Statistics and Its Interface Volume 2 (2009) 145–151

Robust genome-wide scans with genetic model selection using case-control design*

Gang Zheng[†], Jungnam Joo, Xin Tian, Colin O. Wu, Jing-Ping Lin, Mario Stylianou, Myron A. Waclawiw, and Nancy L. Geller

In a genome-wide association study with more than 100,000 (100K) to 1 million single nucleotide polymorphisms (SNPs), the first step is usually a genome-wide scan to identify candidate chromosome regions for further analyses. The goal of the genome-wide scan is to rank all the SNPs based on their association tests or p-values and select the top SNPs. A good ranking procedure ranks the SNPs with true associations as near to the top as possible. This enhances the probability of selecting at least one SNP with a true association. However, if the disease-associated SNPs have moderate genetic effects, the probability that a large number of null SNPs will have extremely small p-values (or large test statistics) is high when screening more than 300K SNPs. Therefore, when selecting a small fraction of top SNPs (usually less than 5%), the probability of selecting at least one SNP with a true association is usually less than 80% unless the sample size is large. Robust statistics have been proposed to rank all the SNPs (e.g., MAX3 and MIN2). In this article we consider genome-wide scans with a genetic model selection and compare this proposed method to the existing approaches. Results from simulation studies are presented.

AMS 2000 subject classifications: Primary 62G35, 62G10; secondary 62G30, 62P10.

Keywords and phrases: Case-control design, Efficiency robustness, Genetic model selection, Genome-wide studies, MAX.

1. INTRODUCTION

In the analysis of genome-wide association studies using 100K to 1 million SNPs, a genome-wide scan (GWS) is the first step, in which an association test is applied to each SNP [13, 20, 23]. Then all the SNPs are ranked either based on their p-values or the values of the association test. In the following, the top SNPs are referred to as those with the largest test statistics (or smallest p-values). After all the ranks are obtained, a small fraction of the top SNPs (often less than 5%) will be chosen for a more powerful analysis in the next step, e.g., haplotype analysis, multi-marker analysis, or re-sequencing [10, 16, 18, 22]. In Klein et al. [13], only the top two SNPs were selected from 100K SNPs. Sladek et al. [20] used the genome-wide threshold level 0.0001 for their GWS with 300K SNPs and selected 59 SNPs for further scrutiny.

In GWS, SNPs with TAs are not always ranked near the top. Factors that affect the ranks of the SNPs with TAs include, but are not limited to, the total number of SNPs, the total number of SNPs with TAs, the genetic effect of the SNPs with TAs, the sample size, the power of test statistics, and linkage disequilibrium between the SNPs and the disease loci [4, 24]. Among these factors, only the total number of SNPs and the power of test statistics can be improved during the GWS, because the other factors are determined either in the design stage or by the underlying science or biology, which are fixed during the analysis stage.

Zaykin and Zhivotovsky [24] and Gail et al. [4] demonstrated by simulation studies that, when the number of SNPs was reduced and/or a more powerful test statistic was used, the probabilities selecting SNPs with TAs were improved. Although a two-stage design may be used to reduce the number of SNPs [1, 19], how to effectively reduce the SNPs to be tested or scanned has not been rigorously studied. On the other hand, it is known that there is no uniform most powerful test for association using case-control samples unless the mode of inheritance (genetic model) of the disease is known. For most common and complicated diseases, the true genetic model is unknown. It could be one of four common genetic models considered in the literature: the recessive, additive, multiplicative or dominant models or none of them. Given that there is no optimal test statistic for GWS, robust test statistics are desirable.

In this article, we first discuss the difference between testing SNPs and ranking SNPs. Then we review some existing test statistics for GWS and compare them to a recent proposed robust test statistic based on genetic model selection. We conduct simulation studies for GWS using 300K SNPs containing 6 SNPs with TAs to compare the ranking methods.

*This article is written in honor of Prof. Joseph Gastwirth whose pioneer work in robust procedures has important impact on our research in genetic association studies.

[†]Corresponding author.

2. TESTING VERSUS RANKING: USING TEST STATISTICS OR P-VALUES

To conduct a genome-wide association study, a test statistic is first applied to each SNP to test for association. This test statistic is denoted as Z. The p-value of Z is obtained for each SNP and compared to the prespecified significance level. The SNPs with p-values smaller than the significance level would be associated with the disease and retained for further analyses. This approach is referred to as a testing approach. An alternative one is a ranking approach. In the ranking approach, the SNPs can be ranked by their p-values and a prespecified proportion of the top ranked SNPs is retained for further analyses regardless of whether or not their p-values are significant.

