Jan Graffelman* and Victor Moreno

The mid *p*-value in exact tests for Hardy-Weinberg equilibrium

Abstract

Objective: Exact tests for Hardy-Weinberg equilibrium are widely used in genetic association studies. We evaluate the mid *p*-value, unknown in the genetics literature, as an alternative for the standard *p*-value in the exact test.

Method: The type 1 error rate and the power of the exact test are calculated for different sample sizes, sigificance levels, minor allele counts and degrees of deviation from equilibrium. Three different *p*-value are considered: the standard two-sided *p*-value, the doubled one-sided *p*-value and the mid *p*-value. Practical implications of using the mid *p*-value are discussed with HapMap datasets and a data set on colon cancer.

Results: The mid *p*-value is shown to have a type 1 error rate that is always closer to the nominal level, and to have better power. Differences between the standard *p*-value and the mid *p*-value can be large for insignificant results, and are smaller for significant results. The analysis of empirical databases shows that the mid *p*-value uncovers more significant markers, and that the equilibrium null distribution is not tenable for both databases.

Conclusion: The standard exact *p*-value is overly conservative, in particular for small minor allele frequencies. The mid *p*-value ameliorates this problem by bringing the rejection rate closer to the nominal level, at the price of ocasionally exceeding the nominal level.

Keywords: Levene-Haldane distribution; power; single nucleotide polymorphism; type I error rate.

***Corresponding author: Jan Graffelman,** Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Avinguda Diagonal 647, 08028 Barcelona, Spain, Phone: +34-934011739, Fax: +34-934016575, e-mail: jan.graffelman@upc.edu

Victor Moreno: Cancer Prevention Program, Catalan Institute of Oncology (ICO) Bellvitge Biomedical Research Institute (IDIBELL), Faculty of Medicine and CIBERESP, Department of Clinical Sciences, University of Barcelona (UB), Gran Via 199, 08908 L'Hospitalet del Llobregat (Barcelona), Spain

1 Introduction

Nowadays, testing genetic markers for Hardy-Weinberg proportions (HWP) is a standard aspect of the analysis of large SNP databases used in genetic association studies. Deviations from HWP may, among other reasons, be the result of genotyping error (Hosking et al., 2004; Attia et al., 2010), population stratification or disease association. Currently, three classes of statistical procedures are in use for testing for HWP. The first class is the classical χ^2-test for goodness of fit, which tests if the genotype counts are compatible with a multinomial distribution given the observed allele frequencies. There are some variations on the χ^2-test (Elston and Forthofer, 1977; Emigh, 1980; Smith, 1986) and results of the test can be affected by the use of the continuity correction, in particular if the minor allele frequency is low (Graffelman and Morales-Camarena, 2008). The second class of tests concerns exact procedures. The exact test for HWP dates back to the work of Levene (1949) and Haldane (1954) and will be discussed in more detail below. The third class comprises Bayesian procedures to test for HWP. The Bayesian approach started with the work of Lindley (1988). Recently, several interesting papers on HWP using a Bayesian approach have appeared (Ayres and Balding, 1998; Shoemaker et al., 1998; Wakefield, 2010). The classical χ^2-test is probably still the most popular way to test HWP (Salanti et al., 2005, Yu et al., 2009), though thanks to the availability of increased computing power and software

for exact tests, the latter is becoming increasingly popular. In the remainder of this paper we restrict our attention to exact test procedures and bi-allelic markers. Exact procedures for multiple alleles have also been developed (Guo and Thompson, 1992; Chakraborty and Zhong, 1994). The structure of the remainder of this paper is as follows. In Section 2 we summarize the exact test for HWP and enumerate different definitions of the *p*-value for this test. In Section 3 we compare the type I error rate and the power of the different versions of the exact test. Section 4 shows the practical implications of using the different *p*-values with HapMap datasets and with a colon cancer dataset. A discussion (Section 5) completes the paper.

2 The exact test for HWP and its *p*-values

The exact test for HWP is based on the conditional distribution of the number of heterozygotes N_{AB}, given the allele count N_A and sample size N. This distribution is given by

$$P(N_{AB}=n_{AB}\,|\,N=n,N_A=n_A)=\frac{n!\,n_A!\,n_B!\,2^{n_{AB}}}{(2n)!\,n_{AB}!\left(\frac{1}{2}(n_A-n_{AB})\right)!\left(\frac{1}{2}(n_B-n_{AB})\right)!}, \tag{1}$$

where n_A and n_B are the sample counts of allele A and B, respectively ($n_A=2n_{AA}+n_{AB}$ and $n_B=2n_{BB}+n_{AB}$), n is the total sample size and n_{AA}, n_{BB} and n_{AB} are the sample counts of the homozygotes and the heterozygotes (Weir, 1996). We refer to this conditional distribution as the Levene-Haldane distribution. Fast recursive procedures exist to compute the probabilities according to Equation (1) for different numbers of heterozygotes (Elston and Forthofer, 1977; Wigginton et al., 2005). For efficient implementation of the recursive procedure, knowledge of the expectation of the Levene-Haldane distribution can be useful. The expectation and the variance of $N_{AB}|N, N_A$ can, using results from Okamoto and Ishii (1961), be shown to be:

$$E(N_{AB}\,|\,N,N_A)=\frac{n_An_B}{2n-1},\ V(N_{AB}\,|\,N,N_A)=\frac{n_A(n_A-1)n_B(n_B-1)(2n-1)+n_An_B(2n-1)(2n-3)+n_A^2n_B^2(2n-3)}{(2n-1)^2(2n-3)}. \tag{2}$$

In general, it is difficult to give an adequate definition of a *p*-value for a two-sided exact test if the null distribution is discrete and asymmetric. The Levene-Haldane distribution is an example of this. The *p*-value of an exact test is usually computed as the sum of the probabilities of all possible samples with the same allele count that are as likely or less likely than the observed sample. However, there are several ways to compute a *p*-value in an exact test, in particular if a two-sided test is desired. We will consider four different *p*-values, the one-sided *p*-value, the doubled one-sided *p*-value, the standard two-sided *p*-value, and the mid *p*-value, and discuss all these in the corresponding subsections below. A graphical representation of the different *p*-values is shown in Figure 1. Here, we first introduce some additional notation. With K we indicate the number of possible samples for the observed number of A alleles. Let p_i be the probability of observing a particular sample under the Levene-Haldane distribution, given by Equation (1) , and let k be the index of the observed sample in the probability array p_i, the latter array ordered according to the number of heterozygotes n_{AB}.

2.1 The one-sided *p*-value

In a one-sided exact test, the *p*-values of the tests for heterozygote dearth and excess are respectively computed as

$$P(N_{AB}\le n_{AB})=\sum_{i=1}^{k}p_i \quad\text{and}\quad P(N_{AB}\ge n_{AB})=\sum_{i=k}^{K}p_i.$$

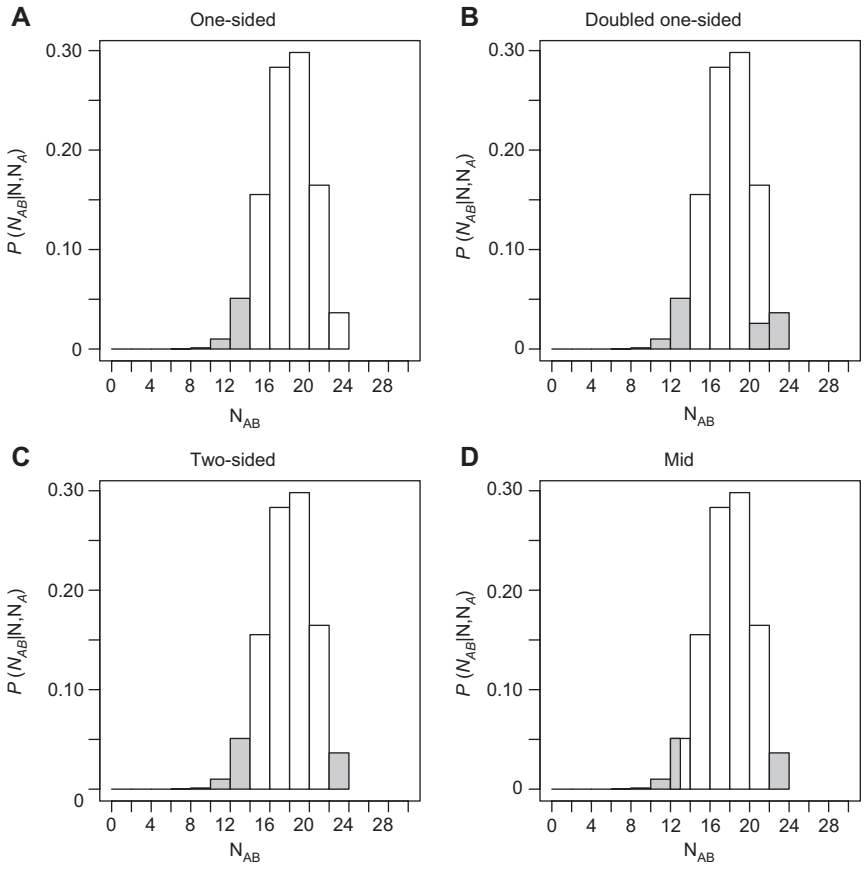

Figure 1 Computation of the *p*-value in an exact test for HWP, for a sample of 50 individuals with a minor allele count of 23, for which 13 heterozygotes were observed. (A) One-sided *p*-value in a test for heterozygote dearth. (B) *p*-value obtained by doubling the one-sided tail. (C) Standard two-sided *p*-value, (D) Mid *p*-value based on half the probability of the observed sample.

In most modern genetic studies there are often no a priori reasons for expecting heterozygote dearth or excess, and two-sided tests therefore form the most natural choice.

2.2 The doubled one-sided *p*-value

The doubled one-sided *p*-value constitutes a more conservative approach to testing markers, and is based on doubling the *p*-value obtained in a one-sided test (Yates, 1984; Graffelman, 2010).

$$P_{\text{DOUBLED ONE-SIDED}} = min(2P(N_{AB} \geq n_{AB}),\ 2P(N_{AB} \leq n_{AB}), 1). \tag{3}$$

Doubling a one-sided tail of a non-symmetric discrete distribution may lead to *p*-values that exceed 1, and for this reason 1 is included in the computation of the minimum in (3). Therefore, the doubled one-sided *p*-value cannot exceed 1.

2.3 The standard two-sided *p*-value

A two-sided *p*-value is currently the standard way to perform an exact test for HWP. Given the allele frequency of the observed sample, the probability of the observed sample can be computed Equation (1). The two-sided *p*-value is the sum of the probabilities of all possible samples with the same allele frequency that are as far

as or farther from HWP in comparison with the observed sample. A two-sided test considers both too large and too small heterozygote counts as evidence against the null. From this point on, we assume the array of probabilities p_i to be ordered by increasing p_i, and we redefine k as the index of the observed sample in this ordered array (in general, the new index value for k will be different from the one used in Section 2.1). The two-sided p-value can then be calculated as

$$P_{\text{TWO-SIDED}} = \sum_{i=1}^{k} p_i. \tag{4}$$

The type I error rate of a test based on the standard two-sided p-value is known never to exceed the nominal significance level α (Wigginton et al., 2005). This means that the test does not suffer from excessively high rejection rates. The classical χ^2-test is known to have too high rejection rates for low minor allele frequencies (Emigh, 1980; Wigginton et al., 2005). When the allele frequency is not too extreme, the rejection rate of the χ^2-test is improved by using the continuity correction (Graffelman, 2010). Even though the two-sided p-value is the most popular and intuitive criterion for an exact test, it suffers certain problems, and could be termed overly conservative for having its rejection rate always below α.

Ideally, the p-value of a statistical test should have a uniform distribution if the null hypothesis is true. This means that, under the null hypothesis, the p-value distribution should have expectation $\frac{1}{2}$ and variance $\frac{1}{12}$. However, with discrete random variables (like the number of heterozygotes in a sample of individuals) this is often not exactly true in practice. It can be shown that the expectation of the standard two-sided p-value always exceeds one half, and that its variance exceeds $\frac{1}{12}$. For the expectation, we have

$$\begin{aligned} E(P_{\text{TWO-SIDED}}) &= \sum_{k=1}^{K} p_k \left(\sum_{j=1}^{k} p_j \right) = \sum_{k=1}^{K} p_k^2 + \sum_{j<k}^{K} p_j p_k \\ &= \sum_{k=1}^{K} p_k^2 + \frac{1}{2} \left\{ \left(\sum_{k=1}^{k} p_k \right)^2 - \sum_{k=1}^{K} p_k^2 \right\} = \frac{1}{2} + \frac{1}{2} \sum_{k=1}^{K} p_k^2, \end{aligned} \tag{5}$$

which implies that the two-sided p-value is, strictly speaking, not uniformly distributed. The real null distribution of the p-value typically has a spike around 1, as is discussed in detail by Rohlfs and Weir (2008).

2.4 The mid *p*-value

In exact tests concerning discrete random variables, Lancaster (1961) proposed the mid p-value as the p-value for use in exact tests. The mid p-value has also been advocated by Agresti (2002), and is further discussed by Barnard (1989), Hirji (1991) and Berry and Armitage (1995). To date, this p-value seems not to have been used in exact tests for HWP, and to be unknown in genetics. The mid p-value is *half* the probability of the observed sample plus the sum of the probabilities of all possible samples that are farther from HWP.

$$P_{\text{MID}} = \frac{1}{2} p_k + \sum_{i=1}^{k-1} p_i = \sum_{i=1}^{k} p_i - \frac{1}{2} p_k$$

This p-value has the advantage that, under the null distribution, its expectation is exactly $\frac{1}{2}$, and its variance is only slightly below $\frac{1}{12}$, as is shown below:

$$E(P_{\text{MID}}) = \sum_{k=1}^{K} p_k \left(\sum_{j=1}^{k} p_j - \frac{1}{2} p_k \right) = \sum_{k=1}^{K} p_k \sum_{j=1}^{k} p_j - \frac{1}{2} \sum_{k=1}^{K} p_k^2 = E(P_{\text{TWO-SIDED}}) - \frac{1}{2} \sum_{k=1}^{K} p_k^2 = \frac{1}{2}.$$

$$V(P_{\text{MID}}) = \frac{1}{12} \left(1 - \sum_{k=1}^{K} p_k^3 \right).$$

The different *p*-values are graphically represented in Figure 1 for a sample of n=50 individuals with a minor allele count of n_A=23, for which 13 heterozygotes were observed. Figure 1 shows the distribution of the number of heterozygotes given the minor allele count, where the different *p*-values are shown as grey shaded areas. The probabilities of all possible samples with n=50 and n_A=23 and the different *p*-values are listed in Table 1.

Table 1 illustrates that mid *p*-values are smaller than standard two-sided *p*-values and can differ considerably. If, for example, a significance level of 1% is adopted, the mid *p*-value would reject HWE for a sample with 11 heterozygotes, whereas the two-sided *p*-value would not.

3 Type I error rate and power comparison

Besides considerations regarding the expectation and the variance of the *p*-value distribution, ideally we would want a statistical test that has a type I error rate that is as close as possible to the nominal significance level (α), and that has high power. For a given sample size, allele frequency and significance level, the rejection rate of the exact test with each type of *p*-value can be calculated exactly (simulation is not needed). The probabilities of all possible K outcomes (p_i=$P(N_{AB}$=$n_{AB}|N$=n, N_A=$n_A)$) under the Levene-Haldane distribution are calculated, as well as their corresponding *p*-values for all these samples (Pv_i). The rejection rate (type I error) is then obtained as:

$$\sum_{i=1}^{k} p_i \cdot I_{Pv_i \leq \alpha}(i), \tag{6}$$

where $I_{Pv_i \leq \alpha}(i)$ is a binary variable indicating whether the ith sample is significant at level α or not. Rejection rates for two-sided, doubled one-sided and mid *p*-values and power were computed as a function of sample size (25, 50, 100 and 1000), the nominal significance level (α=0.05, 0.01 or 0.001) and the minor allele count and are shown in Figure 2.

Figure 2 shows that the mid *p*-value has a rejection rate that is always larger or equal to the doubled one-sided and the standard two-sided rejection rates, and is thus a more liberal way of testing for HWP. The higher rejection rate was to be expected, because mid *p*-values are always smaller than or equal to two-sided *p*-values. However, the most important conclusion from Figure 2 is that the rejection rate of the mid *p*-value is

Table 1 Calculation of the different *p*-values in an exact test for HWE with n=50 and n_A=23.

n_{AA}	n_{AB}	n_{BB}	$P(N_{AB}=n_{AB})$	$P(N_{AB}\leq n_{AB})$	$P(N_{AB}\geq n_{AB})$	$P_{\text{D ONE-SIDED}}$	$P_{\text{TWO-SIDED}}$	P_{MID}	X^2	$P(\chi_1^2 \geq X^2)$
11	1	38	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	44.51	0.0000
10	3	37	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	34.50	0.0000
9	5	36	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	25.75	0.0000
8	7	35	0.0001	0.0001	1.0000	0.0002	0.0001	0.0000	18.29	0.0000
7	9	34	0.0012	0.0012	0.9999	0.0025	0.0012	0.0007	12.09	0.0005
6	11	33	0.0100	0.0113	0.9988	0.0225	0.0113	0.0063	7.18	0.0074
5	13	32	0.0510	0.0622	0.9887	0.1245	0.0987	0.0732	3.54	0.0600
4	15	31	0.1553	0.2175	0.9378	0.4351	0.2540	0.1763	1.17	0.2792
3	17	30	0.2832	0.5008	0.7825	1.0000	0.7019	0.5603	0.08	0.7768
2	19	29	0.2981	0.7989	0.4992	0.9985	1.0000	0.8509	0.27	0.6065
1	21	28	0.1647	0.9636	0.2011	0.4022	0.4187	0.3363	1.73	0.1890
0	23	27	0.0364	1.0000	0.0364	0.0729	0.0477	0.0295	4.46	0.0347

All possible samples with n=50 and n_A=23 are listed with their probabilities ($P(N_{AB}$=$n_{AB})$). All exact *p*-values considered are given ($P(N_{AB}\leq n_{AB})$, $P(N_{AB}\geq n_{AB})$, $P_{\text{DOUBLED ONE-SIDED}}$, $P_{\text{TWO-SIDED}}$ and P_{MID}) as well as the χ^2-statistics (X^2) and *p*-values according to a χ^2-test without continuity correction for HWE.

closer to the nominal significance level, and this holds for all sample sizes and significance levels. This can be appreciated in Figure 2 where the green line corresponding to the mid *p*-value is always the most close to the nominal level. Figure 2 also shows that the rejection rate is closest to the nominal level when the allele frequency is close to 0.5. For low minor allele frequencies the rejection rates are more erratic and differ to a larger extent from the nominal level. Using the mid *p*-value can lead one to exceed the nominal rate, though it typically exceeds the nominal rate by only a small amount. The amount by which the mid *p*-value exceeds the nominal level is always smaller than or equal to the amount by which the other tests underrate the nominal level. Figure 2 also shows that increasing the sample size fastens the convergence of the rejection rate towards the nominal level for all tests.

Besides the type I error rate, tests are compared on the basis of their power. In order to compute the power of a test, the distribution of the number of heterozygotes given the minor allele count under the alternative hypothesis is needed. In a seminal paper, Rohlfs and Weir (2008) introduce this distribution, where the degree of disequilibrium is parametrized by θ, given by:

$$\theta = \frac{P_{AB}^2}{P_{AA}P_{BB}}. \tag{7}$$

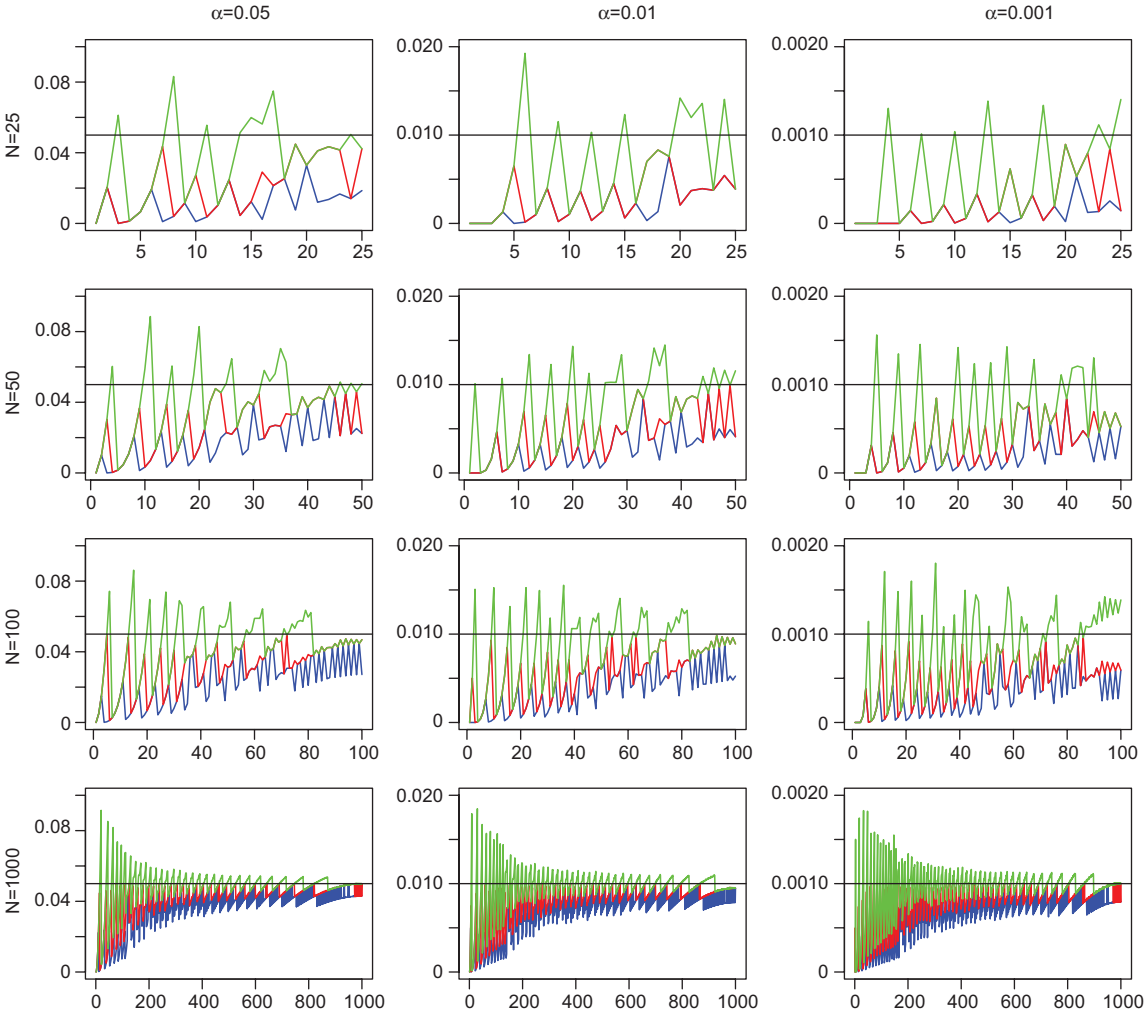

Figure 2 Type I error rate against minor allele count for different sample sizes (25, 50, 100 and 1000) and significance levels (0.05, 0.01, and 0.001) for exact tests with standard two-sided (red), doubled one-sided (blue) and mid *p*-values (green).

Under HWP, we have θ=4. Values of θ>4 imply heterozygote excess, whereas θ<4 means heterozygote dearth. By choosing different values for θ, the degree of disequilibrium (the effect size) can be specified. With (7) Equation (1) can be rewritten as (Rohlfs and Weir, 2008)

$$P(N_{AB}=n_{AB}\,|\,N=n,N_A=n_A)=\frac{\theta^{n_{AB}/2}}{n_{AA}!\,n_{AB}!\,n_{BB}!}\frac{1}{C}, \tag{8}$$

where C is a normalization constant, that depends on the population genotype frequencies. The power of the test, given the value of θ and given a minor allele count, can be computed by summing probabilities according to (8) for all those samples that have a *p*-value below the specified significance level α. These computations were performed for the three types of *p*-value under consideration, using several values of θ (1, 2, 4, 8 and 16) and several sample sizes (25, 50, 100 and 1000) with α=0.05. The resulting power graphics are shown in Figure 3.

Note that for θ=4 the type I error rates shown in Figure 2 are recovered. Also note that for sample sizes below 100, the power of the exact test is in general low. As an indication, with $2\leq\theta\leq 8$ and $N\leq 100$, power is in general below 0.4. The pattern in Figure 3 shows that power typically increases with minor allele frequency. The best power is achieved when *p* is close to 0.5. Figure 3 also shows, as expected, that the exact tests have

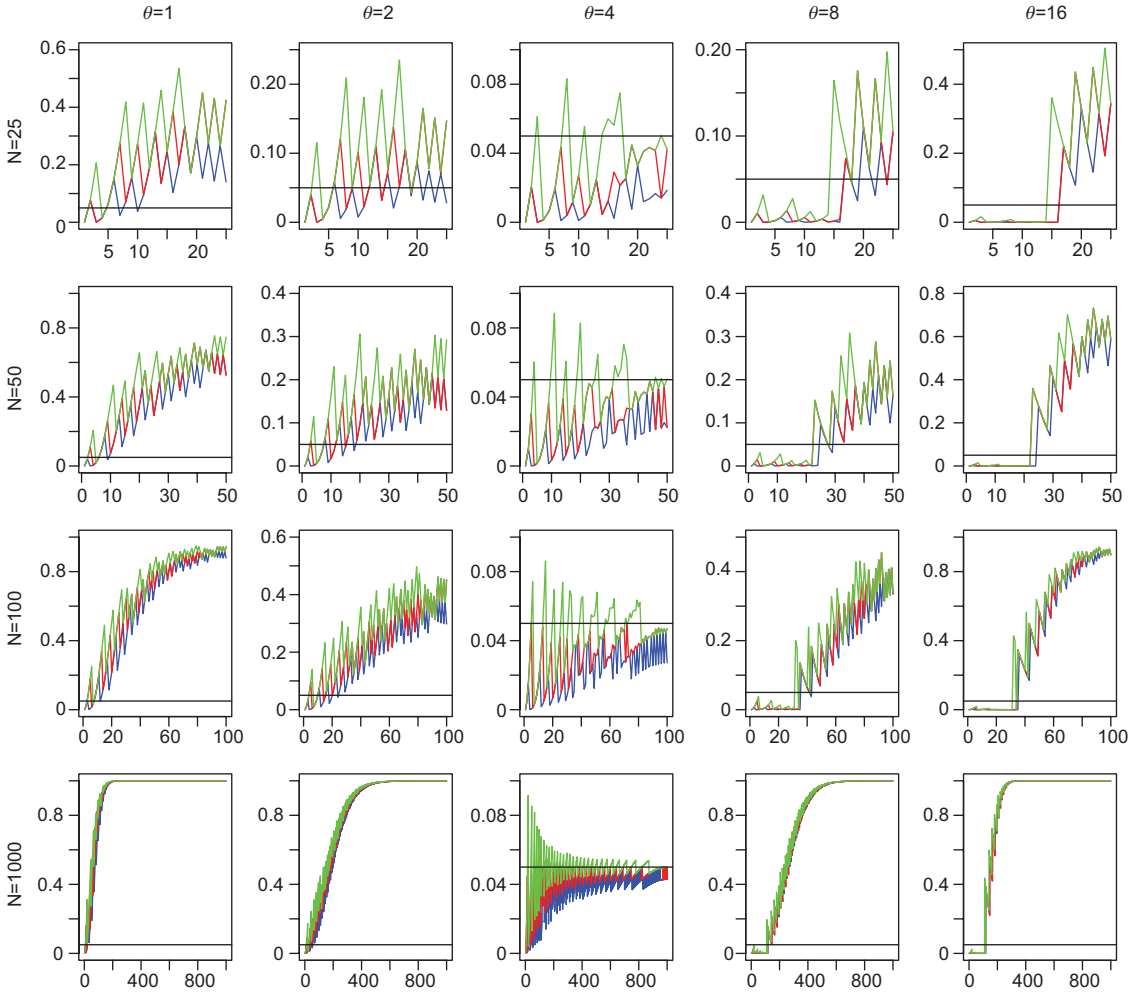

Figure 3 Power of HWP exact tests against minor allele count for different sample sizes (25, 50, 100 and 1000) and degree of disequilibrium (1, 2, 4, 8 and 16). