Genetic Epidemiology 33: 710–716 (2009)

A Note on Permutation Tests for Genetic Association Analysis of Quantitative Traits When Variances Are Heterogeneous

Danielle Posthuma,[1–3]* Dirk-Jan de Koning,[4] Conor Dolan,[5] Michael E. Goddard,[6,7] and Peter M. Visscher[8]

[1]*Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands*
[2]*Department of Medical Genomics, Vrije Universiteit Medical Centre, Amsterdam, The Netherlands*
[3]*Department of Functional Genomics, Vrije Universiteit, Amsterdam, The Netherlands*
[4]*Division of Genetics and Genomics, Roslin Institute and R(D)SVS, University of Edinburgh, Roslin, UK*
[5]*Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands*
[6]*Department of Land and Food Resources, University of Melbourne, Melbourne, Victoria, Australia*
[7]*Department of Primary Industries, Victoria, Australia*
[8]*Queensland Statistical Genetics, Queensland Institute of Medical Research, Brisbane, Australia*

The genetic dissection of quantitative traits, or endophenotypes, usually involves genetic linkage or association analysis in pedigrees and subsequent fine mapping association analysis in the population. The ascertainment procedure for quantitative traits often results in unequal variance of observations. For example, some phenotypes may be clinically measured whilst others are from self-reports, or phenotypes may be the average of multiple measures but with the number of measurements varying. The resulting heterogeneity of variance poses no real problem for analysis, as long as it is properly modelled and thereby taken into account. However, if statistical significance is determined using an empirical permutation procedure, it is not obvious what the units of sampling are. We investigated a number of permutation approaches in a simulation study of an association analysis between a quantitative trait and a single nucleotide polymorphism. Our simulations were designed such that we knew the true *p*-value of the test statistics. A number of permutation methods were compared from the regression of true on empirical *p*-values and the precision of the empirical *p*-values. We show that the best procedure involves an implicit adjustment of the original data for the effects in the model before permutation, and that other methods, some of which seemed appropriate a priori, are relatively biased. *Genet. Epidemiol*. 33:710–716, 2009.

Key words: empirical *p*-value; permutation; WLS; OLS; genetic association

Contract grant sponsor: Netherlands Scientific Organization; Contract grant numbers: NWO 480-05-003; NWO-VIDI-016-065-318; Contract grant sponsor: Australian National Health and Medical Research Council; Contract grant sponsor: Royal Netherlands Academy of Arts and Sciences (KNAW).
*Correspondence to: Danielle Posthuma, Department of Biological Psychology, Vrije Universiteit, van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands. E-mail: danielle@psy.vu.nl
Received 8 September 2008; Accepted 22 February 2009
Published online 13 April 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20423

INTRODUCTION

A contentious theme in conducting whole genome linkage or association studies is the assessment of statistical significance [Curtis, 1996; Lander and Kruglyak, 1995; Storey and Tibshirani, 2003; Witte et al., 1996]. On the one hand, the rate of false positives needs to be controlled due to the inherently multiple testing nature of whole genome linkage and association analyses. On the other hand, whole genome linkage and association signals are expected to be small and are therefore easily overlooked (false negatives). A simple Bonferroni correction, for example, is too conservative if multiple tests are not independent. As map density increases, the number of tests goes to infinity and a Bonferroni correction is not appropriate. In addition, experimental factors that influence the amount of information extracted from the data set such as pedigree structure, the completeness and accuracy of genotype data, and marker density are expected to have substantial effects, both on the power of a study to detect true effects and on the significance of any finding. Several alternatives have been developed to determine the statistical significance of whole genome linkage or association results. These methods include locus counting methods [Wiltshire et al., 2002], Monte Carlo procedures [Lin and Zou, 2004], false discovery rates [Benjamini and Hochberg, 1995], bootstrap methods [Efron, 1979], and simulation procedures [Cheverud, 2001]. Although these methods are elegant and computationally efficient, they are often dependent on certain model assumptions and cannot incorporate experimental-specific characteristics (such as the pattern of missing data, genetic informativeness, multivariate traits, and structure of the pedigree) all of which can influence the distribution of the test statistics. Therefore, it is commonly agreed that accurate estimates of genome-wide statistical significance are best obtained empirically through permutation tests that condition on the observed data [Churchill

and Doerge, 1994; Douglas et al., 2000; Gordon et al., 2000; Hirschhorn et al., 2001; Ott, 1999; Sawcer et al., 1997].

