Copyright © 2008 by the Genetics Society of America
DOI: 10.1534/genetics.108.088005

Distributions of Hardy–Weinberg Equilibrium Test Statistics

R. V. Rohlfs*,[1] and B. S. Weir†

**Department of Genome Sciences, University of Washington, Seattle, Washington 98195-5065 and †Department of Biostatistics, University of Washington, Seattle, Washington 98195-7232*

Manuscript received February 11, 2008
Accepted for publication September 10, 2008

ABSTRACT

It is well established that test statistics and *P*-values derived from discrete data, such as genetic markers, are also discrete. In most genetic applications, the null distribution for a discrete test statistic is approximated with a continuous distribution, but this approximation may not be reasonable. In some cases using the continuous approximation for the expected null distribution may cause truly null test statistics to appear nonnull. We explore the implications of using continuous distributions to approximate the discrete distributions of Hardy–Weinberg equilibrium test statistics and *P*-values. We derive exact *P*-value distributions under the null and alternative hypotheses, enabling a more accurate analysis than is possible with continuous approximations. We apply these methods to biological data and find that using continuous distribution theory with exact tests may underestimate the extent of Hardy–Weinberg disequilibrium in a sample. The implications may be most important for the widespread use of whole-genome case–control association studies and Hardy–Weinberg equilibrium (HWE) testing for data quality control.

MOST analyses of genetic data rely on discrete genetic markers such as single-nucleotide polymorphisms (SNPs), copy number variants (CNVs), or microsatellites, yet most analyses use statistical theory based on continuous distributions such as the normal or chi square. In some cases, use of these theories is satisfactory but in most contemporary genetic analyses there is a need for care, especially with the reported *P*-values for hypothesis tests. The need may be most urgent for the widespread use of whole-genome case–control association studies and Hardy–Weinberg equilibrium (HWE) testing for data quality control.

The issue of approximating discrete distributions with continuous functions has been discussed in the literature. Yates (1934) applied a simple "continuity correction" for goodness-of-fit tests and 50 years later stressed that this made the chi-square test statistic a better approximation to the exact test for 2×2 contingency tables (Yates 1984). However, the correction does not alter the fact that test statistics for discrete data have discrete distributions. One prominent issue raised by discrete test statistics is discrete type I error rates; such a type I error rate has a discrete number of possible values. Tocher (1950) described a stochastic hypothesis rejection method that allows any chosen type I error rate to be achieved when working with discrete test statistics. Innan *et al.* (2005) proposed a similar random procedure for the specific case of the haplotype configuration test. These methods effectively correct the rejection region of discrete statistics; however, the methods are seldom applied, likely because their stochastic nature is unappealing to scientists.

Some other discussions of genetic test statistics have been cognizant of the discrete nature of test statistics and corresponding *P*-values (Slatkin 1994; Raymond and Rousset 1995; Rousset and Raymond 1995; Innan *et al.* 2005). However, the implications of using continuous distribution theory with discrete *P*-values have not been sufficiently discussed. Today testing is done by computer and computational issues are of less importance than in the past, making it possible to evaluate the actual discrete *P*-value distributions and use those for inference. We are particularly interested in how data discreteness affects the null distribution of *P*-values, making them nonuniform. Wigginton *et al.* (2005) looked at HWE testing for various sample sizes, significance thresholds, and minor allele frequencies (MAFs). They found that, even with a sample size of 1000, the actual type I error rates for both goodness-of-fit tests and exact tests may be much different from the nominal values. We build on their work by exploring properties of complete discrete *P*-value distributions under both the null and the alternative hypotheses, using simulated and biological data. We focus particularly on discreteness, power, and MAF affects.

In this article, on the 100th anniversary of the original HWE papers (Hardy 1908; Weinberg 1908), we examine the implications of discrete *P*-values in HWE testing. Evidence for departure from HWE has been used in many applications such as inferring the existence of natural selection (Wallace 1958; Lewontin and Cockerham 1959), challenging the statistical analysis

[1]*Corresponding author:* University of Washington, Department of Genome Sciences, Foege S-250, Box 355065, Seattle, WA 98195-5065. E-mail: rrohlfs@u.washington.edu

of forensic DNA profiles (Cohen *et al.* 1991; Weir 1992), and detecting genotyping errors (Gomes *et al.* 1999; Zou and Donner 2006). We derive the actual null distribution of both the chi-square goodness-of-fit test statistic (Weir 1996) and exact test *P*-values (Weir 1996) by completely enumerating all sets of genotype counts conditional on observed allele counts. These distributions are then used to explore type I error rate, power, an acceptable MAF range, and agreement with real data. Because of the current importance of diallelic SNPs in human genetics, we confine our attention to the two-allele case. We stress that the test statistics have very coarse distributions when the number of copies of the minor allele is small and this calls for caution in applying asymptotic assumptions and in determining significance thresholds in multiple-testing situations such as those in whole-genome scans. For all uses of test statistics and *P*-values, rigorous calculation and accuracy are required when determining the expected distributions to which observed values can be compared.

METHODS

***P*-value distributions under the null hypothesis:** For genotypes *AA*, *Aa*, *aa*, the sample counts are n_{AA}, n_{Aa}, n_{aa}, summing to n. The usual chi-square test statistic for HWE is constructed by comparing these counts to the values expected under HWE: $n\tilde{p}_A^2$, $2n\tilde{p}_A\tilde{p}_a$, $n\tilde{p}_a^2$ where $\tilde{p}_A = (2n_{AA} + n_{Aa})/(2n)$ and $\tilde{p}_a = 1 - \tilde{p}_A$. A convenient form for the test statistic is

$$X^2 = n\left(\frac{4n_{AA}n_{aa} - n_{Aa}^2}{(2n_{AA} + n_{Aa})(2n_{aa} + n_{Aa})}\right)^2. \qquad (1)$$

Under HWE, this test statistic is approximately chi-square distributed with 1 d.f., making the *P*-value for a given data set the area under the $\chi^2_{(1)}$-curve to the right of the calculated test statistic X^2.

The test statistic can also be written as $X^2 = n\hat{f}^2$, where $\hat{f}$ is the (discrete) maximum-likelihood estimate of the (continuous) within-population inbreeding coefficient f. The estimate is just the term in parentheses in Equation 1, and the parameter allows population genotype frequencies to be written in terms of allelic frequencies as $P_{AA} = p_A^2 + fp_Ap_a$, $P_{Aa} = 2(1 - f)p_Ap_a$, $P_{aa} = p_a^2 + fp_Ap_a$. When HWE is true, $f = 0$.

The goodness-of-fit test is equivalent to the 2 × 2 contingency table test of diploid genotype counts, as in Table 1. The test is conditional on the row and column totals, n_A, n_a. For any set of genotype counts there are only as many possible test statistic values as there are tables with the same row and column totals. Without loss of generality, we assume that *A* is the minor allele, meaning that $n_A = n_{Aa} + 2n_{AA} \leq n_a = n_{Aa} + 2n_{aa}$. With fixed row and column totals (n_A and n_a), n_{Aa} ranges over $0, 2, 4, \ldots, n_A$ if n_A is even and over $1, 3, 5, \ldots, n_A$ if n_A is odd. The number of possible n_{Aa} values, and therefore the number of test statistic values, is $\lfloor n_A/2 \rfloor + 1$, where $\lfloor x \rfloor$ indicates the largest integer less than or equal to x.

TABLE 1

Genotype count contingency table

$2n_{AA}$	n_{Aa}	n_A
n_{aA}	$2n_{aa}$	n_a
n_A	n_a	$2n$

Rows and columns in the 2 × 2 top left table sum to n_A or n_a appropriately. For example, $2n_{AA} + n_{Aa} = n_A$, since $n_A + n_a = 2n$, for some n and n_A.

An exact test does not rest on continuous approximations of discrete distributions and is not thought to be problematic with small numbers, as is a chi-square test. Rather, an exact test is based directly on the discrete sampling distribution of the data under the null hypothesis. With random sampling the multinomial distribution is applicable. Under the HWE null hypothesis the probability of the genotype counts n_{AA}, n_{Aa}, n_{aa} is

$$\Pr(n_{Aa} \mid n_A,\ n_a,\ \text{HWE}) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}\frac{2^{n_{Aa}}n_A!n_a!}{(2n)!} \qquad (2)$$

(Weir 1996). Note that the homozygote counts can be parameterized in terms of the heterozygote and allele counts as $n_{AA} = (n_A - n_{Aa})/2$ and $n_{aa} = (n_a - n_{Aa})/2$. We use n_{AA}, n_{aa} for notational simplicity. The *P*-value for this test is calculated as the sum of this probability for an observed data set and the probabilities of all other data sets that have the same or smaller probabilities when HWE is true. The total number of data sets, $\lfloor n_A/2 \rfloor + 1$, is usually small enough to allow for calculations based on a complete enumeration of these values. The more complex methods of Guo and Thompson (1992) are needed only for loci with multiple alleles.

As an illustration, Table 2 shows the eight possible data sets for a sample of 100 individuals with 14 copies of the *A* allele, along with the exact probabilities and *P*-values assuming that there is HWE. The conventional chi-square goodness-of-fit test statistics are also displayed along with the *P*-values from the $\chi^2_{(1)}$-distribution. This chi-square test would reject the HWE hypothesis at the 0.05 significance level if there were ≤10 heterozygotes whereas the exact test would change the rejection region to ≤8 heterozygotes. The "spurious significant" result from the chi-square test is removed if the test statistic is corrected for continuity by replacing $X^2 = \sum(o - e)^2/e$ with $X^2 = \sum(|o - e| - 0.5)^2)/e$ for a set of observed (o) and expected (e) counts (Yates 1934).

The key feature of Table 2 is that the null distributions of the *P*-values are far from uniform for each test. There are only eight distinct *P*-values and, for example, the probabilities that the *P*-values are ≥0.5 are 0.61, 0.00, and 0.93 for the exact, chi-square, and corrected chi-

TABLE 2

HWE chi-square and exact test example for $n = 100$ and $n_A = 14$

			Exact		Chi square		Corrected	
n_{aa}	n_{Aa}	n_{AA}	Probability[a]	P-value	X^2	P-value	X^2	P-value
93	0	7	0.0000	0.0000	100.00	0.0000	86.17	0.0000
92	2	6	0.0000	0.0000	71.64	0.0000	60.01	0.0000
91	4	5	0.0000	0.0000	47.99	0.0000	38.58	0.0000
90	6	4	0.0002	0.0002	29.07	0.0000	21.86	0.0000
89	8	3	0.0051	0.0053	14.87	0.0001	9.86	0.0017
88	10	2	0.0602	0.0654	5.38	0.0204	2.58	0.1082
87	12	1	0.3209	0.3863	0.61	0.4348	0.021	0.8849
86	14	0	0.6136	1.0000	0.57	0.4503	0.018	0.8936

[a] Computed under the null hypothesis.

square tests, respectively, instead of the 0.5 value expected for a uniform distribution.

The example in Table 2 is specific to one observed allele frequency ($\tilde{p}_A = 0.07$). In Figure 1 we show the exact P-value distributions for $\tilde{p}_A = 0.1$, 0.2, 0.3, 0.4, and 0.5 with sample size $n = 1000$. If there were only these five equally likely MAFs, the total P-value distribution would be the average of the five marginal distributions. This total P-value distribution (Figure 1, bottom) is clearly nonuniform. Note that, for any value of $\tilde{p}_A$, the most probable set of genotype counts has an exact P-value of 1.0. The total distribution of P-values over many $\tilde{p}_A$'s will have a spike at 1.0, which may be obscured if P-values are binned before plotting. For any given sample size, the minor allele count is constrained to $1, 2, \ldots, n$, giving n possible values for $\tilde{p}_A$. The distribution over all n allele frequencies for the exact test on 100 individuals is shown in Figure 2. Use of a simple average here assumes that all possible sample allele counts are equally likely.

The number of possible P-values is equal to the number of possible n_{Aa} values, which increases linearly with minor allele count (n_A), which in turn must be less than or equal to the sample size (n). Even as n becomes very large, there are a finite number of possible P-values so the P-value distribution remains discrete, although when binned, it increasingly resembles a uniform distribution.

The chi-square test P-values have a similarly coarse distribution. However, for the chi-square test, the P-value can be 1.0 only if the sample counts are in perfect HWE and this is not usually possible. The null distribution of chi-square test P-values has no spike at 1.0 but it is not uniformly distributed. For the remainder of this article we focus on the exact test.

***P*-value distributions under the alternative hypothesis:** A full description of the behavior of hypothesis tests requires consideration of power: the probability of rejecting the null hypothesis when the alternative hypothesis is true. For the chi-square test, the test statistic still has the value $n\hat{f}^2$ but under the alternative hypothesis it has a noncentral chi-square distribution with 1 d.f. and noncentrality parameter $\lambda = nf^2$. The power of the test can be calculated as the probability to the right of X^2 under the noncentral chi-square distribution. This continuous-distribution theory is convenient for simple sample-size calculations: in this 1-d.f. case, the noncentrality value follows from percentiles of the standard normal distribution. If z_x is the xth percentile of the standard normal, than for significance level α and power $1 - \beta$, $nf^2 = \lambda = (z_{\alpha/2} + z_\beta)^2$. For 90% power and 5% significance level, $\lambda = 10.5$ and $n \geq 10.5/f^2$.

The probability of a given set of genotype counts when HWE is not assumed can be written as

$$\Pr(n_{Aa} \mid n_A,\ n_a) = \frac{\theta^{n_{Aa}/2}}{n_{AA}!n_{Aa}!n_{aa}!} \cdot \frac{1}{C}$$

(Wellek 2004) to show how, for given values of the allele counts, the probability depends only on the heterozygote count (recall that n_{Aa}, n_A, n_a determine n_{AA}, n_{aa}) and a single parameter $\theta = P_{Aa}^2/(P_{AA}P_{aa})$. The normalizing constant is found by summing over all possible values of n_{Aa}:

$$C = \sum_{n_{Aa}} \frac{\theta^{n_{Aa}/2}}{n_{AA}!n_{Aa}!n_{aa}!}.$$

The exact power of the test for some θ is found by summing values of this probability over all genotype counts for which the test rejects the null hypothesis. The rejection region is calculated under the null hypothesis but the power is calculated under the alternative hypothesis. Although the θ-parameterization avoids relying on allele frequencies and is convenient for performing power calculations, it is possible to retain the inbreeding coefficient approach by recognizing that $\theta = 4p_Ap_a(1 - f)^2/[(p_A + fp_a)(p_a + fp_A)]$.

Power calculations are illustrated in Table 3 for the case of $n = 100$, $n_A = 14$ where the rejection region is chosen for the conventional 0.05 significance level (*i.e.*, the exact test rejects for ≤ 8 heterozygotes and the chi-square test rejects for ≤ 10 heterozygotes). A range of θ-values are considered. Note that when $\theta = 4$, the null

