Assessing Genome-Wide Statistical Significance for Large *p* Small *n* Problems

Guoqing Diao*,[1] and Anand N. Vidyashankar*
*Department of Statistics, George Mason University, Fairfax, Virginia 22030

ABSTRACT Assessing genome-wide statistical significance is an important issue in genetic studies. We describe a new resampling approach for determining the appropriate thresholds for statistical significance. Our simulation results demonstrate that the proposed approach accurately controls the genome-wide type I error rate even under the large *p* small *n* situations.

QUANTITATIVE trait loci (QTL) mapping plays an important role in understanding the genetic variations in experimental crosses. A critical issue concerns assessing the genome-wide significance (GWS), since statistical tests are performed at many putative loci. Analytic methods to determine GWS have been investigated by several authors including Lander and Kruglyak (1995) and Zou *et al.* (2004) and these involve specific assumptions on the experimental design and genetic map density. Churchill and Doerge (1994) proposed a permutation test to address these issues. However, their method is computationally intensive due to repeated analyses of the permuted data sets, and its validity relies on the assumption of complete exchangeability under the null hypothesis, which can frequently be violated (see, for instance, Manichaikul *et al.* 2007).

To overcome the limitations of the permutation methods, Zou *et al.* (2004) proposed a resampling procedure requiring one analysis of the data set only, thereby reducing the computational complexity. In this note, we propose a further modification of the resampling approach of Zou *et al.* (2004) to improve the power of the tests while retaining the same computational complexity.

We begin by considering n independent subjects from an experimental cross and statistical testing at p putative loci. Let $\boldsymbol{\beta}_j$ be a vector of the genetic effects at the jth location and H_0: $\boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_p = 0$ denote the null hypothesis of no genetic effects at all loci. It is well known that given a statistical model, the likelihood-ratio test for testing the hypothesis H_{0j}: $\boldsymbol{\beta}_j = 0$ can be approximated by the score statistic

$$W_j = \mathbf{U}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{U}_j,$$

where $\mathbf{U}_j = \sum_{i=1}^n \mathbf{U}_{ij}$, $\mathbf{V}_j = \sum_{i=1}^n \mathbf{U}_{ij}\mathbf{U}_{ij}^{\mathrm{T}}$, and U_{ij} is the efficient score from the ith subject, defined to be the projection of the score function for β_j on the orthocomplement space of the score functions for nuisance parameters (Bickel *et al.* 1993, p. 30). The test statistic for testing H_0 is $\max_{1\le j\le p} W_j$, whose null distribution can be approximated by the resampling algorithm of Zou *et al.* (2004) given below. Theoretical justification for this approximation can be provided along the line of Kuelbs and Vidyashankar (2010). The resampling algorithm for determining the threshold at GWS level α (Zou *et al.* 2004) follows:

$k = 0$
repeat
  $k \leftarrow k + 1$
  $G_i(k) \overset{\text{i.i.d.}}{\sim} N(0,1),\ \ i = 1, \cdots, n$
  $\mathbf{U}_j^*(k) = \sum_{i=1}^n \mathbf{U}_{ij} G_i(k), W_j^*(k) = \mathbf{U}_j^{*\mathrm{T}}(k)\mathbf{V}_j^{-1}\mathbf{U}_j^*(k)$
  $W^\star(k) = \max_{1\le j\le p} W_j^\star(k)$
until $k \ge B$
  Calculate the $100(1-\alpha)$th percentile of $\{W^\star(1), \ldots, W^\star(B)\}$

We modify the above algorithm by generating $G_i(k) \overset{\text{i.i.d.}}{\sim} 2\times\text{Bernoulli}(0.5) - 1$; *i.e.*, $G_i(k)$'s are i.i.d. from the Rademacher distribution, since the error in approximating the distribution of the score statistic is of the order $n^{-3/2}$ when using Rademacher weights (RW) while it is $3n^{-3/2}$ for $N(0, 1)$ weights. This distribution is commonly used in the

doi: 10.1534/genetics.113.150896
Manuscript received February 26, 2013; accepted for publication April 26, 2013
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.150896/-/DC1.
[1]Corresponding author: Department of Statistics, George Mason University, MS 4A7, Fairfax, VA 22030. E-mail: gdiao@gmu.edu