The ascertainment procedure for quantitative traits often results in unequal variance of observations. For example, some phenotypes may be clinically measured whilst others are from self-reports [e.g., Macgregor et al., 2006], or phenotypes may be the average of multiple measures but with the number of measurements varying [e.g., Posthuma et al., 2006]. The resulting heterogeneity of variance poses no real problem for analysis, as long as it is properly modelled and thereby taken into account. For example, in a simple linear model of association analysis, a weighted-least-squares (WLS) analysis is efficient, with the weights inversely proportional to the variance of the observations. However, if statistical significance is determined using an empirical permutation procedure, it is not obvious what the units of sampling are. Should the variance of the observation be permuted with the phenotype, with the indicators of the effects in the model or with the residual? In this study, we considered a number of permutation methods for models in which WLS is efficient. We show which of the methods is best and why it performs better than others that appeared to be appropriate at first.

METHODS

MODELS

Consider a simple model for the additive association between a quantitative trait (y) and the number of alleles ($x_i = 0, 1, 2$) at a biallelic single nucleotide polymorphism (SNP); $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$,or, in matrix notation, $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$. If the errors have constant variance, for example, $\varepsilon \sim N(0,1)$, i.e., $\mathbf{e} \sim N(0, \mathbf{I})$, then the appropriate analysis is ordinary least squares (OLS) and a permutation strategy to test the hypothesis that $\beta_1 = 0$ is to permute x and y repeatedly and estimate β_1 for each sample using OLS for the model $y_i = \beta_0 + \beta_1 x_j + \varepsilon_i$, with subscripts i and j denoting that the sample is permuted. An empirical two-sided p-value can be calculated from the proportion of estimates of β_1 from the permuted samples that are larger than the estimate of β_1 from the original unpermuted sample. In a specific sample, given a sufficiently large number of permutations, the empirical p-value from permuted samples is (asymptotically) identical to the theoretical p-value obtained from comparing the test statistic from the original, i.e., unpermuted, sample to the theoretical distribution of the test statistic under the null hypothesis. In other words, provided all assumptions of the OLS model have been met, the expected difference between the theoretical p-value and empirical p-value is zero and its variance decreases with increasing numbers of permutations.

The predicted variance of the difference between the theoretical and empirical p-values can be easily derived. Given a true value of p, the variance of the estimated p-value from m permutations is $p(1-p)/m$, which follows from the binomial distribution. Over replicate data sets, the true p-values are uniformly distributed, if there is no association between x and y. Therefore, the expected variance of the difference between the theoretical (true) p-value and the p-value from permulations is

$$\int_0^1 \left(\frac{p(1-p)}{m}\right) \mathrm{d}p = \cdots = \frac{1}{6m}.$$

Hence the SD of $p - p_{\mathrm{emp}}$ is $1/\sqrt{6m}$, with p_{emp} representing the empirical p-value. For example, given m = 100, 500, 1,000, and 5,000, the expected SD is 0.0405, 0.0182, 0.0128, and 0.0057, respectively.

Now consider the same model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

but with $\varepsilon_i \sim N(0, v_i)$, or, in matrix notation, $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, $\mathbf{e} \sim N(0, \mathbf{V})$. We assume here that $\mathbf{V}$ is diagonal. Note that the subscript on the variance of ε_i indicates that the residual variance may vary over the level of x_i, i.e., the regression may display heteroskedasticity. Given this, the appropriate analysis is weighted (or general) least squares, with $\mathbf{b} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, as opposed to $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, in the homoskedastic case mentioned above. The question is how to perform a permutation test to obtain the appropriate empirical p-value. The key to any successful permutation procedure is that (i) it should condition on the ''design'' of the experiment and (ii) should mimic the simulation of y-values under the null hypothesis.

METHOD 1

In Method 1, y_i and v_i are permuted over x_i and WLS is performed on the same model as the original model. Thus, we keep y_i and its conditional variance v_i together. This permutation approach breaks up any association between certain genotypic (x) values and the variance (weight) that went with it in the original data, hence the ''design'' element of the experiment is not strictly maintained.

METHOD 2

On the original data, consider an equivalent model, $\mathbf{Ty} = \mathbf{TXb} + \mathbf{Te}$, or $\mathbf{y}^* = \mathbf{X}^*\mathbf{b} + \mathbf{e}^*$, with $(\mathbf{T}'\mathbf{T})^{-1} = \mathbf{V}$. Letting t_{ii} denote a diagonal element of $\mathbf{T}$, we have $t_{ii} = 1/\sqrt{v_i}$. In scalar notation,

$$y_i^* = \beta_0 x_{0(i)} + \beta_1 x_{1(i)}^* + e_i^*$$
$$\text{with } y_i^* = t_{ii} y_i, x_{0(i)} = t_{ii}, x_{1(i)}^* = t_{ii} x_i^*.$$

The transformed data can now be analyzed with OLS because the variance of the residuals is constant. To do the permutation analysis, we permute y^* and x_0 over x_1, and analyze each permuted sample with OLS. This permutation approach is different from Method 1 because it attempts to condition on both the x-values *and* the weights that went with those x-values in the original data. However, it does not fully capture the distribution under the null hypothesis of no association between x and y because the new y^*-value—which is permuted—contains an effect due to the $t_i * \beta_i$ and this is inappropriately attached to a different record after permutation.

METHOD 3

An intuitive method is to adjust the original y-values by the estimates of the mean and regression coefficient, and then to permute residuals $\hat{e}$ over x_1, using WLS, or permute $\hat{e}/\sqrt{\sigma_{\hat{e}}^2}$ over x_1, using OLS. Since we work with residuals, the relationship between the residual and x is zero, before and after a permutation. Therefore, the following algorithm was tested:

We first fit $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $\sigma_{e_i}^2 = v_i$, i.e., fit the standard WLS. The WLS solutions are called $\hat{\beta}_0$ and $\hat{\beta}_1$ (step 1). We then calculate $\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ (step 2).

Hence, residuals are calculated by adjusting for the fixed effects. The second step can also be done by only adjusting for the mean, i.e., $\hat{e}_i = y_i - \hat{\beta}_0$. This would be more convenient if there were many correlated x variables (as in, for example, a genome-wide scan), and does not seem to have an impact on the behavior of the permutation test (results not shown). If the null hypothesis is true (i.e., x and y are unrelated), then adjusting for βx is simply adding noise. If there is a real effect of x on y (i.e., β_1 is not zero) then not adjusting would result in greater residual variance in the permuted samples than in the original data, but this does not seem to have an effect on the p-value (results not shown). We then continue with step 3: Permuting $\hat{e}$ against x and v, after which we calculate $y_j^* = \hat{\beta}_0 + \hat{e}_j\sqrt{v_i/v_j}$, where v_i is the variance pertaining to x_i and v_j is the variance applicable to the $\hat{e}_j$ (step 4). Finally, we fitted $y_j^* = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $\sigma^2_{\hat{e}_i} = v_i$, i.e., the original model (step 5).

This procedure keeps the design constant (x and v together) and creates y-variates under the null hypothesis of no association between y and x.

SIMULATIONS

We simulated data under the null hypothesis of no association between a quantitative trait and a biallelic SNP, with allele frequencies $p = 0.5$ and $q = 1-p$. The genotypic coefficients (x_i; −1, 0, or 1) corresponding to the three possible genotypes were sampled such that their frequencies followed p^2, $2pq$, and q^2. We also simulated data with a continuous x variable by sampling x_i from the uniform distribution, i.e., $x_i \sim U(0,1)$.

We simulated the data by letting the dependent variable (y) to be a function of

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where β_0 is the grand mean (fixed at 100), β_1 is the regression of the genotype on the trait, which was fixed at zero under the null hypothesis, x_i is an individual's genotype (−1, 0, or 1) or a continuous variable, and ε_i is a residual distributed normally with variance which is a function of $z\sqrt{w}$, whereby $z \sim N(0,1)$ and the weights w are drawn from a discrete uniform distribution with

TABLE I. Results for the empirical p-values obtained from simulations according to Methods 1–3, with x_i-values −1, 0, and 1, sampled from a genotypic distribution assuming a biallelic gene and $p = 0.5$

				Method 1		Method 2		Method 3	
nperm	ncases		True p	p	p−true p	p	p−true p	p	p−true p
100	100	Minimum	0.000	0.000	−0.172	0.000	−0.177	0.000	−0.163
		Maximum	1.000	1.000	0.163	1.000	0.150	1.000	0.162
		Mean	0.503	0.503	0.000	0.503	0.000	0.502	−0.001
		Std. deviation	0.288	0.292	0.045	0.291	0.041	0.291	0.041
	5,000	Minimum	0.001	0.000	−0.192	0.000	−0.142	0.000	−0.176
		Maximum	1.000	1.000	0.159	1.000	0.154	1.000	0.156
		Mean	0.503	0.502	0.000	0.502	−0.001	0.503	0.000
		Std. deviation	0.287	0.290	0.041	0.290	0.040	0.290	0.040
500	100	Minimum	0.000	0.000	−0.098	0.000	−0.086	0.000	−0.080
		Maximum	1.000	1.000	0.109	1.000	0.080	1.000	0.075
		Mean	0.497	0.497	0.000	0.498	0.000	0.497	0.000
		Std. deviation	0.287	0.287	0.027	0.288	0.019	0.288	0.019
	5,000	Minimum	0.000	0.000	−0.077	0.000	−0.079	0.000	−0.073
		Maximum	1.000	1.000	0.070	1.000	0.068	1.000	0.080
		Mean	0.498	0.498	0.000	0.498	0.000	0.499	0.000
		Std. deviation	0.289	0.290	0.019	0.289	0.018	0.290	0.018
1,000	100	Minimum	0.000	0.000	−0.119	0.000	−0.052	0.001	−0.050
		Maximum	1.000	1.000	0.121	1.000	0.054	1.000	0.054
		Mean	0.500	0.500	0.000	0.500	0.000	0.499	0.000
		Std. deviation	0.290	0.290	0.024	0.290	0.013	0.290	0.013
	5,000	Minimum	0.000	0.000	−0.062	0.000	−0.055	0.000	−0.055
		Maximum	1.000	1.000	0.053	1.000	0.056	1.000	0.054
		Mean	0.496	0.496	0.000	0.496	0.000	0.496	0.000
		Std. deviation	0.291	0.292	0.013	0.292	0.013	0.292	0.013
5,000	100	Minimum	0.000	0.000	−0.081	0.000	−0.021	0.000	−0.020
		Maximum	1.000	1.000	0.094	1.000	0.024	1.000	0.021
		Mean	0.496	0.496	0.000	0.496	0.000	0.496	0.000
		Std. deviation	0.288	0.288	0.020	0.288	0.006	0.288	0.006
	5,000	Minimum	0.000	0.000	−0.028	0.000	−0.022	0.000	−0.023
		Maximum	1.000	1.000	0.026	1.000	0.021	1.000	0.030
		Mean	0.502	0.502	0.000	0.502	0.000	0.502	0.000
		Std. deviation	0.286	0.286	0.006	0.286	0.006	0.287	0.006

Note: For readability only situations with ncases = 100 or 5,000 are shown.

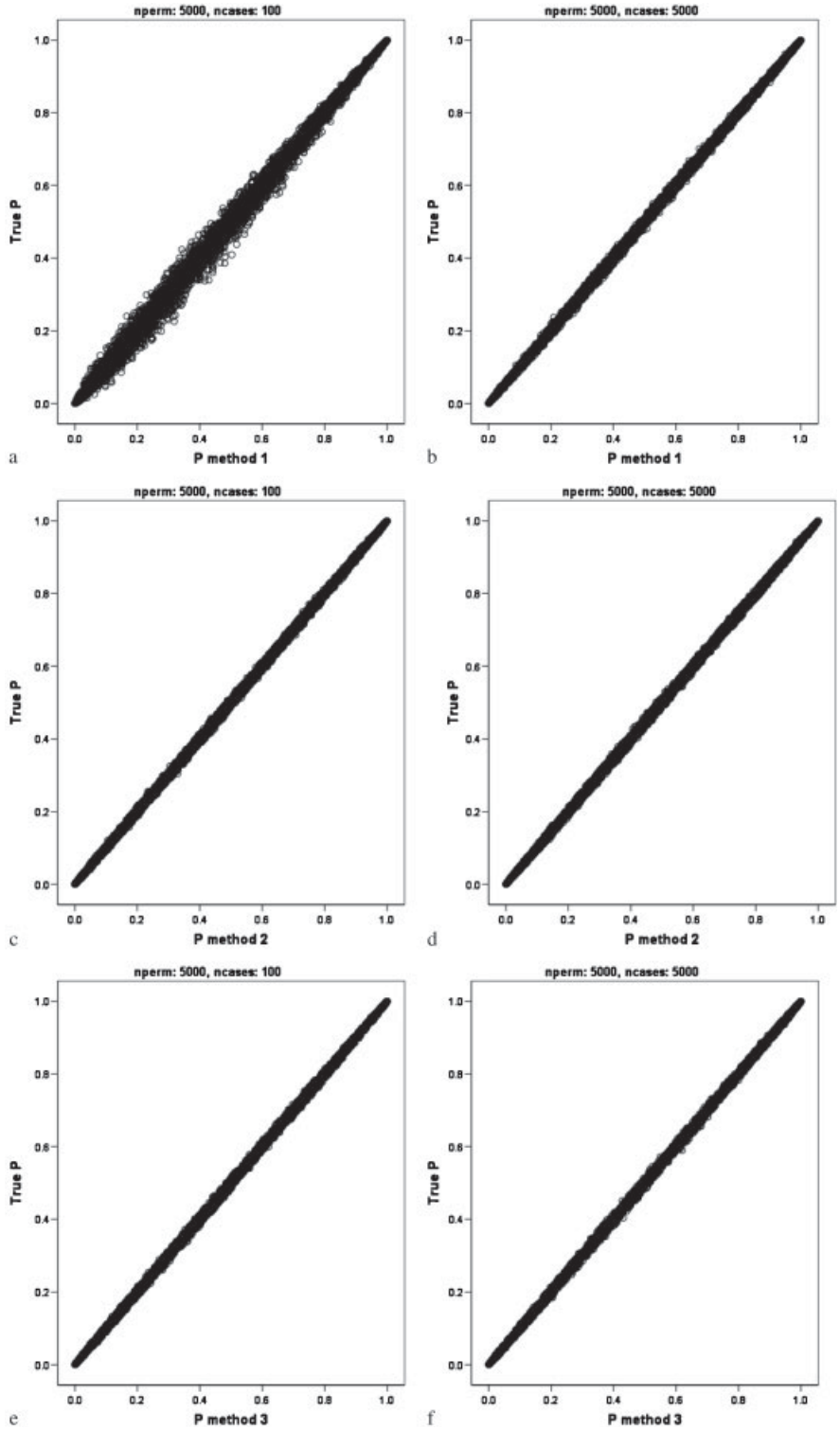

Fig. 1. Scatterplots of empirical *p*-values vs. true *p*-values for Methods 1–3, including 5,000 replicates, 5,000 permutations, and 100 (a, c, e) or 5,000 (b, d, f) cases, when sampling x_i from a genotypic distribution with a biallelic gene and $p = 0.5$.

values from 1 to 10, i.e., $w \sim U(1, 10)$. The variance of the trait is thus proportional to the square root of the weights.

For each simulated data set a WLS analysis was carried out after which the data set was permuted to obtain empirical *p*-values, using the three methods previously described. For these analyses, the sample size (n) was 100, 200, 500, 1,000, or 5,000. The number of permutations (m) was 100, 500, 1,000, or 5,000. A total of 5,000 replicates

were simulated. For each simulated replicate, the *p*-value was obtained from an *F* table. All simulations were carried out using the R software package (http://www.r-project.org/) and were run on the Genetic Cluster Computer (http://www.geneticcluster.org).

RESULTS

Results for the simulations where x_i-values are sampled from a genotypic distribution are provided in Table I and Figure 1a–f. Results for the simulations in which x_i-values are sampled from a continuous normal distribution are provided in Table II and Figure 2a–f.

