPs were simulated using MAF = 6.45%, 500,000 SNPs were simulated using MAF = 10%, and 500,000 SNPs were simulated using MAF = 40%. We observed that, at the same p value levels, SNPs with MAF = 6.45% always have the largest ORs; while at the same OR levels, SNPs with MAF = 40% always have the most significant p values. In addition to the simulation study, we also investigated the expected test statistics of association tests, as well as the corresponding p values, under the alternative hypothesis. Figures 2 and 3 show the graphs of the expected test statistics and the p values, respectively, as the functions of ORs and MAFs, given a fixed sample size. The details of the expected test statistics and the assumptions are described in the Appendix. We observed that, for this specific example, when MAF = 0.4 and OR = 1.5, the expected test statistic was 4.1733 and the corresponding negative logarithm to base 10 of p value was 4.5044; when MAF = 0.1 and OR = 1.8, the expected test statistic was 2.6813 and the corresponding negative logarithm to base 10 of p value was 2.1312. Therefore, using the ranked p values approach with a threshold of 10^{-4}, the SNP with OR = 1.8 and MAF = 0.1 would not be included in the further analysis, although it has a higher effect size; however, with the use of the hybrid ranking approach, this SNP would have a greater chance of being included in the next stage.

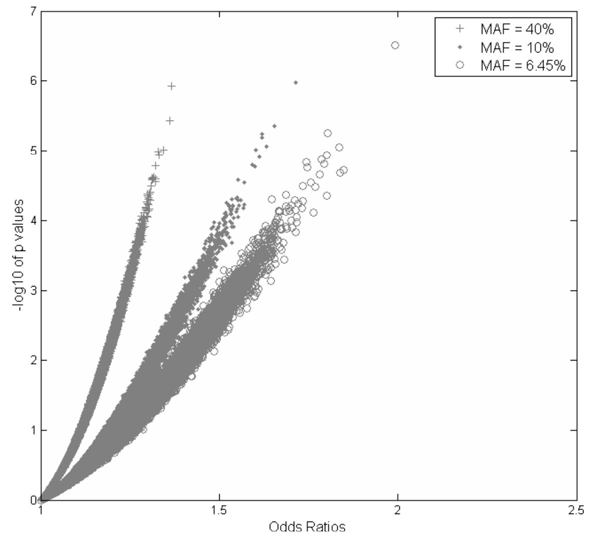

Figure 1 Odds Ratios versus p Values Based on 1,500,000 Simulated Non-Disease-Related SNPs.

Currently, ranked p value is used as a criterion for selecting common variants, but from Figures 1, 2, and 3, we can conclude that ranked OR should be used as an additional criterion for selecting less-common variants. These two approaches are complementary, and therefore, a hybrid approach using both ranked p value and ranked OR should be more powerful for selecting rare variants. It should be noted that the type I error is controlled in the joint analysis because the SNPs selected in stage one with the use of our hybrid ranking approach are not final, and they have to meet the GWA significance (1.7×10^{-7}) in joint analysis. The results from the simulation studies confirmed this statement.

Conclusions

We proposed a hybrid approach for selecting candidate SNPs from the discovery stage of GWA study, based on both p values and ORs, and conducted a simulation study to demonstrate the performance of our approach. The power comparison results show that our hybrid ranking approach is more powerful than the existing ranked p value approach for identifying relatively less-common SNPs. Therefore, GWA studies should consider including SNPs based not only on ranked p values but also on ranked ORs. Furthermore, with the rapid development of sequencing techniques, much denser SNP chips with more low-MAF SNPs may be available in the near future. With these improved technologies, our hybrid ranking approach for selecting top SNPs offers a promising direction for future research in GWA studies.