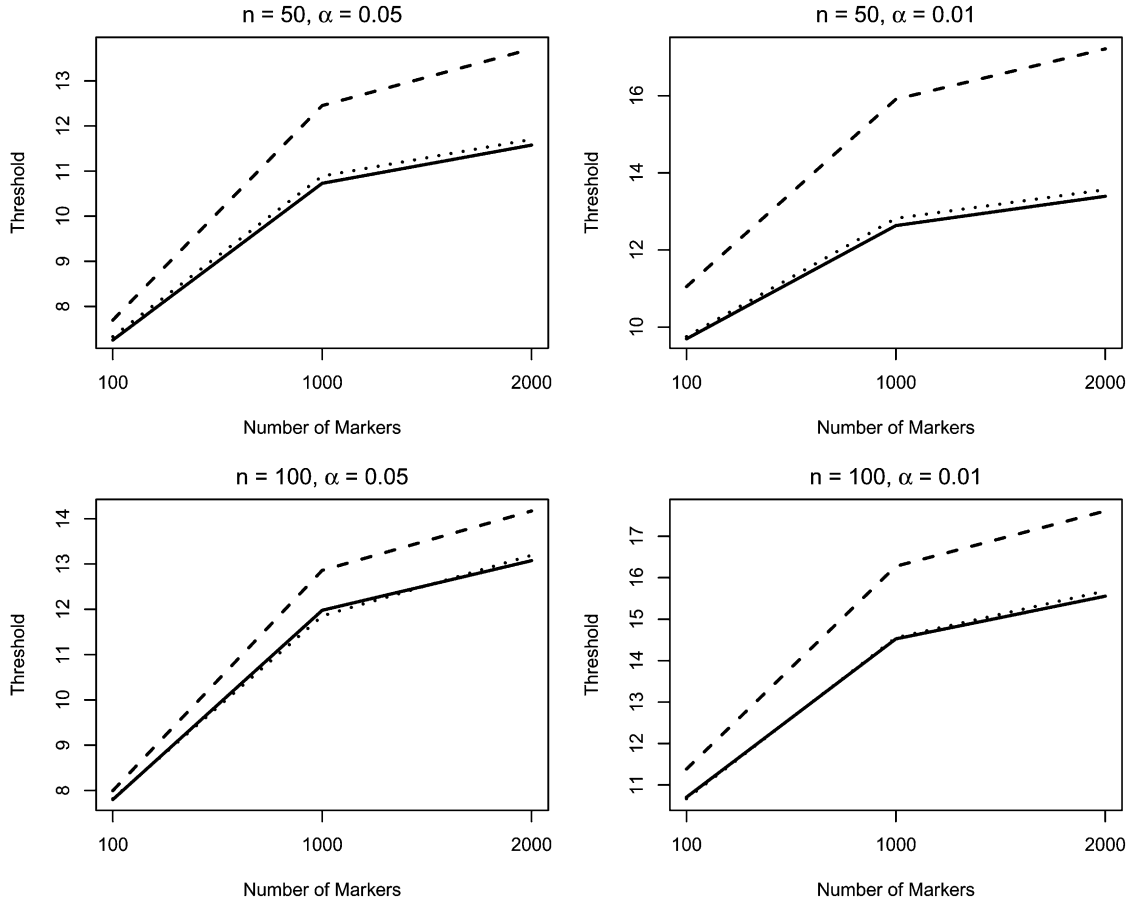

Figure 1 Thresholds at the targeted GWS levels of α. The solid, dashed, and dotted curves correspond to the average thresholds based on the proposed method and the method of Zou *et al.* (2004) from 10,000 simulated data sets and the empirical thresholds based on 10,000 simulated data sets under the null hypothesis, respectively.

multiplier bootstrap literature (Praestgaard 1990), econometrics, and learning theory (Bartlett and Mendelson 2003; Koltchinskii and Panchenko 2000) and measures how well correlated the most-correlated hypothesis is to a random labeling of the efficient scores.

We conducted simulation studies to study the effect of using RW with sample of sizes 50 and 100 and genetic maps of 1, 10, and 20 chromosomes. Each chromosome has a length of 100 cM and 100 equally spaced markers. We use the function sim.cross in R/qtl (Broman *et al.* 2003) to generate the genotype data. Under H_0, we generate the quantitative traits from $N(0, 1)$ while under H_1, we generate from $N(\mu, 1)$, $\mu \in [0.2, 1.0]$, representing an additive effect at 35 cM on chromosome one.

