A Simulation Study of Permutation, Bootstrap, and Gene Dropping for Assessing Statistical Significance in the Case of Unequal Relatedness

Riyan Cheng* and Abraham A. Palmer*,†,1

*Department of Human Genetics and †Department of Psychiatry and Behavioral Neuroscience, the University of Chicago, Chicago, Illinois 60637

ABSTRACT We used simulations to evaluate methods for assessing statistical significance in association studies. When the statistical model appropriately accounted for relatedness among individuals, unrestricted permutation tests and a few other simulation-based methods effectively controlled type I error rates; otherwise, only gene dropping controlled type I error but at the expense of statistical power.

DETERMINING statistical significance thresholds is an essential part of quantitative trait locus (QTL) mapping. Computationally efficient methods have been proposed to obtain significance thresholds via approximating the test statistic by an Ornstein–Uhlenbeck diffusion process (Lander and Botstein 1989; Dupuis and Siegmund 1999; Zou *et al.* 2001) or Davis' approximation (Davis 1987; Rebaï 1994; Piepho 2001) or by estimating the effective number of independent tests (Cheverud 2001; Moskvina and Schmidt 2008). However, these methods may not provide satisfactory results (Zou *et al.* 2001; Dudbridge and Gusnanto 2008). Simulation-based tests are still recommended (Lander and Schork 1994) and have been used extensively in QTL mapping. Permutation tests (Fisher 1935) have been a standard method with which to estimate significance thresholds in QTL mapping since they were introduced for this purpose by Churchill and Doerge (1994). Problems may arise when complex mapping populations or complicated statistical analyses are used (Zou *et al.* 2006; Churchill and Doerge 2008). In these situations, naive application of unrestricted permutation tests may lead to invalid inference because the fundamental assumption of exchangeability is violated. This problem typically occurs in mapping populations where individuals share varying degrees of genetic relatedness and has raised questions about whether permutation tests should be applied in such situations (Abney *et al.* 2002; Zou *et al.* 2005; Peirce *et al.* 2008; Cheng *et al.* 2010).

In this study, we performed extensive simulations to evaluate the permutation test as well as several other simulation-based methods: parametric bootstrapping (Efron 1979), gene dropping and genome reshuffling for advanced intercross permutation (GRAIP), for assessing significance using linear mixed effect models and advanced intercross lines (AIL) (Darvasi and Soller 1995), where individuals are known to be genetically unequally related. The primary purpose of this work was to investigate the performance of these methods with respect to type I error rates and statistical power in the context of statistical modeling and to provide useful insight in the choice of methods for estimating significance thresholds when subjects are genetically unequally related. In contrast to Valdar *et al.* (2009), which focused on modeling, our study focuses on methods for determining significance thresholds when relatedness is a concern. We report our main findings while leaving the details in Supporting Information, File S1, File S2, and File S3.

Simulation Results

We generated an AIL pedigree and sampled 576 individuals from F_{26} (Table S1). The phenotype was generated such that polygenic variation approximately accounted for 56, 46, or 32% of the total phenotypic variation, corresponding to the standard deviation 0.7, 1, or 1.5 of the residual effect.

doi: 10.1534/genetics.112.146332

Manuscript received October 1, 2012; accepted for publication December 6, 2012

Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.146332/-/DC1.

[1] Corresponding author: Department of Human Genetics, University of Chicago, 920 E. 58th St., CLSC-507D, Chicago, IL 60637. E-mail: aap@uchicago.edu