Suppose Z asymptotically follows a standard normal distribution $N(0,1)$ under the null hypothesis. Denote the value of Z for the ith SNP as Z_i with p-value P_i for $i = 1, \ldots, M$, where M is the total number of SNPs, e.g., $M = 300,000$. Then, M SNPs can be ranked either by Z_i^2, $i = 1, \ldots, M$ or by P_i, $i = 1, \ldots, M$. Both ranking procedures result in the same ranks for all M SNPs. When Z does not asymptotically follow $N(0,1)$, the p-value may not be easily obtained. For example, the asymptotic distributions of the robust statistics that we will discuss later are complex and may also depend on the minor allele frequency of each SNP. Therefore, the test statistic for one SNP may not be comparable to that of another, even though the difference may be small [15]. In this situation, only p-values are comparable among the SNPs. However, finding the p-values for these robust tests is computationally intensive, in particular for genome-wide association studies with 300K or more SNPs. In this case, ranking by statistics and ranking by p-values may not result in the same ranks, but ranking SNPs by the values of the test statistic is much easier than ranking them by the p-values.

In GWS, the goal is not to claim if a specific SNP is associated with the disease. Instead, it aims to obtain the rank of one SNP relative to the others through competing for the top ranks. Therefore, how small the p-value is does not provide additional information in GWS when all SNPs are ranked. Hence, we propose to rank SNPs by the values of test statistics. Li et al. [15] compared the rankings by the MAX statistic (see the description of MAX below) and its p-value, and found, although there were discrepancies, the difference was minor.

3. EXISTING GWS: FROM SINGLE MODEL TESTS TO ROBUST TESTS

3.1 Notation and genetic models

The common test statistics for GWS include: the allele-based test [13], the Cochran-Armitage trend test (CATT) [17, 23], Pearson's Chi-square test with 2 degrees of freedom [23], and a robust test, MAX [20]. Although the allele-based test is different from the CATT for single marker analysis even when Hardy-Weinberg equilibrium (HWE) holds in the population [29], the performance of GWS using the allelic test and the CATT should be similar. Moreover, the allelic test is not valid when HWE does not hold [9, 17]. Hence we consider the CATT here rather than the allelic test.

Table 1. Genotype data for a single SNP

	Genotypes			
	AA	AB	BB	Total
Case	r_0	r_1	r_2	r
Control	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

Consider a single SNP with two alleles A and B. The genotypes for r cases and s controls are presented in Table 1 where total numbers of genotypes $G_0 = AA$, $G_1 = AB$ and $G_2 = BB$ are n_0, n_1 and n_2, respectively. Let the population frequencies for the alleles be $\Pr(B) = p$ and $\Pr(A) = q = 1 - p$. Assuming HWE holds in the population, the population frequencies of the three genotypes are $g_0 = \Pr(G_0) = q^2$, $g_1 = \Pr(G_1) = 2pq$ and $g_2 = \Pr(G_2) = p^2$.

Denote the penetrance by $f_j = \Pr(\text{case}|G_j)$ for $j = 0, 1, 2$. Then the recessive, additive, multiplicative and dominant models refer to $f_1 = f_0$, $f_1 = (f_0 + f_2)/2$, $f_1 = (f_0 f_2)^{1/2}$, and $f_1 = f_2$, respectively. The assumption for the above genetic model definitions is that allele B is the risk allele. This assumption is required for only one of the approaches we will discuss later. For the other approaches, we consider two-sided tests. Denote the disease prevalence by $k = \Pr(\text{case})$. The genotype counts (r_0, r_1, r_2) in cases and (s_0, s_1, s_2) in controls are random samples from the multinomial distributions $Mul(r; p_0, p_1, p_2)$ and $Mul(s; q_0, q_1, q_2)$, respectively, where $r = r_0 + r_1 + r_2$, $s = s_0 + s_1 + s_2$, $p_j = g_j f_j / k$ and $q_j = g_j(1 - f_j)/(1 - k)$. Under the null hypothesis of no association, $H_0 : p_j = q_j$ for $j = 0, 1, 2$, i.e. $H_0 : f_0 = f_1 = f_2 = k$. Assuming $f_0 > 0$, denote the genotype relative risks (GRRs) by $\lambda_i = f_i / f_0$ for $i = 1, 2$.

3.2 Existing test statistics

Using the case-control data in Table 1, the CATT can be written as

$$Z_{\text{CATT}}(x) = \frac{n^{1/2}(n \sum_{j=0}^{2} x_j r_j - r \sum_{j=0}^{2} x_j n_j)}{[rs\{n(n_1 + 4n_2) - (n_1 + 2n_2)^2\}]^{1/2}},$$

where $(x_0, x_1, x_2) = (0, x, 1)$ is a set of the increasing scores prespecified for the three genotypes, where $x \in [0, 1]$. The choice of x depends on the underlying genetic model, and $x = 0$, $1/2$, and 1 for the recessive, additive (or multiplicative), and dominant models [3, 17, 25]. Under H_0, the CATT asymptotically follows $N(0,1)$, denoted by

$Z_{\text{CATT}}(x) \sim N(0,1)$. When the true genetic model is unknown, $Z_{\text{CATT}}(1/2)$ is often used because it is most robust among the three CATTs [3, 17]. We focus on the CATT with $x = 1/2$ ($Z_{\text{CATT}}(1/2)$).

Pearson's Chi-square test can be written as

$$T_{\chi_2^2} = \sum_{j=0}^{2} \frac{(r_j - n_j r/n)^2}{n_j r/n} + \sum_{j=0}^{2} \frac{(s_j - n_j s/n)^2}{n_j s/n},$$

which has an asymptotic Chi-squared distribution with 2 degrees of freedom.

It is known that when the underlying genetic model is additive or dominant, the CATT (with $x = 1/2$) is either optimal or robust. However the CATT is not robust if the genetic model is recessive. Pearson's test, $T_{\chi_2^2}$, on the other hand, does not depend on the genetic model. Therefore, it is robust across the four genetic models, although it is less powerful than the CATT under the additive and dominant models.

A more robust test, based on the efficiency robust theory of Gastwirth [5, 6], is MAX (or MAX3 in some literature), which takes the maximum of the three CATTs across the three genetic models, given by

$$\text{MAX} = \max(|Z_{\text{CATT}}(0)|, |Z_{\text{CATT}}(1/2)|, |Z_{\text{CATT}}(1)|).$$

The asymptotic null distribution of MAX can be simulated [3, 20] or approximated [14] so that its p-value can be obtained. In general, the p-value of MAX depends on the minor allele frequency of each SNP. Thus, in principle, the asymptotic distribution of MAX is different from SNP to SNP. Zheng et al. [26] showed by simulation studies that MAX is always more powerful than $T_{\chi_2^2}$ when the genetic model is restricted to the four common models. See also Zheng et al. [30] for the relationship among the CATT, Pearson's and MAX. Implementing MAX, however, is difficult for GWS due to intensive computation of the p-value of MAX [15]. Zheng et al. [27] and Li et al. [15] proposed using the MAX statistic directly to rank the SNPs rather than using the p-values. Simulation results from Li et al. [15] showed that ranking SNPs with TAs among 300K SNPs resulted in comparable ranks by using MAX (MAX-rank) and using p-values (P-rank).

In addition to the above statistics, we also consider a recent robust test used by Wellcome Trust Case-Control Consortium (WTCCC) [23], which is referred to as MIN2 in Joo et al. [11]. Let the p-values of Pearson's test and the CATT optimal for the additive model denote by $P_{\chi_2^2}$ and P_{CATT}, respectively. Then, MIN2 is written as

$$\text{MIN2} = \min\left(P_{\chi_2^2}, P_{\text{CATT}}\right),$$

which takes the minimum p-values of the two tests. Note that MIN2 itself is not a valid p-value. Joo et al. [11] derived an asymptotic null distribution for MIN2 and showed that it is independent of the allele frequency of each SNP. Further, they showed that the statistic MIN2 is increasing in the p-value of MIN2. Therefore, using the values of MIN2 to rank SNPs is equivalent to using the p-values of MIN2 to rank the SNPs. The p-values of Pearson's test and the CATT are easily obtained, so MIN2 is easy to apply in GWS. The results from the simulation studies of candidate-genes in Joo et al. [11] indicate that MIN2 often combines the strength of both Pearson's test and the CATT, and has similar power performance with MAX. But previous results of MIN2 are based on single marker analysis for a given genetic model. Here we will examine its performance in GWS by competing ranks among the SNPs with different genetic models.

4. GENOME-WIDE SCAN WITH GENETIC MODEL SELECTION

Genetic model selection (GMS) is another recently proposed method for testing association using case-control samples [28]. GMS selects an underlying genetic model based on the sign and value of Hardy-Weinberg disequilibrium (HWD) coefficient between cases and controls [21] followed by testing association using the CATT with the selected model.

Define the HWD coefficient in cases and controls by $\Delta_p = p_2 - (p_2 + p_1/2)^2$ and $\Delta_q = q_2 - (q_2 + q_1/2)^2$, respectively. Denote the HWD coefficient in the population by $\Delta = \Pr(BB) - \{\Pr(BB) + \Pr(AB)/2\}^2$. Under HWE in the population, $\Delta = 0$. It is noted by Song and Elston [21] that the null hypothesis of association can be tested by $H_0 : \Delta_p = \Delta_q$ against the alternative hypothesis $H_1 : \Delta_p \neq \Delta_q$ under HWE ($\Delta = 0$). The test statistic using the difference of $\hat{\Delta}_p$ and $\hat{\Delta}_q$, referred to as the HWD trend test (HWDTT), is given by [21]

$$Z_{\text{HWDTT}} = \frac{\widehat{\Delta}_p - \widehat{\Delta}_q}{\left[\{1 - n_2/n - n_1/(2n)\}\{n_2/n + n_1/(2n)\}\right]^{1/2}},$$

where $\widehat{p}_j = r_j/r$ and $\widehat{q}_j = s_j/s$ for $j = 0, 1, 2$. Under H_0, $Z_{\text{HWDTT}} \sim N(0,1)$. The test Z_{HWDTT}, however, cannot be used to test for association under the multiplicative model, because its expected value under H_1 with the multiplicative model is zero, and it has lower power under the additive model.

Zheng and Ng [28] also noticed that Z_{HWDTT} can be used to indicate the recessive and dominant models. That is, if B is the risk allele, then $Z_{\text{HWDTT}} \gg 0$ indicates the recessive model while $Z_{\text{HWDTT}} \ll 0$ indicates the dominant model. Then they proposed a GMS procedure defined by $Z_{\text{GMS}} = Z_{\text{CATT}}(0)$ if $Z_{\text{HWDTT}} > 1.64$, $Z_{\text{GMS}} = Z_{\text{CATT}}(1)$ if $Z_{\text{HWDTT}} < -1.64$, and $Z_{\text{GMS}} = Z_{\text{CATT}}(1/2)$, otherwise. Zheng and Ng [28] provided formulas to calculate the p-value of the GMS which involves integrations. They also conducted simulation studies and showed that, when the risk allele was known (also the minor allele), the GMS is always more powerful than MAX in candidate-gene analysis.

In our simulation studies that will be described later, we determine the risk allele first before applying the GMS. By doing this, we do not need to know the risk allele nor to assume the minor allele is the risk allele, which may not be true in practice. This is because when the two alleles are switched the sign and value of Z_{HWDTT} does not change. However, the sign of the CATT is changed after switching the alleles while the absolute value of the statistic is the same. The impact of examining the risk allele is that it could inflate the Type I error rate slightly (Joo et al. [12]). But, as we discussed before, our goal in GWS is to rank SNPs rather than conclude whether or not the SNPs are significantly associated with the disease. For the same reason, instead of using the p-value of GMS to rank SNPs, we rank SNPs directly by Z_{GMS}. Hence, it is very simple to apply GMS to GWS.

5. SIMULATION STUDIES

5.1 Design of simulation studies

We conduct simulation studies to compare GWS with the CATT (using $x = 1/2$), χ^2_2, MAX, MIN2, and GMS. We assume a genome-wide association study using 300K SNPs among which 6 SNPs are associated with the disease. Zaykin and Zhivotovsky conducted the GWS simulation with joint effects among the SNPs with TAs and showed the impact of multi-marker effects due to linkage disequilibrium among the SNPs was small. Hence, for simplicity, we also assume that these 6 SNPs are independent.