Figure 1 presents the thresholds for the single-marker analysis at the GWS level of 0.05 and 0.01 and compares them to both the empirical thresholds and that of Zou *et al.* (2004) based on 10,000 replicates and $B = 10,000$ in the algorithm. When $n = p$, the thresholds based on both methods match the empirical thresholds, under both H_0 and H_1 (data not presented under H_1). When n is small and $<p$, the thresholds using RW still match the empirical thresholds, whereas the thresholds from Zou *et al.* (2004) are overestimated. The standard errors of the thresholds were also calculated using the function *quantileSE* in the R package *broman* (detailed results are presented in supporting information, File S1). Figure 2 presents the sizes and powers of the two resampling approaches. The proposed approach has type I error rates close to the nominal level under all situations and is substantially more powerful than Zou *et al.* (2004) under the large p small n scenarios.

In summary, we proposed a new resampling approach for assessing GWS in QTL mapping. This new approach retains all the attractive features of the resampling approach of Zou *et al.* (2004) and outperforms it under the large p small n situation. Additional simulation studies with $n = 500$ and $P = 2000$ showed that the two methods yielded similar results (detailed results are presented in File S1).

Acknowledgments

The authors thank the editor and a referee for helpful comments. The first author was supported in part by National Science Foudnation (NSF) DMS-1107108 and National Institutes of Health CA150698. The second author was supported by NSF DMS-1107108.

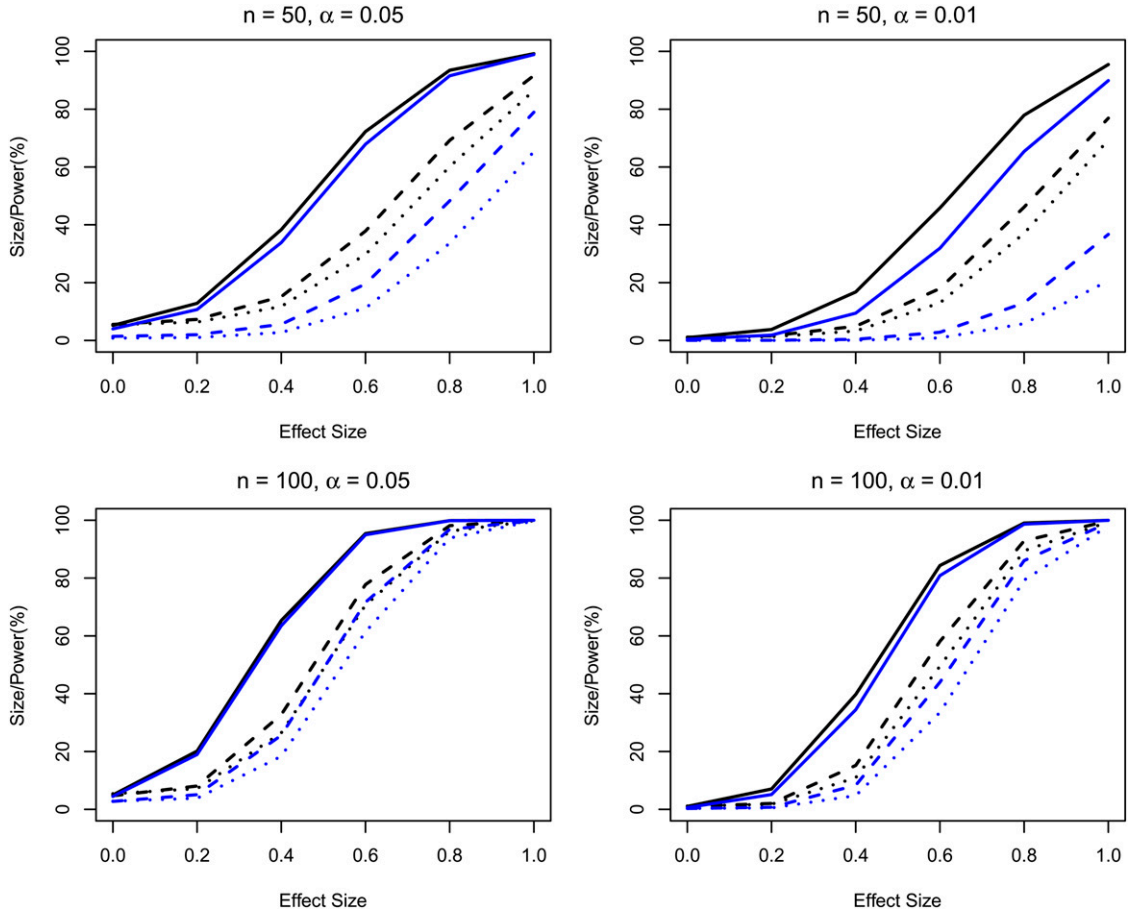

Figure 2 Sizes/powers(%) at nominal GWS level of α. The black and blue curves correspond to sizes/powers from the proposed method and the method of Zou *et al.* (2004), respectively. The solid, dashed, and dotted curves correspond to the sizes/powers under the scenarios when p = 100, 1000, and 2000, respectively.