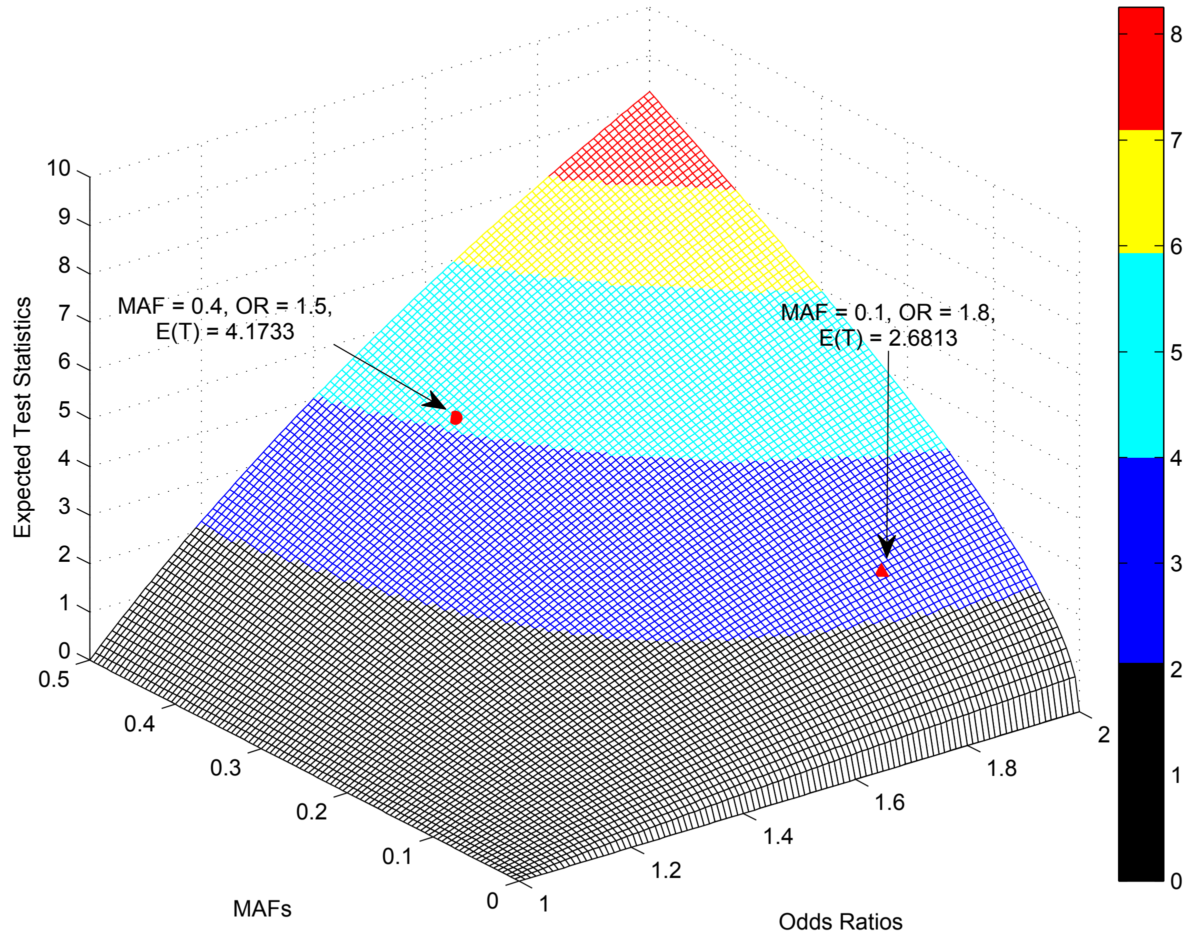

Figure 2 The Expected Test Statistics E(*T*) under the Alternative Hypothesis as a Function of Minor Allele Frequencies (MAFs, *ρ*) and Odds Ratios (ORs, exp (*β*)) given a Fixed Sample Size *n* = 2000. The expected test statistics under the alternative hypothesis were estimated using Equation (1), where we used *σ* = 4.6, *n* = 2000, and assumed an additive underlying genetic model. We observed that, when MAF = 0.4 and OR = 1.5, E(*T*) = 4.1733; when MAF = 0.1 and OR = 1.8, E(*T*) = 2.6813.

Appendix

Without loss of generality, we considered a simple linear regression model:

$$y_i = \beta x_i, \quad i = 1, \ldots n.$$

The hypothesis we are interested in testing is H_0: β = 0 versus H_1: $\beta \neq 0$. The estimate $\hat{\beta}$ has the estimate variance $s^2 / \sum x_i^2$, where s is its estimated standard deviation. Therefore, we can obtain the test statistics T as follows:

$$T = \hat{\beta} \times \sqrt{\sum x_i^{\,2}} \,/\, s.$$

Under the null hypothesis H_0, T follows a t-distribution with $\upsilon = n$-2 degrees of freedom, and if $\beta \neq 0$, T has a noncentral t-distribution with $\upsilon = n$-2 degrees of freedom and non-centrality parameter $\mu = \beta\sqrt{\sum x_i^2} \,/\, \sigma$ [29]. The expected value of the non-central t-distribution test statistics is

$$E(T) = \mu\sqrt{\upsilon / 2}\,\frac{\Gamma((\upsilon - 1)/2)}{\Gamma(\upsilon / 2)}, \quad \upsilon > 1, \qquad (1)$$

Where $\Gamma(\cdot)$ is the gamma function. In order to draw the graphs of the expected test statistics and the corresponding p values, we used σ = 4.6 and n = 2000, and assumed an additive genetic model. Therefore, in the above equation, the degree of freedom is $\upsilon = n$-2 = 1998, and the non-centrality parameter is $\mu = \log(OR)\sqrt{n \times (2(\rho^2 + \rho))} \,/\, 4.6$, where ρ is the MAF, and OR is odds ratio.