Three patterns are considered. In pattern 1, the six SNPs with TAs contain 2 SNPs with the recessive model, 2 SNPs with the dominant model, 1 SNP with the additive model and 1 SNP with the multiplicative model. In pattern 2, there are 2 SNPs with the additive model and 2 SNPs with the multiplicative model, but 1 SNP with the recessive model and 1 SNP with the dominant model. Pattern 3 changes the two SNPs in Pattern 1 with the recessive and dominant models respectively to the overdominant models ($\lambda_1 < \lambda_0$ or $\lambda_1 > \lambda_2$ depending on which allele is the risk allele). Table 2 presents the distributions of the 6 SNPs with TAs along with their minor allele frequencies (MAFs). The MAFs reported in Table 2 were also used by Li et al. [15]. These MAFs were taken from the genome-wide association studies of WTCCC [23], except that the order of MAFs for the four genetic models is different from that in Li et al. [15]. For example, in Li et al. [15], MAF=0.1078 was assigned to the recessive model. Here, in Table 2, MAF=0.1821 or 0.2943 are assigned to the recessive model, so that the rare recessive SNP is not considered. The MAFs for the 299,944 (=300,000-6) null SNPs are independently generated from the uniform distribution (0.1,0.5).

Table 2. The minor allele frequencies (MAFs) of the 6 SNPs with TAs among 300K SNPs that were used by Li et al. [15]. The MAFs of the null SNPs were independently simulated from the uniform distribution (0.1, 0.5)

MAFs	Pattern 1	Pattern 2	Pattern 3
0.1821	Recessive	Recessive	Overdominant
0.2943	Recessive	Additive	Recessive
0.1078	Additive	Additive	Additive
0.4459	Multiplicative	Multiplicative	Multiplicative
0.1620	Dominant	Multiplicative	Dominant
0.1825	Dominant	Dominant	Overdominant

HWE is assumed in the simulation. Genotype counts for cases and controls are generated from the multinomial distributions with prevalence $k = 0.1$, GRR $\lambda_2 = 1.25$ or $\lambda_2 = 1.5$ for all 6 SNPs, and λ_1 is calculated for each of the 6 SNPs with TAs using a genetic model and λ_2. The GRRs for the null SNPs are $\lambda_1 = \lambda_2 = 1$. For the two SNPs with the overdominant model in Pattern 3, $\lambda_1 = 0.85 < 1$ and $\lambda_1 = \lambda_2 + 0.15 > \lambda_2$ are chosen. After the data are generated with 200 replicates, each method is applied to rank the SNPs. The top 5,000 ranked SNPs are considered. The ranks of the 6 SNPs with TAs among the top 5,000 SNPs are recorded. In some cases, when there is no SNP with TAs among the top 5,000 SNPs, we record it as missing.

Several statistics are used to evaluate the ranking method. The first is the probability that at least one of the 6 SNPs with TA is among the top 5,000 SNPs. If this probability is low, it indicates a high chance that the ranking method may not include at least one SNP with TA for further more powerful analysis. Among the 200 replicates, we first estimate the proportion of missing (missing rate). Then the estimate of the probability that at least one of 6 SNPs with TA is among the top 5,000 SNPs is 1 minus the missing rate. Because this probability does not tell how many SNPs with TAs are ranked among the top 5,000 SNPs, we also estimate the average number of SNPs with TAs that are ranked among the top 5,000 SNPs in the 200 replicates. This average number is in the range of 0 (all missing in each replicate) to 6 (no missing in all 200 replicates). A third statistic we consider is the minimum rank of the SNPs with TAs that are also ranked in the top 5,000 SNPs. The smaller the minimum rank, the closer to the top the ranks of the SNPs with TA will be. Among the 200 replicates, both the mean and median of the minimum rank of the 6 SNPs with TA are reported.

5.2 Results

The results from the simulation studies are summarized in Table 3 for Pattern 1, Table 4 for Pattern 2, and Table 5 for Pattern 3.

First, let us compare the probability that at least one SNP with TAs is included in the top 5,000 SNPs. From the three tables, the GMS usually has the largest probability (except when $\lambda_2 = 1.25$ in Tables 4 and 5) while Pearson's test has the smallest probability (except when $\lambda_2 = 1.50$ in Tables 3 and 5).