METHOD 1

The results of the simulations show that this method is unbiased (i.e., the regression of empirical on theoretical *p*-values is unity), but has random error around the prediction of empirical *p*-values, even for large numbers of permutations. Hence, for a given sample, and thus for a given fixed set of *x*-values, there is a ''prediction error variance'' that does not seem to reduce to zero with an increasing number of permutations. This is because the requirements for the permutation test (constant design and sample under the null hypothesis) are not met. See Figure 1a, b. That is, *y*-values that contain a residual with large variance become attached, after permutation, to a record with supposedly small residual variance.

METHOD 2

The simulations show that if x_1 is grouped (we looked at three groups with $x_1 = -1$, 0, or 1 with probabilities of 0.25, 0.50, and 0.25, respectively) this method works better than Method 1. That is, the method was also unbiased, and gave a smaller error around the regression line of empirical *p*-values on theoretical *p*-values. However, when we simulated x_1 from a uniform $(0, 1)$ distribution (see Table II and Figure 2a–f), i.e., all x_1 are different, then the method performed worse in that it produced larger

TABLE II. Results for the empirical *p*-values obtained from simulations according to Methods 1–3, with continuous x_i-values sampled from a uniform distribution (0, 1)

				Method 1		Method 2		Method 3	
nperm	ncases		True *p*	*p*	*p*−true *p*	*p*	*p*−true *p*	*p*	*p*−true *p*
100	100	Minimum	0.000	0.000	−0.190	0.000	−0.180	0.000	−0.160
		Maximum	1.000	1.000	0.180	1.000	0.360	1.000	0.150
		Mean	0.503	0.504	0.001	0.504	0.001	0.503	0.000
		Std. deviation	0.288	0.290	0.046	0.294	0.061	0.291	0.041
	5,000	Minimum	0.001	0.000	−0.170	0.000	−0.180	0.000	−0.180
		Maximum	1.000	1.000	0.170	1.000	0.480	1.000	0.190
		Mean	0.496	0.496	0.000	0.495	−0.001	0.496	−0.001
		Std. deviation	0.291	0.293	0.041	0.297	0.060	0.294	0.040
500	100	Minimum	0.000	0.000	−0.110	0.000	−0.100	0.000	−0.080
		Maximum	1.000	1.000	0.130	1.000	0.490	1.000	0.070
		Mean	0.502	0.503	0.001	0.502	0.000	0.501	−0.001
		Std. deviation	0.291	0.291	0.027	0.294	0.049	0.291	0.018
	5,000	Minimum	0.001	0.000	−0.070	0.000	−0.120	0.000	−0.070
		Maximum	1.000	1.000	0.070	1.000	0.350	1.000	0.070
		Mean	0.498	0.497	0.000	0.497	0.000	0.498	0.000
		Std. deviation	0.288	0.289	0.018	0.292	0.049	0.289	0.018
1,000	100	Minimum	0.000	0.000	−0.100	0.000	−0.090	0.000	−0.050
		Maximum	1.000	1.000	0.120	1.000	0.430	1.000	0.050
		Mean	0.497	0.497	0.000	0.497	0.000	0.498	0.000
		Std. deviation	0.286	0.287	0.024	0.289	0.047	0.287	0.013
	5,000	Minimum	0.000	0.000	−0.050	0.000	−0.090	0.000	−0.050
		Maximum	0.999	1.000	0.050	1.000	0.370	1.000	0.060
		Mean	0.495	0.495	0.000	0.495	0.000	0.494	0.000
		Std. deviation	0.288	0.288	0.013	0.292	0.047	0.288	0.013
5,000	100	Minimum	0.000	0.000	−0.080	0.000	−0.080	0.000	−0.030
		Maximum	1.000	1.000	0.100	1.000	0.350	1.000	0.020
		Mean	0.497	0.497	0.000	0.497	0.000	0.497	0.000
		Std. deviation	0.292	0.292	0.021	0.295	0.045	0.292	0.006
	5,000	Minimum	0.000	0.000	−0.030	0.000	−0.060	0.000	−0.020
		Maximum	0.999	1.000	0.020	1.000	0.390	1.000	0.020
		Mean	0.499	0.499	0.000	0.499	0.000	0.499	0.000
		Std. deviation	0.291	0.291	0.006	0.294	0.044	0.291	0.006

Note: For readability only situations with ncases = 100 or 5,000 are shown.