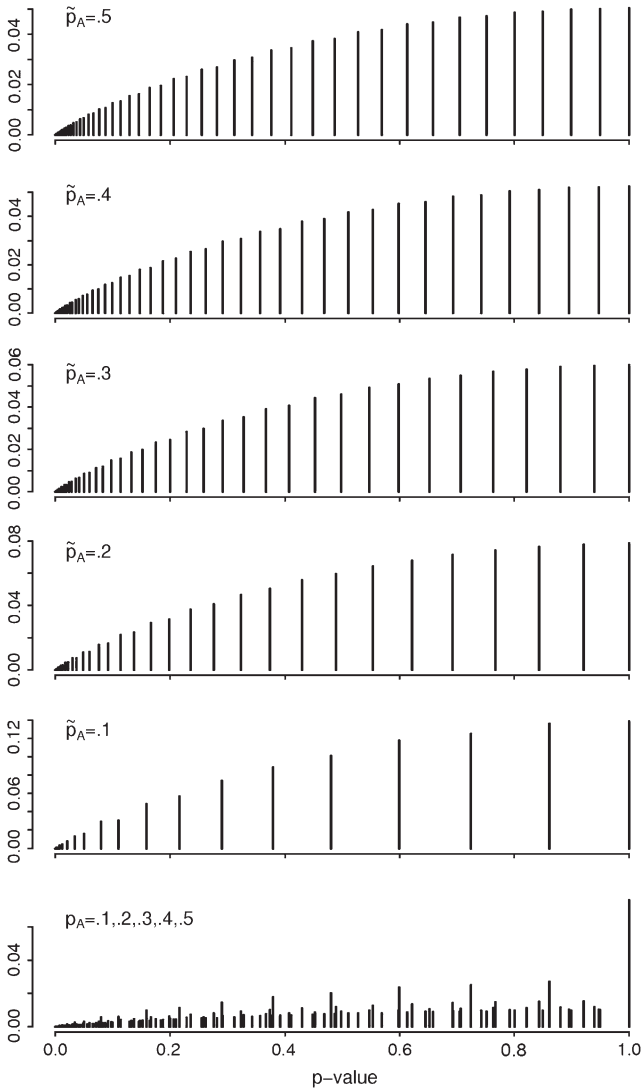

FIGURE 1.—Exact *P*-value distributions conditional on MAF. The top five plots show distribution functions for exact *P*-values for 1000 individuals with the indicated MAF. The bottom plot is the normalized sum of the plots above.

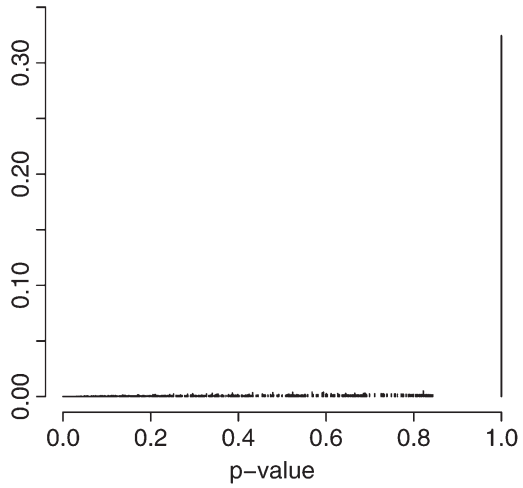

FIGURE 2.—Total exact *P*-value distribution. The HWE exact *P*-value probability density function is under the null hypothesis where $n = 100$ and minor allele frequencies are uniformly distributed.

hypothesis of HWE is true and the power is equivalent to the type I error rate. When $\theta < 4$ and $\theta > 4$, there is an excess of homozygotes and heterozygotes, respectively. In the example shown in Table 3, the most likely genotype configuration under the null hypothesis is 14 heterozygotes and no minor allele homozygotes, as shown in Table 2. So in the case of excess homozygotes ($\theta < 4$), the tests gain power. However, it is not possible to observe an excess of heterozygotes, so this sort of departure is not detectable by either test, causing them to lose power when $\theta > 4$. Figure 3 shows how type I error rate and power vary across n_A, n, and θ. Type I error and power increase and decrease simultaneously so that if power is high for a particular MAF, type I error is also high. Because the chi-square rejection region is not smaller than the exact test region, the chi-square test is always at least as powerful as the exact test, even as the power of each varies with MAF.

AN APPLICATION

In the exact *P*-value distribution over all $\tilde{p}_A$ for $n = 100$ shown in Figure 2, we assumed that MAFs are uniformly distributed, but this may not be true in some SNP panels. When MAFs in a data set are not uniformly distributed, the empirical MAF distribution can be used to weight the null distributions of *P*-values for each MAF. If the sample size is n for every SNP, then there are only n possible MAF values and marginal null distributions. The number of times each occurs in the data can be used as a weight in constructing the total null distribution. We performed exact HWE tests on real data to demonstrate this point and to explore the consequences of approximating the expected *P*-value distribution as uniform.

The Wellcome Trust Case–Control Consortium (WTCCC) kindly provided us with genomewide SNP data in 1504 individuals from the 1958 birth cohort (Wellcome Trust Case–Control Consortium 2007). This sample, consisting of all individuals born in England, Scotland, and Wales in 1 week in 1958, was previously found to have negligible population structure (Wellcome Trust Case–Control Consortium 2007).

To avoid complications of varying sample size due to missing data, we excluded SNPs with any missing genotypes, for a total of 34,625 polymorphic SNPs. Using unfiltered complete SNPs makes our final SNP panel susceptible to genotyping errors and severe ascertainment bias. This SNP panel is not appropriate for biological analysis or as a surrogate for the 1958 birth cohort data set; however, it is a mix of actual SNPs in

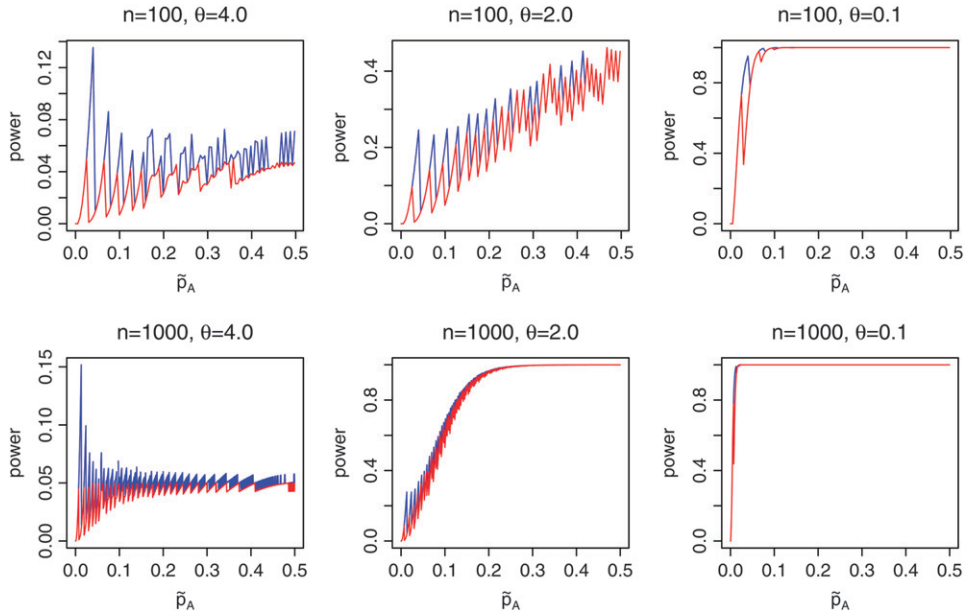

FIGURE 3.—HWE test power. Power is shown over all p_A for $n = 100$, 1000 and $\theta = 4$, 2, 0.1. Note that $\theta = 4$ implies HWE, so the power is equivalent to the type I error rate. Red lines show χ^2-test power, and blue lines show exact test power.

HWE and Hardy–Weinberg disequilibrium (HWD) and thus is useful for comparing observed *P*-values with different expected *P*-value distributions. This SNP set may contain a higher proportion of non-HWE SNPs than is typical in quality-controlled data, but that is not relevant when comparing analyses across expected distributions. The nonuniformity of MAFs in this SNP set (Figure 4) justify weighting the expected marginal *P*-value distributions by the observed MAF distribution.

When comparing observed and expected values, such as the HWE exact *P*-values, a *Q-Q* plot proves a useful diagnostic by plotting the observed ranked *P*-values against the ranked *P*-values sampled independently from the theoretical null distribution of *P*-values. If the observed and expected values are similarly distributed, all the points lie near the diagonal. For a more detailed examination of the very small *P*-values, $-\log(p)$ can be used in a *Q-Q* plot. Figure 5 shows *Q-Q* plots comparing *P*-values resulting from tests of the observed SNPs to expected *P*-values drawn from a uniform distribution and from our calculated null distribution with $n = 1504$ and the observed MAF distribution. Many of the observed SNPs are in strong Hardy–Weinberg disequilibrium. The farthest outliers are likely to be caused by genotyping errors and many other outliers may be due to population structure or selection. Clearly, expected *P*-values drawn from the calculated null distribution match the bulk of the observed *P*-values much closer than *P*-values drawn from a uniform distribution. Many of the observed *P*-values are 1.0, resulting in a horizontal bar in the *Q-Q* plots against uniform *P*-values. Using a uniform distribution as the expected distribution in *Q-Q* plots leads to inaccurate assessment of expected and observed data agreement.

HWE testing to control SNP quality has often used a somewhat arbitrary significance level, such as 0.001, and excluded markers with *P*-values below that threshold from further analysis. Using this 0.001 significance threshold, our analysis showed an observed rejection rate of 0.002484, which is twice the type I error rate expected under the continuous assumption (0.001) but four times the exact expected rate of 0.0005827. A simple 1-d.f. chi-square test shows that the observed positive rate is significantly higher than expected with asymptotic assumptions ($X^2 = 76.23$), more so if the calculated expected value is used ($X^2 = 214.81$). There may exist a case where assuming uniform *P*-values leads to the conclusion that there are not significantly more positives than expected under the null hypothesis when, in fact, there are.

As mentioned previously, the type I error rate can be set exactly using randomized tests that reject the null hypothesis with a specific probability for *P*-values closest to the significance threshold (TOCHER 1950; INNAN *et al.* 2005; GIBBONS 2006). However, randomization can cause the same analysis to call a marker significant on one run and not significant on the next. This inconsistency is unsatisfying and can be avoided by deriving the true *P*-value distribution.

TABLE 3

Comparative power

	$\theta = 8$	$\theta = 4$	$\theta = 2$	$\theta = 0.1$
Chi-square test	0.0199	0.0654	0.1848	0.9900
Exact test	0.0008	0.0053	0.0285	0.9185

The exact power is computed with $P(n_{Aa} \mid n_A, n_a)$ for the exact and chi-square tests where $n = 100$, $n_A = 14$, and 0.05 is the nominal significance level. θ parameterizes deviance from HWE so that when $\theta = 4$ there is HWE and the power is equivalent to type I error.

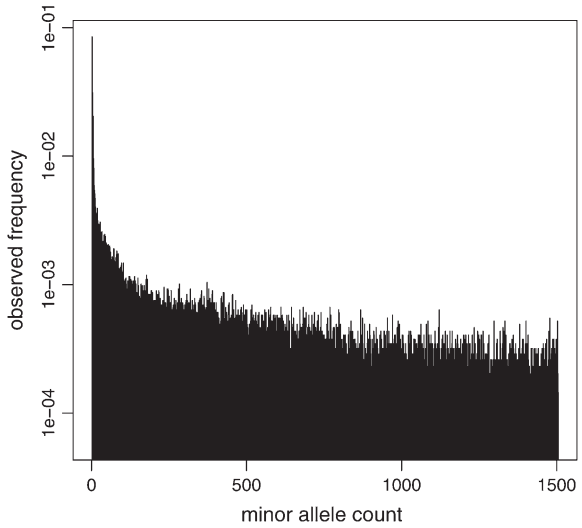

FIGURE 4.—Observed MAF distribution. This plot shows the distribution of observed minor allele counts among SNPs with no missing data in the WTCCC 1958 birth cohort.

In the 1958 birth cohort analysis outlined above, some SNPs are in HWE and some are in HWD but the true proportions of HWE and HWD SNPs are unknown. To show the relevance of expected distribution choice when all SNPs are in HWE, we performed a simulation study of 100,000 SNPs in 100 individuals with uniformly distributed MAFs under HWE. Figure 6 shows *Q-Q* plots comparing exact HWE *P*-values calculated on the basis of the simulated data with the calculated null *P*-values and uniform *P*-values. Similar to the real data analysis, the calculated null *P*-values fit the observed *P*-values much better than do the uniform *P*-values. When comparing observed *P*-values to the uniform, an investigator may be surprised to see a deficit of low observed *P*-values, indicating that many SNPs have a closer fit to HWE than expected under HWE. When comparing the observed *P*-values to the calculated null *P*-values, the distributions align well, indicating that the SNPs are in HWE.

DISCUSSION

The small number of possible sets of genotype counts for a given set of allele counts can lead to very coarse discrete distributions of *P*-values that may not be adequately approximated by a continuous uniform distribution. Analyses that assume *P*-value uniformity may be reconsidered using the true, discrete, coarse *P*-value distribution. These analyses include type I error rate, power, false discovery rate (FDR) estimation, and distribution agreement assessment.

In most studies across many diallelic markers, a lower bound is set on acceptable MAF. Figure 7 illustrates the effects of MAF lower bounds on power, type I error rate, and FDR. FDR is the proportion of significant tests

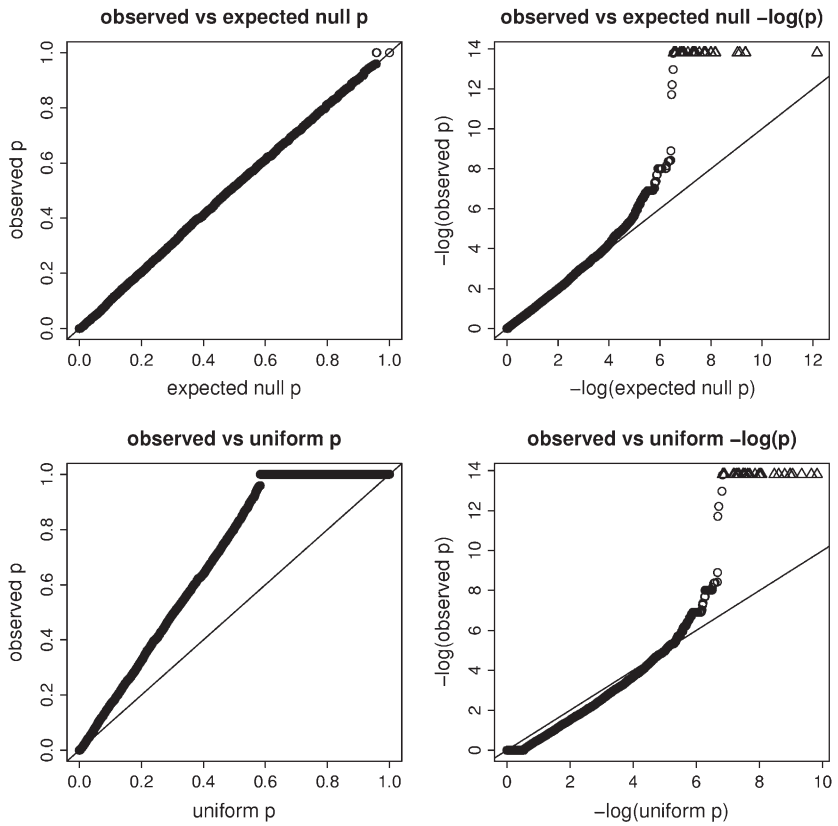

FIGURE 5.—Observed *vs.* expected exact *P*-values for the 1958 birth cohort data set. The top *Q-Q* plots compare *P*-values observed in 34,625 SNPs over 1504 individuals in the WTCCC 1958 birth cohort to *P*-values sampled from the calculated expected null distribution for $n = 1504$ and observed MAF weights. The bottom *Q-Q* plots compare the same observed *P*-values to a uniform distribution. SNPs with *P*-values $<1e\text{-}6$ are represented as triangles at the top of the $-\log(p)$ plots.

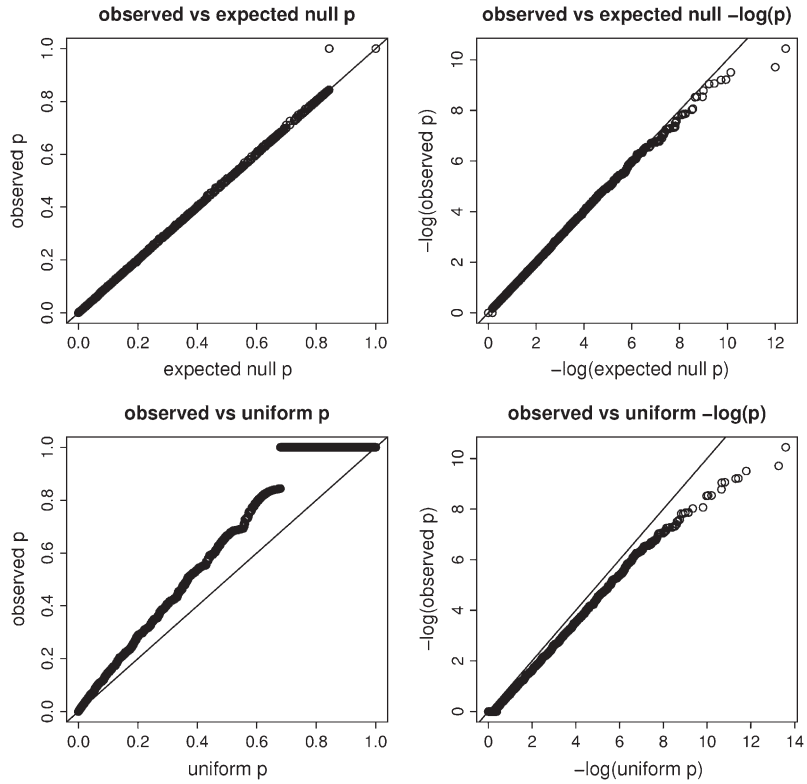

FIGURE 6.—Observed *vs.* expected exact *P*-values for simulated SNPs. The top *Q-Q* plots compare *P*-values calculated for 100,000 simulated SNPs over 100 individuals with uniform MAFs to *P*-values sampled from the calculated expected null distribution for $n = 100$ with uniformly distributed MAFs. The bottom *Q-Q* plots compare the same calculated *P*-values to a uniform distribution.

falsely deemed significant and is often used in multiple-testing situations such as whole-genome scans. Applying a lower bound on MAF actually increases type I error rate and power. Depending on the specific parameters (n, n_A, θ, proportion of truly null and alternative markers), the net effect may be calculated as change in FDR for a specific lower bound. Often MAF lower bounds are used to avoid genotyping errors that are more pronounced in low MAF markers. Even in these cases, we advise informing the decision with an exact analysis of genotyping error rate, type I error rate, power, and FDR depending on MAF lower bound.

The results shown in this article are specific to HWE tests, but the same concept applies to any statistical test based on discrete data. In the field of statistical genetics, this includes most tests using genetic marker data such as simple case–control association (NEUHÄUSER 2002), linkage disequilibrium (LD) (ZAPATA and ALVAREZ 1997), and general allelic association (ZAYKIN *et al.* 1995) testing. Data for these problems may be framed in a contingency table and the probability of a particular table given the marginal counts can be computed under the null and alternative hypotheses. For example, ZAPATA and ALVAREZ (1997) show the contingency

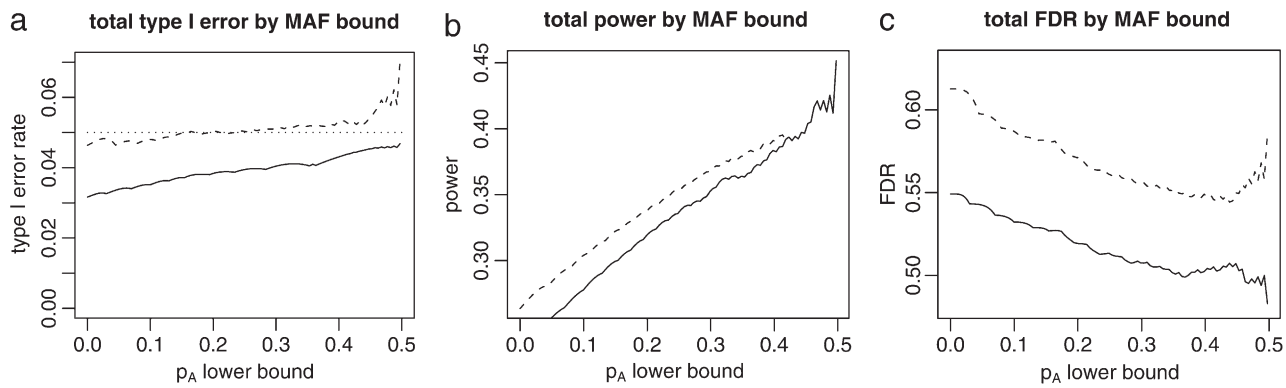

FIGURE 7.—MAF-bound effects. The plots show total (a) type I error rate, (b) power, and (c) FDR for every possible lower bound on $\tilde{p}_A$, where $n = 100$, $\theta = 2$, and the rate of truly alternative SNPs is 0.1. Dashed lines indicate the chi-square test and solid lines Fisher's exact test. The horizontal dotted line in the type I error plot shows the significance threshold of 0.05.

table for LD testing and the procedure for the exact LD test. They show the probability of a particular table under the null hypothesis of no LD with n haplotypes,

$$\Pr(n_{CD} \mid n,\ n_C,\ n_D,\ H_0) = \frac{n_C!n_c!n_D!n_d!}{n!n_{CD}!n_{Cd}!n_{cD}!n_{cd}!}, \qquad (3)$$

where C and D are loci with alleles C/c and D/d. Note that if n_C, n_c, n_D, n_d are held constant, the contingency table consisting of all haplotype counts can be completely specified with n_{CD}, the number of CD haplotypes present. A certain degree of linkage disequilibrium can be parameterized as $\kappa = (p_{CD}p_{cd})/(p_{Cd}p_{cD})$, where p_{CD} is the probability of observing the genotype CD and so on. The probability of a particular two-locus haplotype configuration for some value of κ can be calculated as

$$\Pr(n_{CD} \mid n,\ n_C,\ n_D,\ \kappa) = \frac{n!}{n_{CD}!n_{Cd}!n_{cD}!n_{cd}!}(\kappa)^{n_{CD}} \cdot \frac{1}{C},$$

where

$$C = \sum_{n_{CD}} \frac{n!}{n_{CD}!n_{Cd}!n_{cD}!n_{cd}!}(\kappa)^{n_{CD}}.$$

Using these probabilities and the enumeration of all contingency tables conditional on the marginals, a test statistic or P-value distribution can be derived for all possible marginals. This provides the precise test statistic or P-value distribution under a null or an alternative hypothesis.

We thank T. Lumley and W. G. Hill for their thoughtful discussion on this topic. This study makes use of data generated by the Wellcome Trust Case–Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. This work was supported in part by National Institutes of Health grants R01 GM 75091 and T32 GM07735. Funding for the project was provided by the Wellcome Trust under award 076113.

LITERATURE CITED

Cohen, J. E., M. Lynch and C. E. Taylor, 1991 Forensic DNA tests and Hardy-Weinberg equilibrium. Science **253:** 1037–1038.

Gibbons, J. D., 2006 Randomized tests, in *Encyclopedia of Statistical Sciences*, edited by S. Kotz, C. B. Read, N. Balakrishnan and B. Vldakovic. John Wiley & Sons, New York.

Gomes, I., A. Collins, C. Lonjou, N. S. Thomas, J. Wilkinson *et al.*, 1999 Hardy-Weinberg quality control. Ann. Hum. Genet. **63:** 535–538.

Guo, S. W., and E. A. Thompson, 1992 Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics **48:** 361–372.

Hardy, G. H., 1908 Mendelian proportions in a mixed population. Science **28:** 49–50.

Innan, H., K. Zhang, P. Marjoram, S. Tavaré and N. A. Rosenberg, 2005 Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. Genetics **169:** 1763–1777.

Lewontin, R. C., and C. C. Cockerham, 1959 The goodness-of-fit test for detecting natural selection in random mating populations. Evolution **13:** 561–564.

Neuhäuser, M., 2002 Exact tests for the analysis of case-control studies of genetic markers. Hum. Hered. **54:** 151–156.

Raymond, M., and F. Rousset, 1995 An exact test for population differentiation. Evolution **49:** 1280–1283.

Rousset, F., and M. Raymond, 1995 Testing heterozygote excess and deficiency. Genetics **140:** 1413–1419.

Slatkin, M., 1994 Linkage disequilibrium in growing and stable populations. Genetics **137:** 331–336.

Tocher, K. D., 1950 Extension of the Neyman-Pearson theory of tests to discontinuous variates. Biometrika **37:** 130–144.

Wallace, B., 1958 The comparison of observed and calculated zygotic distributions. Evolution **12:** 113–115.

Weinburg, W., 1908 Über den Nachweis der Vererbung beim Menchen. *Jahresh. Verein* f vaterl Naturk in Wüttemberg **64:** 368–382 (translated in *Papers on Human Genetics*, 1963, pp. 4–15. Prentice-Hall, Englewood Cliffs, NJ).

Weir, B. S., 1992 Population genetics in the forensic DNA debate. Proc. Natl. Acad. Sci. USA **89:** 11654–11659.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Wellcome Trust Case–Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447:** 661–683.

Wellek, S., 2004 Tests for establishing compatibility of an observed genotype distribution with Hardy-Weinberg equilibrium in the case of a biallelic locus. Biometrics **60:** 694–703.

Wigginton, J. E., D. J. Cutler and G. R. Abecasis, 2005 A note on exact tests of Hardy-Weinberg equilibrium. Am. J. Hum. Genet. **76:** 887–893.

Yates, F., 1934 Contingency tables involving small numbers and the X^2 test. J. R. Stat. Soc. **1**(Suppl.): 217–235.

Yates, F., 1984 Tests of significance for 2 × 2 tables. J. R. Stat. Soc. Ser. A **147:** 426–463.

Zapata, C., and G. Alvarez, 1997 On Fisher's exact test for detecting gametic disequilibrium between DNA polymorphisms. Ann. Hum. Genet. **61:** 71–77.

Zaykin, D., L. Zhivotovsky and B. S. Weir, 1995 Exact tests for association between alleles at arbitrary numbers of loci. Genetica **96:** 169–178.

Zou, G. Y., and A. Donner, 2006 The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: a cautionary note. Ann. Hum. Genet. **70:** 921–933.

Communicating editor: M. W. Feldman