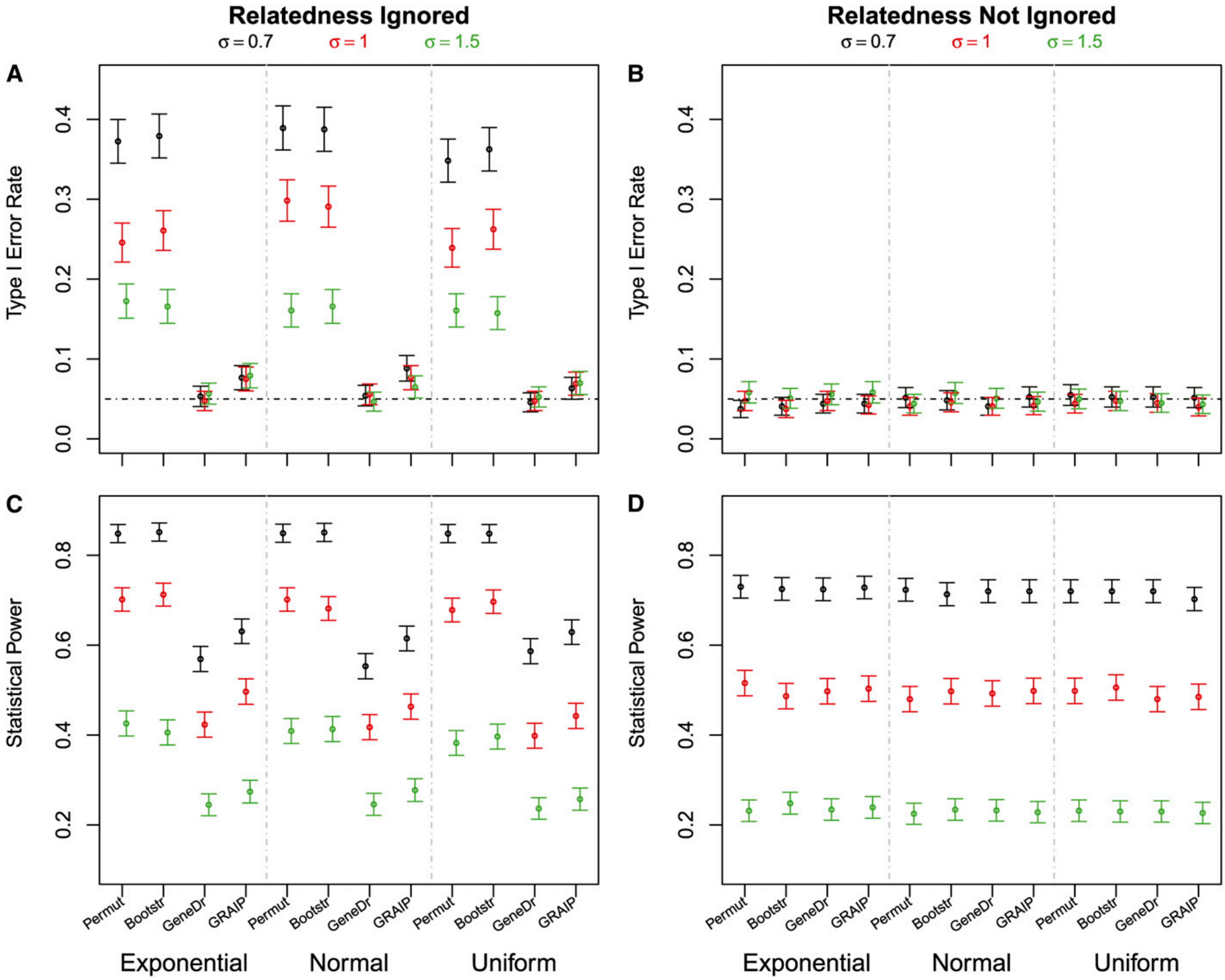

Figure 1 Type I error rates and statistical power. Type I error rates (A and B) and statistical power (C and D) estimated at genome-wide significance level 0.05 by each of the following methods: permuting genotypic data (Permut), bootstrapping phenotypic data (Bootstr), gene dropping (GeneDr), and GRAIP. The distribution of the residual was exponential, normal, or uniform, each with a standard deviation 0.7, 1, or 1.5.

Type I error

First, we ignored polygenic variation. Only the gene-dropping method effectively controlled the type I error rates; all other methods produced inflated type I error rates (Figure 1A). The larger the polygenic variation was relative to the environmental variation, the more seriously the type I error rates were inflated. GRAIP performed much better than either bootstrap or permutation but was still not able to control false positives at the expected significance level.

Next we took polygenic variation into account. All the methods controlled type I error rates at the expected levels (Figure 1B). Misspecification of the residuals produced somewhat overly conservative results, but had little impact overall (Table S2).

Statistical power

One QTL was generated with a heritability of ~2.8, 2.3, or 1.6%, corresponding to the standard deviation 0.7, 1, or 1.5 of the residual effect. Figure 1C reports power even when type I error is not controlled (*e.g.*, permutation, bootstrapping). This reflects a combination of both true and false positives. The power was comparable for all of the four methods when polygenic variation was accounted for in the model (Figure 1D). Notably, gene dropping has a higher statistical power when the relatedness was accounted for (Figure 1, C and D).

Simulations with different family sizes and subpopulation structure

We performed additional simulations by randomly choosing 288 individuals from the F_{26} sample and 288 individuals from a real data set (see below). The results were similar (data not shown), suggesting that variable family size did not negatively affect the procedures. We then considered different allele (A/a) frequencies at the founder generation: 3/1 for F_{26} *vs.* 1/3 for F_{34}. Under these conditions both permutation and bootstrap failed to control type I error when the

Table 1 Estimated Type I Error Rate and Statistical Power

		Type I error rate			Statistical power		
Distr[a]	Method[b]	α = 0.10	α = 0.05	α = 0.01	α = 0.10	α = 0.05	α = 0.01
Exp	Permut	0.191***	0.113***	0.028***	0.493	0.387	0.235
	Bootstr	0.145***	0.078***	0.022***	0.451	0.360	0.235
	GeneDr	0.108	0.045	0.009	0.402	0.312	0.164
	GRAIP	0.116*	0.052	0.012	0.416	0.315	0.164
Norm	Permut	0.129***	0.059	0.012	0.478	0.379	0.241
	Bootstr	0.090	0.048	0.007	0.409	0.343	0.223
	GeneDr	0.090	0.051	0.010	0.416	0.355	0.239
	GRAIP	0.086*	0.044	0.010	0.418	0.342	0.217
Unif	Permut	0.136***	0.079***	0.014	0.488	0.397	0.241
	Bootstr	0.104	0.057	0.011	0.435	0.352	0.219
	GeneDr	0.104	0.062*	0.011	0.429	0.351	0.220
	GRAIP	0.105	0.060	0.011	0.430	0.352	0.246

Allele (A/a) frequencies at the founder generation: 3/1 for F_{26} *vs.* 1/3 for F_{34}. Estimated from 1200 simulations at genome-wide significance level α = 0.10, 0.05 or 0.01. *, **, and *** indicate that the estimated type I error rate is significantly different from the expected significance levels 0.10, 0.05, and 0.01, respectively.