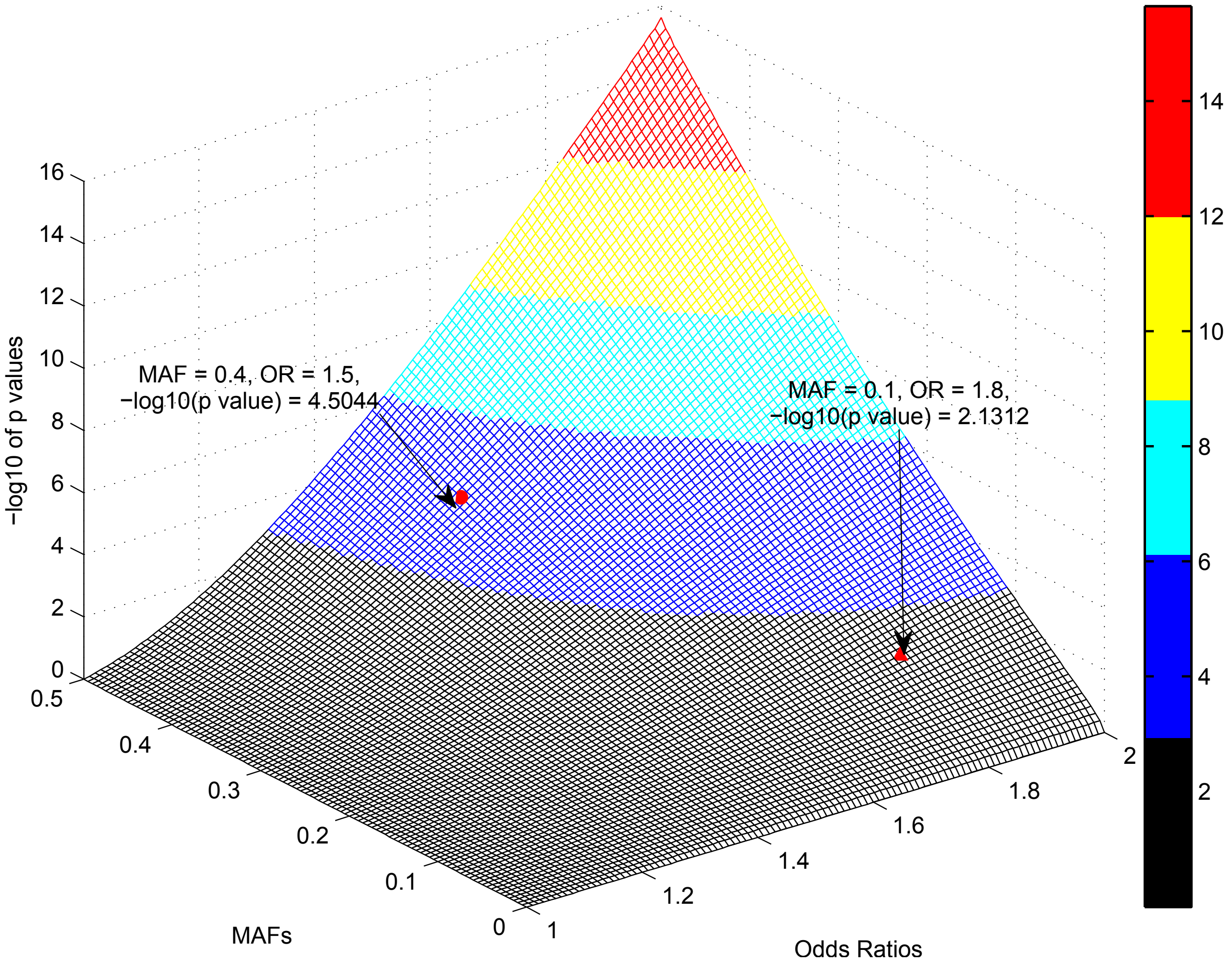

Figure 3 The Negative Logarithm to Base 10 of p Values under the Alternative Hypothesis as a Function of Minor Allele Frequencies (MAFs, *ρ*) and Odds Ratios (ORs, exp (*β*)) given a Fixed Sample Size *n* = 2000. The negative logarithm to base 10 of p values is corresponding to the expected test statistics in Figure 2 under the alternative hypothesis. We observed that, when MAF = 0.4 and OR = 1.5, -log10 of p value = 4.5044; when MAF = 0.1 and OR = 1.8, -log10 of p value = 2.1312.

Acknowledgements
We appreciate the two anonymous reviewers for their insightful and constructive comments. This work was supported by National Institutes of Health [grant number 1R01CA131324].