Table 3. Genotype scans of 300K SNPs containing 6 SNPs with TAs (2 REC, 1 ADD, 1 MUL and 2 DOM) by five methods. Minor allele frequencies are reported in Table 2. Only the top 5,000 ranked SNPs are selected where the following measures are obtained only for the SNPs with TAs among the top 5,000 SNPs: prob = probability at least one SNP with TA; ave. no. of true SNPs = the average number of true SNPs; mean of min ranks = the average of the minimum rank of the SNPs with TAs; and med. of min ranks = the median of the minimum ranks of the SNPs. The results are based on 200 replicates

GRR λ_2	Methods	Prob (%)	Ave. no. of true SNPs	Mean of min ranks	Med. of min ranks
1.25	CATT	92.0	1.79	971	409
	GMS	94.5	1.90	838	381
	MAX	90.5	1.80	909	411
	MIN2	89.5	1.79	934	409
	χ^2_2	86.5	1.69	960	423
1.50	CATT	99.5	2.71	186	24
	GMS	100.0	2.99	178	20
	MAX	99.5	2.83	205	25
	MIN2	100.0	2.78	234	28
	χ^2_2	100.0	2.71	286	42

Table 4. Genotype scans of 300K SNPs containing 6 SNPs with TAs (1 REC, 2 ADD, 2 MUL and 1 DOM) by five methods. The descriptions of the table entries are the same as those in Table 3

GRR λ_2	Methods	Prob (%)	Ave. no. of true SNPs	Mean of min ranks	Med. of min ranks
1.25	CATT	88.0	1.72	897	374
	GMS	87.0	1.79	797	355
	MAX	82.5	1.64	846	396
	MIN2	86.0	1.66	932	483
	χ^2_2	83.0	1.50	1030	563
1.50	CATT	99.0	2.46	349	75
	GMS	99.5	2.61	355	45
	MAX	98.0	2.34	379	59
	MIN2	99.5	2.35	434	62
	χ^2_2	97.0	2.21	485	99

In terms of the number of SNPs with TAs in the top 5,000 SNPs, the GMS continues to have the best performance while Pearson's test still has the worst performance. The GMS always selects the most SNPs with TAs, and MIN2 and MAX select similar numbers of SNPs with TAs. Note that MIN2 selects more SNPs with TAs on average than the CATT does except under Pattern 2 in which 2 SNPs with TAs have the additive model and 2 SNPs with TAs have the dominant model under which the CATT is optimal or quite robust.

Table 5. Genotype scans of 300K SNPs containing 6 SNPs with TAs (1 REC, 1 ADD, 1 MUL, 1 DOM and 2 overdominant) by five methods. The descriptions of the table entries are the same as those in Table 3. For the two SNPs with the overdominant model, $\lambda_1 = 0.85$ and $\lambda_1 = \lambda_2 + 0.15$, respectively

GRR λ_2	Methods	Prob (%)	Ave. no. of true SNPs	Mean of min ranks	Med. of min ranks
1.25	CATT	91.5	1.96	1051	596
	GMS	90.5	1.99	989	585
	MAX	86.5	1.84	1140	682
	MIN2	88.0	1.99	993	616
	χ^2_2	84.0	1.80	1219	752
1.50	CATT	97.5	2.38	432	79
	GMS	98.0	2.69	367	82
	MAX	96.5	2.42	422	98
	MIN2	98.5	2.53	491	100
	χ^2_2	97.5	2.31	560	169