[a] Residual distribution: exponential (Exp), normal (Norm), or uniform (Unif).

[b] Permuting marker data (Permut), bootstrapping phenotypic data (Bootstr), or gene dropping (GeneDr).

residual was exponentially distributed and permutation also failed to control type I error when the residual was uniformly distributed (Table 1). This is broadly consistent with our main point, which is that when the model used to analyze the data are correctly chosen, permutation is an effective strategy for analyzing the data.

Real data example

We used a data set from a 34th generation of a mouse AIL, which consisted of body weight measurements and genotypes for 688 mice at 3105 SNPs (Cheng *et al.* 2010; Parker *et al.* 2011). We did not perform the exact GRAIP procedure; instead, we shuffled simulated F_{33} haplotype pairs within sex and then simulated F_{34} genotypes. This simplified the analysis while maintaining the key property of GRAIP, *i.e.*, its ability to retain relatedness solely for full sibship. The estimated thresholds were similar when polygenic variation was accounted for in the model (Table S3). Both permutation and bootstrap produced similar thresholds regardless of whether polygenic variation was ignored or accounted for in the model. In contrast, both gene dropping and GRAIP yielded significantly larger thresholds when polygenic variation was ignored.

Discussion

There has been widespread concern about the use of permutation tests in complex mapping designs (Abney *et al.* 2002; Zou *et al.* 2005; Churchill and Doerge 2008; Peirce *et al.* 2008). In a previous publication we observed that permutation and gene dropping produced similar thresholds in the analysis of an AIL when polygenic variation was incorporated in the model (Cheng *et al.* 2010); however, that article did not explore the finding, consider alternative methods, or explore statistical power. Here we studied four simulation-based methods for obtaining empirical significance thresholds: permuting genotypes, bootstrapping phenotypes, gene dropping, and GRAIP. The permutation test has been a standard simulation-based method in QTL mapping, the bootstrap test is among the most useful empirical methods in statistics and has been recommended in mixed effect models (Pinheiro and Bates 2000; Valdar *et al.* 2009), and gene dropping is appropriate when pedigree information is available. We found that all these methods worked well when polygenic variation was appropriately taken into account in the model; however, when polygenic variation was ignored, only gene dropping was able to control type I error rates and this came at the expense of statistical power (Figure 1, C and D). Thus, it is important to specify an appropriate statistical model in QTL mapping, especially in complex populations such as AIL; an inappropriate model can invalidate statistical inference. These principles should extend to general cases where unequal relatedness or a population structure exists.

We found that the estimated distribution of the test statistic under the null hypothesis (no real QTL) was similar whether or not polygenic variation was accounted for in the model for some of the methods we examined but not for others (Table S4). In particular, the estimated distribution was significantly different when using gene dropping and GRAIP but not when using bootstrap or permutation. The take-home message is that if the model is appropriate for a genome-wide scan, we may ignore the random polygenic effect to reduce computation when performing permutation tests to estimate the significance threshold. We also found that when the polygenic variation was accounted for in the model, the estimated distributions of the test statistic for all the four methods were not significantly different from one another. One possible explanation for this is that the trait values of genetically related individuals tend to be similar and thus the test statistic is inflated because of the confounding effect between the genotype and the phenotype adjusted for other effects in the model when the polygenic variation is ignored. Gene dropping (or to a lesser extent GRAIP) retains the relationship and is therefore capable of controlling the false-positive rate regardless of the inclusion of polygenic variation. The permutation (or bootstrap) test largely dissolves the confounding and therefore provides similar thresholds regardless of whether or not the polygenic variation is accounted for in the model, and it cannot control the false-positive rate if the polygenic variation is ignored.

Our observations were mainly based on AIL data. It is worth pointing out that the permutation test, as well as the bootstrap test, should be used with caution. Model appropriateness such as independency, normality, and constancy of residuals is a general concern in statistical modeling. We showed that the permutation test was not robust to misspecification of the residual distribution when the population was structured with different allele frequencies (Table 1). In addition, a major QTL (or a polygene with relatively large effects) may result in false positives due to uncontrolled confounding between the QTL (or polygene) and a scanning locus. In such a case, incorporating major QTL and possibly a few loci with relatively large effects as covariates in the model may address this concern (Valdar *et al.* 2009; Segura *et al.* 2012).

Acknowledgments

We acknowledge the valuable input of Mark Abney and Andrew Skol on topics related to this work. We also appreciate the useful discussions with Gary Churchill, Karl Broman, Saunak Sen, and William Valdar. This project was supported by National Institutes of Health grants DA024845, DA021336, and MH079103.

Literature Cited

Abney, M., C. Ober, and M. S. McPeek, 2002 Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the hutterites. Am. J. Hum. Genet. 70: 920–934.

Cheng, R., J. E. Lim, K. E. Samocha, G. Sokoloff, M. Abney *et al.*2010 Genome-wide association studies and the problem of relatedness among advanced intercross lines and other highly recombinant populations. Genetics 185: 1033–1044.