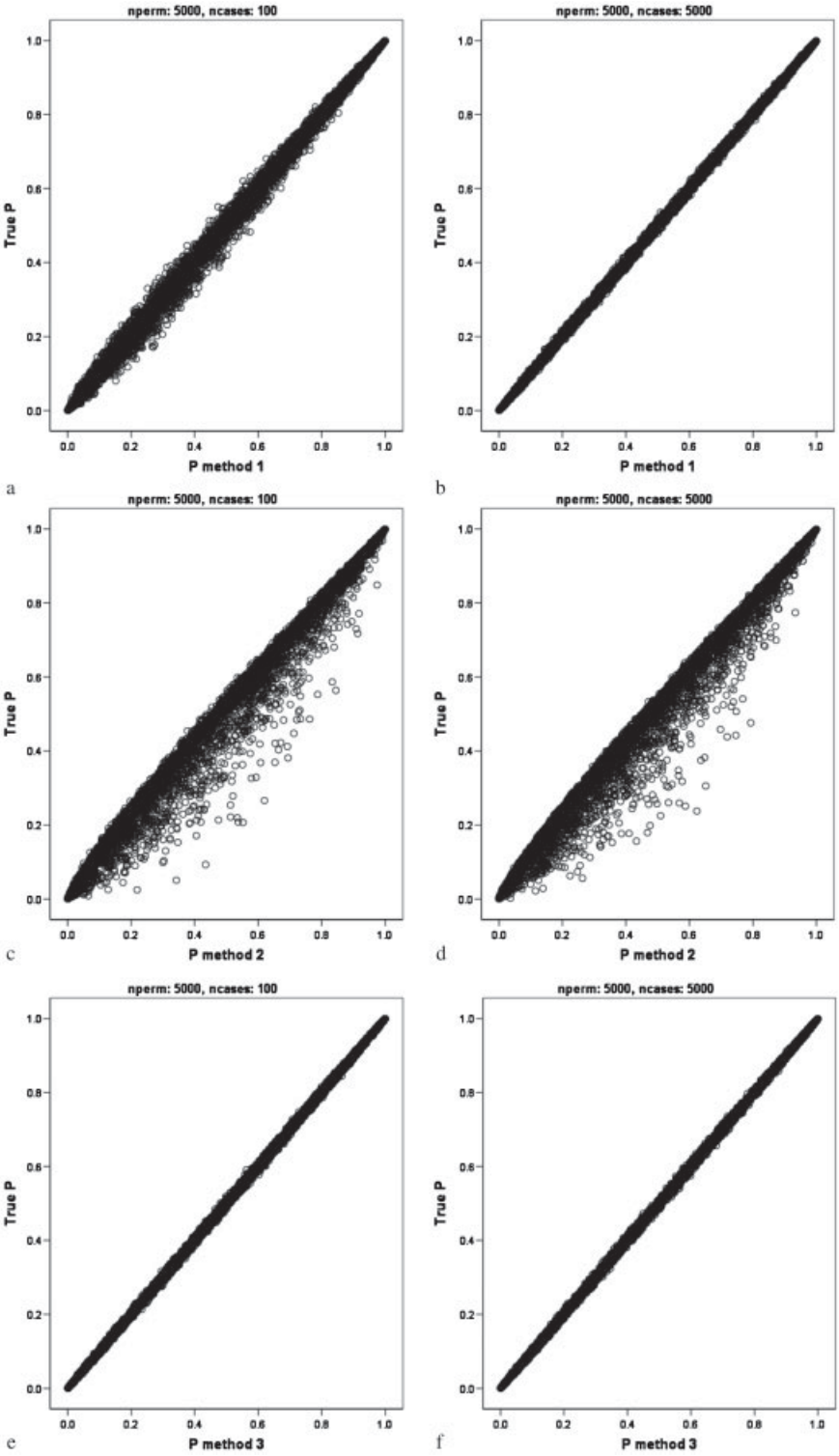

Fig. 2. Scatterplots of empirical *p*-values vs. true *p*-values for Methods 1–3, including 5,000 replicates, 5,000 permutations, and 100 (a, c, e) or 5,000 (b, d, f) cases, when sampling x_i from a uniform distribution.

error around the regression line. Not only that, but this method seems sensitive to misspecification of the variance: when we simulated error terms from a χ^2 instead of a normal, but ignored this in the analyses, then Method 2 gave biased and inaccurate results, whereas Method 1 remained unbiased. This suggests that this method also

does not fully condition on the experimental design. As pointed out earlier, the deficiencies of Method 2 arise because the new y^*-value, which is permuted, contains an effect due to the $t_i * \beta_i$ and this is inappropriately attached to a different record after permutation.

In Method 1, the average variance of the regression slope from permuted samples is approximately the same as the variance of the estimate of the slope from the simulations. That is, the sampling variance from the permuted samples reflects the sampling process of the simulated samples. In Method 2 this is not the case, and the variance of the estimates of the slope from permutations is less than the variance of the estimates of the slope from simulations. This suggests again that Method 2 does not properly condition on the design.

METHOD 3

Results for Method 3 are given in Figures 1e, f and 2e, f. Method 3 is the ''best'' method in that it has the desirable properties of being unbiased and having the smallest prediction error. For the biallelic simulated marker, Methods 2 and 3 are indistinguishable (Table I). However, when the x variable was sampled from a continuous, Method 3 had smaller prediction error variance when the number of permutations was 100 or 500 (Table II).

DISCUSSION

The ascertainment procedure for quantitative traits often results in heterogeneity of variance. Although this can be properly modelled and poses no real problem for analysis, it may complicate evaluation of empirical significance as it is not obvious what the units of sampling are. We compared three different permutation procedures to investigate which method has the most desirable properties when variances are heterogeneous. We have shown that the best permutation analysis when variances are heterogeneous is the one in which the data are adjusted for the effects of the model and residuals are permuted against the original fixed effects structure and variance structure. The aim of a permutation test is to retain the design of the experiment, including any heterogeneity of variance, and to simulate y-values under the null hypothesis. Method 3 does this by permuting the residuals but rescaling them to match the record to which they are now attached.

ACKNOWLEDGMENTS

The statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org) which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003). D.P. is supported by NWO-VIDI-016-065-318. P.M.V. is supported by the Australian National Health and Medical Research Council. Collaboration between D.P. and P.M.V. was supported through a Visiting Professorship grant from the Royal Netherlands Academy of Arts and Sciences (KNAW).

REFERENCES

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300.

Cheverud JM. 2001. A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87:52–58.

Churchill GA, Doerge RW. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963–971.

Curtis D. 1996. Genetic dissection of complex traits. Nat Genet 12:356–358.

Douglas JA, Boehnke M, Lange K. 2000. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. Am J Hum Genet 66:1287–1297.

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

Gordon D, Leal SM, Heath SC, Ott J. 2000. An analytic solution to single nucleotide polymorphism error-detection rates in nuclear families: implications for study design. Pac Symp Biocomput 5:663–674.

Hirschhorn JN, Lindgren CM, Daly MJ, Kirby A, Schaffner SF, Burtt NP, Altshuler D, Parker A, Rioux JD, Platko J, Gaudet D, Hudson TJ, Groop LC, Lander ES. 2001. Genomewide linkage analysis of stature in multiple populations reveals several regions with evidence of linkage to adult height. Am J Hum Genet 69:106–116.

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

Lin DY, Zou F. 2004. Assessing genomewide statistical significance in linkage studies. Genet Epidemiol 27:202–214.

Macgregor S, Cornes BK, Martin NG, Visscher PM. 2006. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. Hum Genet 120:571–580.

Ott J. 1999. Analysis of Human Genetic Linkage. Baltimore: The Johns Hopkins University Press.

Posthuma D, Visscher PM, Willemsen G, Zhu G, Martin NG, Slagboom PE, de Geus EJ, Boomsma DI. 2006. Replicated linkage for eye color on 15q using comparative ratings of sibling pairs. Behav Genet 36:12–17.

Sawcer S, Jones HB, Judge D, Visser F, Compston A, Goodfellow PN, Clayton D. 1997. Empirical genomewide significance levels established by whole genome simulations. Genet Epidemiol 14:223–229.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440–9445.

Wiltshire S, Cardon LR, McCarthy MI. 2002. Evaluating the results of genomewide linkage scans of complex traits by locus counting. Am J Hum Genet 71:1175–1182.

Witte JS, Elston RC, Schork NJ. 1996. Genetic dissection of complex traits. Nat Genet 12:355–356.