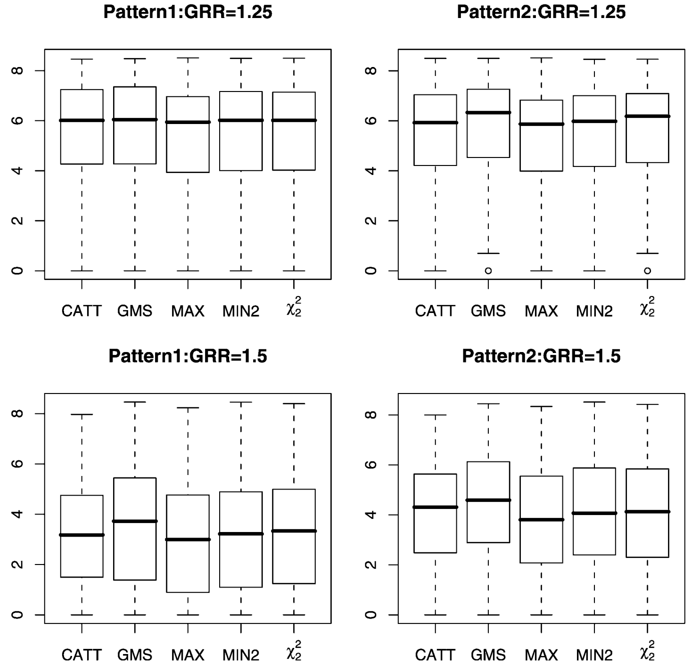

Figure 1. Plots of minimum ranks of the SNPs with TAs that are contained in the top 5,000 SNPs. Pattern 1 with GRR = 1.25 (GRR = 1.50) is on the left column and first (second) row. Pattern 2 with GRR = 1.25 (GRR = 1.50) is on the right column and first (second) row.

Next, we compare the five methods based on the median or mean of the minimum ranks of the SNPs with TAs among the 5,000 SNPs across the replicates. From the last two columns of Tables 3–5, the ranking data are quite skewed. We can conclude that GMS has the best performance while Pearson's test has the worst performance. GMS tends to rank the SNPs with TAs nearer the top than other approaches. Figure 1 presents box plots of the minimum ranks (on a log-scale) among the 200 replicates, if not missing, with Patterns 1 and 2 and with the two different GRRs.

6. DISCUSSION

In this contribution, we study several robust genome-wide scan methods for genome-wide association studies using case-control data. The genome-wide scan is different from single-marker analysis where the latter tests association between the disease and each SNP. The goal of the genome-wide scan is to rank all the SNPs and select the top SNPs for further association studies. The significance (p-value) of each SNP has not been used in the genome-wide scans. Some published results from genome-wide association studies have shown that some SNPs with true associations may not reach the genome-wide significance level. In regular simulation studies comparing powers of various test statistics, one often compares several test statistics under a single genetic model. In the simulations for the genome-wide scans, we simulated 6 SNPs with true associations with different genetic models. Hence, we can examine the performance of each approach when they compete for the top ranks.

We focused on genome-wide scans for 300K SNPs. It is known that there is no optimal test for testing association when the underlying genetic model is unknown. The efficiency robust methods of Gastwirth [5, 6] for categorical data analysis, survival analysis and legal statistics [2, 7, 8] were applied here to genetic studies and genome-wide association studies. We also demonstrated that genome-wide ranking can be done by ranking statistics rather than p-values as was previously studied by Zheng et al. [27] and Li et al. [15] using MAX (or MAX3). In this contribution, we extend the genetic model selection approach [28] to genome-wide scans. The robust test of WTCCC [23] and Joo et al. [11] is also applied to genome-wide scans for comparison. Our simulation results show that the genetic model selection approach in general outperforms the other methods even though the difference is minor. Based on the performance of the genetic model selection approach for both single marker analysis (Zheng and Ng [28] and Joo et al. [12]) and the genome-wide scans that we investigated in this article, we recommend it for screening all SNPs in initial genome-wide association studies.

ACKNOWLEDGEMENTS

We would like to thank two reviewers for their helpful suggestions and comments which greatly improved our presentations.

Received 9 September 2008

REFERENCES

[1] Elston, R. C., Lin, D. and Zheng, G. (2007). Multistage sampling for genetic studies. *Annu. Rev. Genomics Hum. Genet.* **8** 327–342.

[2] Freidlin, B., Podgor, M. J. and Gastwirth, J. L. (1999). Efficiency robust tests for survival or ordered categorical data. *Biometrics* **55** 883–886. MR1705680

[3] Freidlin, B., Zheng, G, Li, Z. and Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum. Hered.* **53** 146–152.

[4] Gail, M. H., Pfeiffer, R. M., Wheeler, W. and Pee, D. (2007). Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. *Biostatistics* **9** 201–215.

[5] Gastwirth, J. L. (1966). On robust procedures. *J. Am. Stat. Assoc.* **61** 929–948. MR0205397

[6] Gastwirth, J. L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *J. Am. Stat. Assoc.* **80** 380–384. MR0792737

[7] Gastwirth, J. L. (1997). Statistical evidence in discrimination cases, *J. Roy. Stat. Soc. Ser. A* **160** 289–303.

[8] Gastwirth, J. L. (2002). Comment on the Age Discrimination Example. *Jurimetrics J.* **42** 333–340.

[9] Guedj, M., Nuel, G. and Prum, B. (2008). A note on allelic tests in case-control association studies. *Ann. Hum. Genet.* **72** 407–409.

[10] Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.* **4** 701–709.

[11] Joo, J., Kwak, M., Ahn, K. and Zheng, G. (2008). A robust genome-wide scan statistic of Wellcome Trust Case-Control Consortium. *Biometrics*, in press.

[12] Joo, J., Kwak, M. and Zheng, G. (2008). Improving power for testing genetic association in case-control studies by reducing alternative space. *Biometrics*, revised for invited revision.

[13] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C. and Hoh, J. (2005). Complement factor H polymorphism in aged-related macular degeneration. *Science* **308** 385–389.

[14] Li, Q., Zheng, G., Li, Z. and Yu, K. (2008). Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies. *Ann. Hum. Genet.* **72** 397–406.

[15] Li, Q., Yu, K., Li, Z. and Zheng, G. (2008). MAX-rank: a simple and robust genome-wide scan for case-control association studies. *Hum. Genet.* **123** 617–623.

[16] Marchini, J., Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37** 413–417.

[17] Sasieni, P. D. (1997). From genotype to genes: doubling the sample size. *Biometrics* **53** 1253–1261. MR1614374

[18] Schaid, D. J., McDonnell, S. K., Hebbring, S. J., Cunningham, J. M. and Thibodeau, S. N. (2005). Nonparametric tests of association of multiple genes with human diseases. *Am. J. Hum. Genet.* **76** 780–793.

[19] Skol, A. D., Scott, L. J., Abecasis, G. R. and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38** 209–213.

[20] Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vinvent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445** 881–885.

[21] Song, K. and Elston, R. C. (2006). A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Stat. Med.* **25** 105–126. MR2222077

[22] Wang, T., Zhu, X. and Elston, R. C. (2007). Improving power in contrasting linkage disequilibrium patterns between cases and controls. *Am. J. Hum. Genet.* **80** 911–920.

[23] Wellcome Trust Case-Control Consortium (WTCCC) (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

[24] Zaykin, D. V. and Zhivotovsky, L. A. (2005). Ranks of genuine associations in whole-genome scans. *Genetics* **171** 813–823.

[25] Zheng, G., Freidlin, B., Li, Z. and Gastwirth, J. L. (2003). Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometrical J.* **45** 335–348. MR1973305

[26] Zheng, G., Freidlin, B. and Gastwirth, J. L. (2006). Comparison of robust tests for genetic association using case-control studies. *IMS Lecture Notes – Monograph Series* (2nd special issue in honor of E. L. Lehmann), 320-336. MR2338547

[27] Zheng, G., Joo, J., Lin, J.-P., Stylianou, M., Waclawiw, M. A. and Geller, N. L. (2007). Robust ranks of true associations in case-control association studies. *BMC Proceed.* **1** (Suppl 1) S165.

[28] Zheng, G. and Ng, H. K. T. (2008). Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics* **9** 391–399.

[29] Zheng, G. (2008). Can the allelic test be retired from analysis of case-control association studies? *Ann. Hum. Genet.* **72** 848–851.

[30] Zheng, G., Joo, J. and Yang, Y. (2008). Pearsons test, trend test, and MAX are all trend tests with different types of scores. *Ann. Hum. Genet.*, revised for minor revision.

Gang Zheng
6701 Rockledge Drive, Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892-7913, USA
E-mail address: zhengg@nhlbi.nih.gov

Jungnam Joo
6701 Rockledge Drive, Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892-7913, USA
E-mail address: jooj@nhlbi.nih.gov

Xin Tian
6701 Rockledge Drive, Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892-7913, USA
E-mail address: tianx@nhlbi.nih.gov

Colin O. Wu
6701 Rockledge Drive, Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892-7913, USA
E-mail address: wuc@nhlbi.nih.gov

Jing-Ping Lin
6701 Rockledge Drive, Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892-7913, USA
E-mail address: linj@nhlbi.nih.gov

Mario Stylianou
6701 Rockledge Drive, Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892-7913, USA
E-mail address: stylianm@nhlbi.nih.gov

Myron A. Waclawiw
6701 Rockledge Drive, Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892-7913, USA
E-mail address: waclawim@nhlbi.nih.gov

Nancy L. Geller
6701 Rockledge Drive, Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892-7913, USA
E-mail address: gellern@nhlbi.nih.gov