Cheverud, J. M., 2001 A simple correction for multiple comparison in interval mapping genome scans. Heredity 87: 52–58.

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

Churchill, G. A., and R. W. Doerge, 2008 Naive application of permutation testing leads to inflated type i error rates. Genetics 178: 609–610.

Darvasi, A., and M. Soller, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141: 1199–1207.

Davis, R. B., 1987 Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 74: 33–43.

Dudbridge, F., and A. Gusnanto, 2008 Estimation of significance thresholds for genomewide association studies. Genet. Epidemiol. 32: 227–234.

Dupuis, J., and D. Siegmund, 1999 Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151: 373–386.

Efron, B., 1979 Bootstrap methods: another look at the jackknife. Ann. Stat. 7(1): 1–26.

Fisher, R. A., 1935 *The Design of Experiment*. Hafner Press, New York.

Lander, E. S., and D. Botstein, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199.

Lander, E. S., and N. J. Schork, 1994 Genetic dissection of complex traits. Science 265: 2037–2048.

Moskvina, V., and K. M. Schmidt, 2008 On multiple-testing correction in genome-wide association studies. Genet. Epidemiol. 32: 567–573.

Parker, C. C., R. Cheng, G. Sokoloff, J. E. Lim, A. D. Skol *et al.*, 2011 Fine-mapping alleles for body weight in LG/J × SM/J F2 and F34 advanced intercross lines. Mamm. Genome 22: 563–571.

Peirce, J. L., K. W. Broman, L. Lu, E. J. Chesler, G. Zhou *et al.*, 2008 Genome reshuffling for advanced intercross permutation (GRAIP): simulation and permutation for advanced intercross population analysis. PLoS ONE 3(4): e1977.

Piepho, H.-P., 2001 A quick method for computing approximate thresholds for quantitative trait loci detection. Genetics 157: 425–432.

Pinheiro, J. C., and D. M. Bates, 2000 *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, New York.

Rebaï, A. B., 1994 Approximate thresholds of interval mapping tests for QTL detection. Genetics 138: 235–240.

Segura, V., B. J. Vilhjalmsson, A. Platt, A. Korte, U. Seren *et al.*, 2012 An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44(7): 825–830.

Valdar, W., C. C. Holmes, R. Mott, and J. Flint, 2009 Mapping in structured populations by resample model averaging. Genetics 182: 1263–1277.

Zou, F., J. A. L. Gelfond, D. C. Airey, L. Lu, K. F. Manly *et al.*, 2005 Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. Genetics 170: 1299–1311.

Zou, F., B. S. Yandell, and J. P. Fine, 2001 Statistical issues in the analysis of quantitative traits in combined crosses. Genetics 158: 1339–1346.

Zou, F., Z. Zu, and T. Vision, 2006 Assessing the significance of quantitative trait loci in replicable mapping populations. Genetics 174: 1063–1068.

Communicating editor: G. A. Churchill

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.146332/-/DC1

A Simulation Study of Permutation, Bootstrap, and Gene Dropping for Assessing Statistical Significance in the Case of Unequal Relatedness

Riyan Cheng and Abraham A. Palmer

DOI: 10.1534/genetics.112.146332

File S1

Supplemental Material: A Simulation Study of Permutation, Bootstrap and Gene Dropping for Assessing Statistical Significance in the Case of Unequal Relatedness

This supplement contains a number of sections that are meant as reference material that extends on the level of detail provided in the main text. It is not designed to be read from beginning to end and does not conform to a narrative format in the way a journal article might.

Statistical Model

A typical genetic model for mapping a diploid population with alleles A and a at a locus is as follows

$$y_i = \boldsymbol{x}_i{}'\boldsymbol{\beta} + x_i^* a^* + z_i^* d^* + u_i + \epsilon_i, \quad i = 1, 2, \cdots, n \tag{1}$$

where y_i is the trait value for the i-th individual, $\boldsymbol{x}_i$ represents covariates (e.g. sex) and $\boldsymbol{\beta}$ are the corresponding effects, x_i^* is $1, 0$ or -1 if the genotype at the putative QTL is AA, Aa or aa and a^* is the additive effect of the putative QTL, z_i^* is 1 if the genotype at the putative QTL is heterozygous or 0 if the genotype is homozygous and d^* is the dominance effect, u_i represents polygenic variation, and ϵ_i denotes the residual effect. Assume that $\epsilon_i \sim N(0, \sigma^2), i = 1, 2, \cdots, n$ are independent, and $\boldsymbol{u} = (u_1, u_2, \cdots, u_n)' \sim N_n(\boldsymbol{0}, \boldsymbol{G})$ with $\boldsymbol{G} = (g_{ij})$ and is independent of $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$. It is known (Jackquard, 1974; Abney et al., 2000) that in general

$$\begin{aligned} g_{ij} &= 2\Phi_{ij}\sigma_a^2 + \Delta_{ij,7}\sigma_d^2 + (4\Delta_{ij,1} + \Delta_{ij,3} + \Delta_{ij,5})Cov(a,d) \\ &\quad + \Delta_{ij,1}\sigma_h^2 + (\Delta_{ij,1} + \Delta_{ij,2} - f_i f_j)\mu_h^2 \\ &\overset{def}{=} g_{ij}^{(a)}\sigma_a^2 + g_{ij}^{(d)}\sigma_d^2 + g_{ij}^{(ad)}Cov(a,d) \\ &\quad + g_{ij}^{(h)}\sigma_h^2 + g_{ij}^{(m)}\mu_h^2 \end{aligned} \tag{2}$$