Literature Cited

Bartlett, P. L., and S. Mendelson, 2003 Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. 3: 463–482.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner, 1993 *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Broman, K. W., H. Wu, S. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889–890.

Churchill, G. A., and R. W. D. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics 138: 963–971.

Koltchinskii, V., and D. Panchenko, 2000 Rademacher processes and bounding the risk of function learning. Progr. Probab. 47: 443–458.

Kuelbs, J., and A. N. Vidyashankar, 2010 Asymptotic inference for high-dimensional data. Ann. Stat. 38: 836–869.

Lander, E., and L. Kruglyak, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. 11: 241–247.

Manichaikul, A., A. A. Palmer, S. Sen, and K. W. Broman, 2007 Significance thresholds for quantitative trait locus mapping under selective genotyping. Genetics 177: 1963–1966.

Praestgaard, J., 1990 *Bootstrap with general weights and multiplier central limit theorems. Technical Report 195*, Department of Statistics, University of Washington.

Zou, F., J. Fine, J. Hu, and D. Y. Lin, 2004 An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. Genetics 168: 2307–2316.

Communicating editor: F. Zou

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.150896/-/DC1

Assessing Genome-Wide Statistical Significance for Large *p* Small *n* Problems

Guoqing Diao and Anand N. Vidyashankar

DOI: 10.1534/genetics.113.150896

File S1

Supplementary materials for "On Assessing Genome-Wide Statistical Significance for Large p Small n Problems"

Guoqing Diao and Anand N. Vidyashankar

Department of Statistics, George Mason University, Fairfax, VA 22030

Additional Simulation Studies

GWAS study with SNP data

We conducted additional simulation studies to evaluate the performance of the proposed method for genome-wide association studies with single nucleotide polymorphisms (SNPs) data. We considered a study that scans 10 independent genome regions with 200 biallelic SNPs in each region. For each SNP, we assume Hardy-Weinberg equilibrium and set the minor allele frequency to be 0.4. Within each genome region, the linkage disequilibrium (LD) between successive two loci varied from 0 to 0.18. Under the null hypothesis, we generated the quantitative traits from a standard normal distribution; under the alternative hypothesis, we assume that the 35th SNP in genome region 1 has an additive effect. The effect sizes were set to be 1.2 and 0.8 for sample sizes of 50 and 100, respectively. The number of resamples B was set to be 2,000.

Figure 1 presents the sizes and powers of the two resampling approaches at genome-wide significance (GWS) level of 0.05 and 0.01 based on 10,000 replicates. Under all scenarios, the method of our paper has type I error rate close to the nominal level while the approach of Zou *et al*. (2004) tends to be conservative especially for $n = 50$ and significance level of 0.01. The proposed method substantially improves the power of the test over that of Zou *et al*. (2004). For example, with