Authors' contributions

JW performed the simulation studies and drafted the manuscript. SS conceptualized the study, provided advice and revised the manuscript. All authors read and approved the final manuscript.

Received: 9 July 2010 Accepted: 6 January 2011
Published: 6 January 2011

References

1. Marchini J, Donnelly P, Cardon LR: **Genome-wide strategies for detecting multiple loci that influence complex diseases.** *Nat Genet* 2005, **37**:413-417.
2. Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Ballinger D, Daly M, Donnelly P, Faraone SV, Frazer K, Gabriel S, *et al*: **New models of collaboration in genome-wide association studies: the Genetic Association Information Network.** *Nat Genet* 2007, **39**:1045-1051.
3. Pearson TA, Manolio TA: **How to interpret a genome-wide association study.** *JAMA* 2008, **299**:1335-1344.
4. Manolio TA, Brooks LD, Collins FS: **A HapMap harvest of insights into the genetics of common disease.** *J Clin Invest* 2008, **118**:1590-1605.
5. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, *et al*: **Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1.** *Nat Genet* 2008, **40**:616-622.
6. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, *et al*: **A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25.** *Nature* 2008, **452**:633-637.
7. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, *et al*: **A variant associated with nicotine dependence, lung cancer and peripheral arterial disease.** *Nature* 2008, **452**:638-642.
8. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J, *et al*: **Multiple newly identified loci associated with prostate cancer susceptibility.** *Nat Genet* 2008, **40**:316-321.

9. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, *et al*: **Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24.** *Nat Genet* 2007, **39**:631-637.
10. Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, Yu K, Chatterjee N, Welch R, Hutchinson A, *et al*: **Multiple loci identified in a genome-wide association study of prostate cancer.** *Nat Genet* 2008, **40**:310-315.
11. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, *et al*: **Genome-wide association study of prostate cancer identifies a second risk locus at 8q24.** *Nat Genet* 2007, **39**:645-649.
12. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, Jakobsdottir M, Bergthorsson JT, Gudmundsson J, Aben KK, *et al*: **Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer.** *Nat Genet* 2008, **40**:703-706.
13. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, *et al*: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.** *Nat Genet* 2007, **39**:870-874.
14. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, *et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci.** *Nature* 2007, **447**:1087-1093.
15. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, Friedman E, Narod S, Olshen AB, Gregersen P, *et al*: **Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33.** *Proc Natl Acad Sci USA* 2008, **105**:4340-4345.
16. Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B, Simon M, Marie Y, Boisselier B, Delattre JY, *et al*: **Genome-wide association study identifies five susceptibility loci for glioma.** *Nat Genet* 2009, **41**:899-904.
17. Wrensch M, Jenkins RB, Chang JS, Yeh RF, Xiao Y, Decker PA, Ballman KV, Berger M, Buckner JC, Chang S, *et al*: **Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility.** *Nat Genet* 2009, **41**:905-908.
18. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, *et al*: **Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.** *Science* 2007, **316**:1336-1341.
19. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, *et al*: **A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.** *Science* 2007, **316**:1341-1345.
20. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, *et al*: **Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.** *Science* 2007, **316**:1331-1336.
21. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, *et al*: **A genome-wide association study identifies novel risk loci for type 2 diabetes.** *Nature* 2007, **445**:881-885.
22. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, Hirota Y, Mori H, Jonsson A, Sato Y, *et al*: **Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus.** *Nat Genet* 2008, **40**:1092-1097.
23. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, Ng DP, Holmkvist J, Borch-Johnsen K, Jorgensen T, *et al*: **SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations.** *Nat Genet* 2008, **40**:1098-1102.
24. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, *et al*: **A variant in CDKAL1 influences insulin response and risk of type 2 diabetes.** *Nat Genet* 2007, **39**:770-775.
25. Dudek SM, Motsinger AA, Velez DR, Williams SM, Ritchie MD: **Data simulation software for whole-genome association and other studies in human genetics.** *Pac Symp Biocomput* 2006, 499-510.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**:559-575.
27. Skol AD, Scott LJ, Abecasis GR, Boehnke M: **Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies.** *Nat Genet* 2006, **38**:209-213.
28. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N: **Estimation of effect size distribution from genome-wide association studies and implications for future discoveries.** *Nat Genet* 2010, **42**:570-575.
29. Bickel PJ, Doksum KA: *Mathematical Statistics: Basic Ideas and Selected Topics* Oakland, California: Holden-Day, Incorporated; 1977.

doi:10.1186/1471-2156-12-3
Cite this article as: Wang and Shete: **A powerful hybrid approach to select top single-nucleotide polymorphisms for genome-wide association study.** *BMC Genetics* 2011 **12**:3.