where Φ_{ij} is the kinship coefficient between the i-th and j-th individuals, f_i is the inbreeding coefficient for the i-th individual, Δ_{ij}'s are identity coefficients as defined in Lynch and Walsh (1998, pp.133) and can be calculated from the pedigree data, and $g_{ij}^{(a)}$ denotes $2\Phi_{ij}$ etc. Abney et al. (2000) suggested that the last three polygenic variance components, σ_h^2, $Cov(a,d)$ and μ_h^2, in g_{ij} are negligible, and we ignored these three variance components for ease of computation. Though it is common to only consider the additive polygenic variance component (e.g. Yu et al., 2006; Kang et al., 2008), we prefer to keep both the additive and dominance polygenic variance components.

Permutation, Bootstrap, Gene Dropping and Genome Reshuffling for Advanced Intercross Permutation

The following four simulation-based methods for estimating significance thresholds were used:

Permutation tests A permutation test is a randomization test. It is a re-sampling procedure. Typically, the data points are randomly reassigned to subjects and then the permuted data is reanalyzed to obtain the test statistic. The process is repeated many times. The values of the test statistic obtained from the permuted data are treated as a sample from the distribution of the test statistic of the original data under the null hypothesis, and the threshold at significance level α is then estimated by the $100(1-\alpha)$th percentile of this set of values.

A fundamental requirement for valid permutation is exchangeability, which should be ensured by the design of an experiment or be assumed under the null hypothesis (Anderson, 2001; Nichols and Holmes, 2001). A permutation test is exact when permutation is performed within exchangeable units. Exact permutation tests do not exist when data points are not exchangeable, for instance, in a linkage analysis where a continuous variable is used as a covariate. In this case, one may consider approximate permutation tests. Different strategies have been proposed to perform approximate permutation tests, including permutation of the raw data or residuals under null hypothesis (see e.g. Anderson, 2001), restricted permutation (Zou et al., 2005), and permutation of transformed residuals (Abney et al., 2002). The performance of approximate permutation tests varies in different experimental designs (Anderson and Braak, 2003).

Permuting the phenotypic data and permuting the genotypic data are two different ways to perform permutation in QTL mapping. We permuted genotypic data, which would retain the relationship between the trait and other predictors (e.g. sex) and could result in better estimation (O'Gorman, 2005).

Bootstrap tests Bootstrap is another popular re-sampling procedure. Bootstrap has a wide range of statistical applications including hypothesis testing (e.g. Efron and Tibishirani, 1993). There are two versions of bootstrap: non-parametric bootstrap and parametric bootstrap. While non-parametric bootstrap draws samples from the original data with replacement, parametric bootstrap generates data from a fitted model. We now briefly discuss how to use parametric bootstrap in our situation. Under the hypothesis of no QTL, model (1) reduces to $y_i = \boldsymbol{x}_i{}'\boldsymbol{\beta} + u_i + \epsilon_i, \quad i = 1, 2, \cdots, n$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)' \sim N_n(\boldsymbol{x}\boldsymbol{\beta}, \boldsymbol{G} + \boldsymbol{I}\sigma^2)$ with $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)'$ and $\boldsymbol{G} = (g_{ij}^{(a)}\sigma_a^2 + g_{ij}^{(d)}\sigma_d^2)$. We can fit the model and obtain parameter estimates $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_a^2$, $\hat{\sigma}_d^2$ and $\hat{\sigma}^2$, and then generate a sample $\boldsymbol{y}^{(b)} = (y_1^{(b)}, y_2^{(b)}, \cdots, y_n^{(b)})'$ from $N_n(\boldsymbol{x}\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{G}} + \boldsymbol{I}\hat{\sigma}^2)$ with $\hat{\boldsymbol{G}} = (g_{ij}^{(a)}\hat{\sigma}_a^2 + g_{ij}^{(d)}\hat{\sigma}_d^2)$. When polygenic variation is ignored, $\boldsymbol{y}^{(b)}$ is generated from $N_n(\boldsymbol{x}\hat{\boldsymbol{\beta}}, \boldsymbol{I}\hat{\sigma}^2)$ instead. We then analyze $\boldsymbol{y}^{(b)}$ the same way as we analyze the original data $\boldsymbol{y}$. The values of the test statistic calculated from a number (say 1000) of bootstrap samples are pooled to estimate significance thresholds in the same way we described for permutation tests. Our approach should be similar to what is described in ``Alternative mapping methods 2'' in Valdar et al. (2009).

Gene dropping tests Gene dropping is yet another re-sampling procedure. Instead of re-sampling phenotypes, it uses pedigree information and Mendelian segregation principles to generate genotypic data. The idea is straightforward. If we know the haplotypes in a pair of parents and recombination rates between loci, we can simulate haplotypes (and thus genotypes) in an offspring by simulating meiosis. If we know the haplotypes in the founders, a full pedigree and a genetic map, we can simulate genotypes for any individuals in the pedigree (see Cheng et al., 2010, for more details). Gene dropping has been used to assess significance in a wide range of applications such as genetic variability (MacCluer et al., 1986; Pardo et al., 2005; Thomas, 1990), inbreeding and allele sharing (Suwanlee et al., 2007; Jung et al., 2006), and genome-wide association studies (Cheng et al., 2010). A limitation of gene dropping is the need for a pedigree.