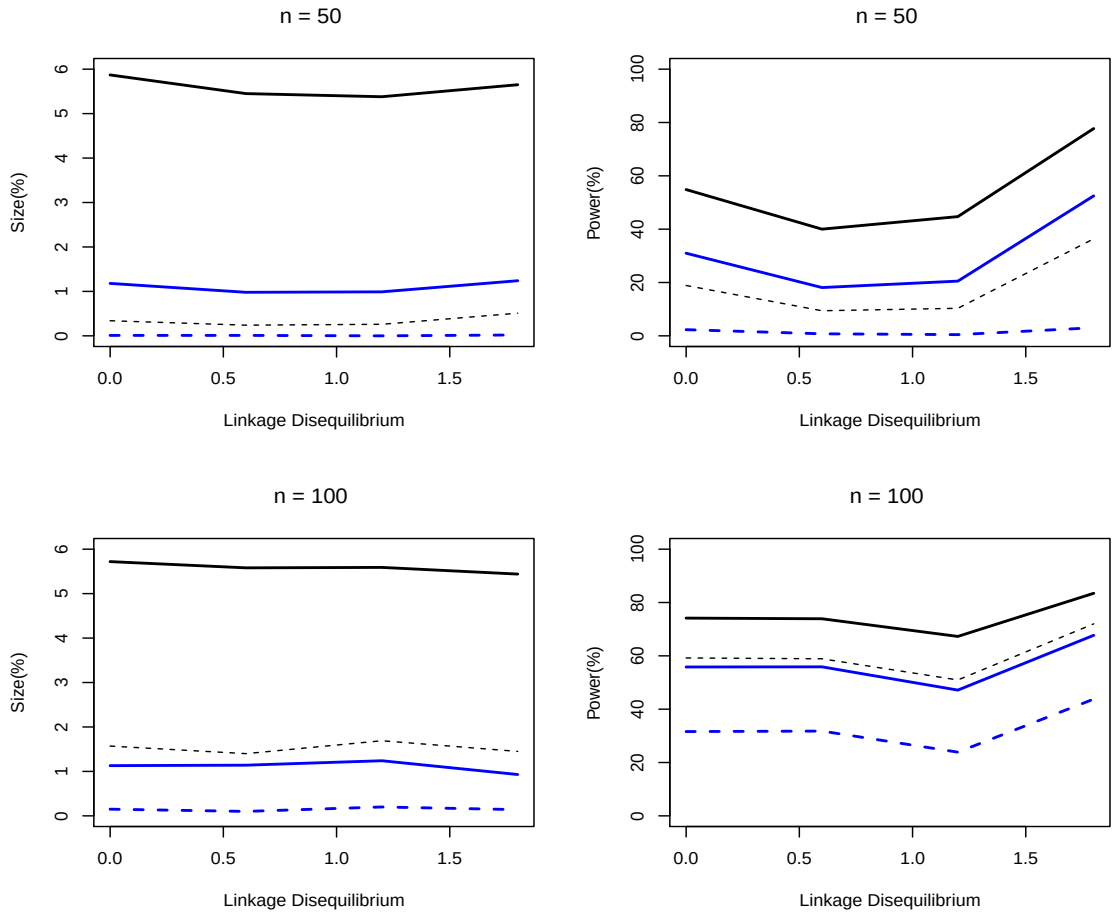

Figure 1: Sizes/powers(%) at nominal genome-wide significance level of 0.05 and 0.01. The black solid and black dashed curves correspond to the sizes/powers of the proposed method at significance levels of 0.05 and 0.01, respectively. The blue solid and blue dashed curves correspond to the sizes/powers of the method of ZOU *et al.* (2004) at significance levels of 0.05 and 0.01, respectively.

$n = 50$ and a LD of 0.18, the powers were 77.73% and 52.49% at significance levels of 0.05 and 0.01, respectively, compared to 36.37% and 3.04% of ZOU *et al.* (2004).

Simulation studies for $n = 500$

We have conducted additional simulation studies for the case of $n = 500$ and $p = 100, 1000$, and 2000. Table 1 in the Supplementary Materials presents the sizes and powers of the proposed method and the resampling approach of ZOU *et al.* (2004). These two methods yielded similar results. The reason for this is that when n is large and p is not much larger than n, the asymptotic theory takes effect. It would be desirable to conduct simulation studies to compare the two methods under the scenario of $p \gg n$ for large n. While it is feasible to analyze a real data set with both large n and large p, it is computationally prohibitive to conduct simulation studies given the current computing technology.

TABLE 1

Sizes/powers(%) at nominal genome-wide significance level of α with $n = 500$

Setup		Proposed[c]		ZOU *et al.* (2004)[d]	
p[a]	μ[b]	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
100	0.0	4.80	0.96	4.78	0.90
	0.2	73.99	50.88	73.66	50.04
1000	0.0	4.74	1.08	4.43	0.95
	0.2	44.95	25.73	43.88	24.13
2000	0.0	5.06	1.08	4.59	0.88
	0.2	37.21	20.32	35.76	18.64

[a] Total number of markers.

[b] Additive effect.

[c] Sizes/powers based on the proposed resampling method.

[d] Sizes/powers based on the resampling method of ZOU *et al.* (2004).

Computation of the standard error estimates

For the simulations described in the main manuscript, we also computed the standard error estimates on the thresholds by using the function *quantileSE* in the R package *broman*, which implements the method described in COX and HINKLEY (1974). The average of the standard error estimates based on the proposed method agree well with the standard error estimates of the empirical thresholds, obtained from 10,000 replicates under the null hypothesis. For example, for $n = 50$, $p = 100$, and $\alpha = 0.05$, the empirical threshold was 7.33 (SE=0.071) and the average of the proposed threshold was 7.25 and the average of the standard error estimates was 0.071.

References

COX, D. R., and D. V. HINKLEY, 1974 *Theoretical Statistics*. Chapman and Hall, London.

ZOU, F., J. FINE, J. HU, and D. Y. LIN, 2004 An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. Genetics **168**: 2307–2316.