GRAIP Genome reshuffling for advanced intercross permutation, or GRAIP, was proposed by Peirce et al. (2008) for situations where relatedness is a concern but a complete pedigree is not available. The haplotype pairs in the parents of the last generation are permuted across the parents within each sex and then genotypic data for the individuals in the last generation are generated from the permuted haplotypes by gene dropping, using the pedigree information about nuclear families only. As the haplotypes in the parents are unknown in practice, one needs to derive phase data for the parents. This was not an issue in our studies because the haplotype (and thus genotype) data were generated using the gene dropping procedure so phase was known.

Simulation Details

Additional details of our simulation studies are provided here:

Generate a pedigree We used advanced intercross lines (AIL) as our mapping population. We created a pedigree of twenty-six generations from two inbred founder strains. In F$_n$ $(2 \leq n < 25)$, there were 144 breeding pairs and each pair produced one female and one male progeny. The 144 female progeny randomly paired with the 144 male progeny to breed the next generation. Each breeding pair in F$_{25}$ had four progeny, which created our sample of size 576. This pedigree resulted in varying relatedness among F$_{26}$ individuals (supplemental table S1).

Simulate genotypic and phenotypic data It was assumed that there were twenty chromosomes and 101 markers were evenly distributed every 1 cM on each chromosome. One of every five markers on the second ten chromosomes were chosen as polygenic QTL to generate polygenic variation. The additive and dominance effects of the polygenic QTL were randomly uniformly distributed in $(-0.2, 0.2)$ and $(-0.04, 0.04)$ respectively.

Phenotypic data were generated from equation (1), with an overall mean 0 and polygenic effects as stated above. The relatedness measurements were calculated from the pedigree as described in Cheng et al. (2010). The standard deviation σ of the residual ϵ_i was 0.7, 1 or 1.5, and the corresponding polygenic effects on average approximately accounted for 56%, 46%, or 32% of the total variation in the phenotype. Genotypic data were generated by gene dropping using the pedigree.

To investigate robustness of a test to misspecification of the residual's distribution, we generated data from exponential and uniform distributions in addition to normal distributions.

Obtaining significance thresholds We used four methods to test for QTL: permutation, parametric bootstrap (e.g. Efron and Tibishirani, 1993), gene dropping and genome reshuffling for advanced intercross permutation (GRAIP) Peirce et al. (2008). In the permutation test, we permuted genotypic data without restriction unless specified otherwise. We were especially interested to investigate the performance of the permutation test in the context of statistical modeling. In applications, one may choose restricted permutation if appropriate.

Type I error The genome scan for QTL under the null hypothesis of no QTL was performed on the first ten chromosomes, where there were no QTL. For each set of parameter values, 1200 datasets were generated and each dataset was analyzed using the likelihood ratio test (LRT). The type I error rate was estimated by the proportion of the 1200 datasets for which one or more of the scanned markers were identified as QTL, meaning that the test statistic exceeded the genome-wide significance threshold at a given significance level. We generated 6000 datasets to estimate significance thresholds for each of the four methods and each set of parameter values.

The data were analyzed with polygenic variation either being ignored or being accounted for. If polygenic variation was ignored, the model to analyze the data was $y_i = \mu + x_i^* a^* + z_i^* d^* + \epsilon_i, i = 1, 2, \cdots, n$; this was model (1) without the random polygenic effect.

Statistical power In new sets of simulations, a QTL was placed in the middle of the first chromosome. The QTL had an additive effect 0.4 and a dominance effect 0.1. The QTL accounted for approximately 2.8%, 2.3%, or 1.6% of the total variance, corresponding to $\sigma = 0.7$, 1 or 1.5. Again, the genome scan for QTL under the null hypothesis of no QTL was performed on the first ten chromosomes. A QTL was identified if the test statistic at any of the scanning loci exceeded the genome-wide threshold at a given significance level. For each of the four methods and each set of parameter values, the power was estimated by the proportion of 1200 simulations where a QTL was identified. The threshold was estimated in the same way as for type I error rates.

Pooling Procedure

In practice when we have one dataset, we can permute the data N times to estimate a threshold for the test statistic. When we replicate a simulation K times, the test statistic in all the replicates follows the same distribution. Therefore, we only need one threshold for all the replicates. Suppose we permute the data N_i times in the i-th replicate simulation and get $S_i = \{x_{ij}, j = 1, 2, \cdots, N_i\}$, $i = 1, 2, \cdots, K$. Then

$$E\{\frac{\sum_{i=1}^{K}\sum_{j=1}^{N_i} I_{x_{ij}>x}}{\sum_{i=1}^{K} N_i}\} = \frac{\sum_{i=1}^{K} \alpha N_i}{\sum_{i=1}^{K} N_i} = \alpha$$

where x is the $100(1-\alpha)$th percentile of S_i and $I_{x_{ij}>x} = 1$ if $x_{ij} > x$ or 0 otherwise. This means that we can pool S_i $(i = 1, 2, \cdots, K)$ to estimate the threshold for the test statistic in all the replicate simulations.

Computational Approximation

In general there is no analytical solution to maximum likelihood estimates (MLE) for model (1). Genome scans are extremely computationally intensive and sometimes impractical without computational simplification. Note that the random effect u in model (1) is only used to control background genetic variation. A reasonable approximation will be good enough. Assume in equation (2) $g_{ij}^{(a)}\sigma_a^2 + g_{ij}^{(d)}\sigma_d^2 + g_{ij}^{(ad)}Cov(a,d) + g_{ij}^{(h)}\sigma_h^2 + g_{ij}^{(m)}\mu_h^2 = (g_{ij}^{(a)}c_1 + g_{ij}^{(d)}c_2 + g_{ij}^{(ad)}c_3 + g_{ij}^{(h)}c_4 + g_{ij}^{(m)}c_5)\sigma^2$. Then the variance-covariance matrix of $\boldsymbol{y}$ is $\boldsymbol{\Sigma} = (\boldsymbol{G}^{(a)}c_1 + \boldsymbol{G}^{(d)}c_2 + \boldsymbol{G}^{(ad)}c_3 + \boldsymbol{G}^{(h)}c_4 + \boldsymbol{G}^{(m)}c_5 + \boldsymbol{I})\sigma^2$ where $\boldsymbol{G}^{(a)} = (g_{ij}^{(a)})$ etc. If c's are known, then $\frac{1}{\sigma^2}\boldsymbol{\Sigma}$ is a known matrix and an analytical MLE solution exists. In applications, c's are unknown; however, we can estimate them under the null hypothesis and use the estimates as known values. Approximating random effects by their estimates is a known strategy in mixed-effect model models (Pinheiro and Bates, 2000) and works well in our situation.

Computational Efficiency

The permutation test as well as the other three methods is computationally intensive, which is a trade-off between reliability and computation. However, the computation is still manageable with the previous computational approximation even if there are thousands of markers. In our simulations, there were 1010 SNP markers and the sample size was 576; one genome scan took only a few seconds on a conventional desktop computer. Parallel computing can make it realistic to perform permutation tests even when there are hundreds of thousands of SNP markers.

References

Abney, M., M. S. McPeek, and C. Ober (2000). Estimation of variance components of quantitative traits in inbred populations. *Am. J. Hum. Genet. 141*, 629--650.

Abney, M., C. Ober, and M. S. McPeek (2002). Quantitative-trait homozygosity and association mapping and empirical genome-wide significance in large, complex pedigrees: fasting serum-insulin level in the hutterites. *Am. J. Hum. Genet. 70*, 920--934.

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Can. J. Fish. Aquat. Sci. 58*, 626--639.

Anderson, M. J. and C. J. F. T. Braak (2003). Permutation tests for multi-factorial analysis of variance. *J. Stat. Comput. Simul. 73*, 85--113.

Cheng, R., J. E. Lim, K. E. Samocha, G. Sokoloff, M. Abney, A. D. Skol, and A. A. Palmer (2010). Genome-wide association studies and the problem of relatedness among advanced intercross lines and other highly recombinant populations. *Genetics 185*, 1033--1044.

Efron, B. and R. Tibishirani (1993). *An introduction to the bootstrap*. Chpman & Hall, Inc.

Jackquard, A. (1974). *The genetics structure of populations*. Springer-Verlag, NY.

Jung, J., D. E. Weeks, and E. Feingold (2006). Gene-dropping vs. empirical variance estimation for allele-sharing linkage statistics. *Genet. Epidem. 30*, 652--665.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin (2008). Efficient control of population structure in model organism association mapping. *Genetics 178*, 1709--1723.

Lynch, M. and B. Walsh (1998). *Genetics and analysis of quantitative traits*, Volume 5. Sinauer Associates, Inc.

MacCluer, J. W., J. L. VandeBerg, B. Read, and O. A. Ryder (1986). Pedigree analysis by computer simulation. *Zoo Biology 5*, 147--160.

Nichols, T. E. and A. P. Holmes (2001). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping 15*, 1--25.

O'Gorman, T. W. (2005). The performance of randomization tests that use permutations of independent variables. *Commun. Stat. Sim. Comput. 34*, 895--908.

Pardo, L. M., I. MacKay, B. Oostra, C. M. van Duijin, and Y. S. Aulchenko (2005). The effect of genetic drift in a young genetically isolated population. *Ann. Hum. Genet. 69*, 288--295.

Peirce, J. L., , K. W. Broman, L. Lu, E. J. Chesler, G. Zhou, D. C. Airey, A. E. Birmingham, and R. W. Williams (2008). Genome reshuffling for advanced intercross permutation (GRAIP): simulation and permutation for advanced intercross population analysis. *PLoS ONE 3*(4), e1977.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-effects models in S and S-PLUS*. Springer-Verlag, New York.

Suwanlee, S., R. Baumung, J. Sl̈kner, and I. Curik (2007). Evaluation of ancestral inbreeding coefficients: Ballou's formula versus gene dropping. *Conserv. Genet. 8*, 489--495.

Thomas, A. (1990). Comparison of an exact and a simulation method for calculating gene extinction probabilities in pedigrees. *Zoo Biology 9*, 259--274.

Valdar, W., C. C. Holmes, R. Mott, and J. Flint (2009). Mapping in structured populations by resample model averaging. *Genetics 182*, 1263--1277.

Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet. 38*, 203--208.

Zou, F., J. A. L. Gelfond, D. C. Airey, L. Lu, K. F. Manly, R. W. Williams, and D. W. Threadgill (2005). Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. *Genetics 170*, 1299--1311.

Table S1 Summary of Relatedness[a]

	Min.	1st Qu.	Median	3rd Qu.	Max.
$g_{ij}^{(a)}$	0.76810	0.77070	0.77180	0.77480	1.42500
$g_{ij}^{(d)}$	0.13590	0.15480	0.15540	0.15590	0.61600
$g_{ij}^{(ad)}$	0.66420	0.66930	0.67160	0.67640	1.69800
$g_{ij}^{(h)}$	0.07814	0.07917	0.07964	0.08059	0.42460
$g_{ij}^{(m)}$	0.00714	0.00816	0.00821	0.00830	0.24430

[a] Defined in equation (2) among the simulated F_{26} individuals. The different levels of relatedness means that the assumption of exchangeability is incorrect.

Table S2 Type I Error Rates[a]

Distr[b]	Method[c]	$\sigma = 0.7$			$\sigma = 1$			$\sigma = 1.5$		
		$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
Exp	Permut	0.0933	0.0375**	0.0050**	0.0925	0.0475	0.0075	0.1108	0.0583	0.0133
	Bootstr	0.0925	0.0408	0.0050**	0.0817**	0.0375**	0.0075	0.1092	0.0508	0.0117
	GeneDr	0.1017	0.0442	0.0050**	0.0883	0.0475	0.0075	0.1083	0.0558	0.0125
	GRAIP	0.0925	0.0442	0.0075	0.0875	0.0425	0.0075	0.1125	0.0583	0.0158*
Norm	Permut	0.1100	0.0525	0.0108	0.1058	0.0475	0.0100	0.0958	0.0475	0.0083
	Bootstr	0.1067	0.0517	0.0108	0.0958	0.0408	0.0100	0.0958	0.0442	0.0058*
	GeneDr	0.1000	0.0525	0.0133	0.0958	0.0417	0.0100	0.0958	0.0467	0.0075
	GRAIP	0.0967	0.0525	0.0117	0.0958	0.0450	0.0100	0.1058	0.0450	0.0058
Unif	Permut	0.0908	0.0408	0.0092	0.0942	0.0483	0.0125	0.0917	0.0517	0.0108
	Bootstr	0.0892	0.0400*	0.0083	0.0950	0.0467	0.0117	0.0992	0.0558	0.0125
	GeneDr	0.0908	0.0400*	0.0083	0.0925	0.0467	0.0117	0.0883	0.0517	0.0125
	GRAIP	0.0908	0.0467	0.0092	0.0950	0.0467	0.0125	0.0958	0.0525	0.0125

[a] Estimated from 1200 simulations at genome-wide significance levels $\alpha =$ 0.10, 0.05 and 0.01. Symbol *, ** or *** indicates the estimated type I error rate is significantly different from the expected level at significance level 0.10, 0.05 or 0.01.

[b] Permuting genotypic data (Permut), bootstrapping phenotypic data (Bootstr), gene dropping (GeneDr) or GRAIP.

[c] The distribution of the residual was exponential (Exp), normal (Norm) or uniform (Unif), each with a standard deviation 0.7, 1 or 1.5.

Table S3 Estimated Genome-wide Thresholds for the Body Weight Data

	Relatedness Ignored			Relatedness Not Ignored		
α level	0.1	0.05	0.01	0.1	0.05	0.01
Permut	19.45	21.09	24.52	18.70	20.23	23.56
Bootstr	19.49	21.01	24.25	19.49	21.00	24.20
GeneDr	65.17	70.46	84.48	19.53	21.08	24.45
GRAIP	57.69	62.20	72.67	19.72	21.26	24.50

Estimated from 5000 simulations at genome-wide significance levels $\alpha = 0.1$, 0.05 and 0.01 by the following methods: permuting genotypic data (Permut), bootstrapping phenotypic data (Bootstr), gene dropping (GeneDr) and GRAIP, using the likelihood ratio test (LRT).

Table S4 P-values by the Kolmogorov-Smirnov Test

	Permut	Bootstr	GeneDr	GRAIP
$\sigma = 0.7$	0.60200	0.32428	0.00000	0.00000
$\sigma = 1$	0.44558	0.44988	0.00000	0.00000
$\sigma = 1.5$	0.43282	0.10871	0.00000	0.00000

Based on 6000 simulations under the null hypothesis that when no QTL effects existed, the distribution estimated by a testing method when relatedness was ignored was identical to the distribution estimated by the same method when relatedness was taken into account. Data was generated by each of the testing methods: permuting genotypic data (Permut), bootstrapping phenotypic data (Bootstr), gene dropping (GeneDr) and GRAIP. The distribution of the residual was normal with a standard deviation 0.7, 1 or 1.5.

Supporting Data and R Scripts

Available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.146332/-/DC1.

File S2   R Scripts
File S3   Raw Data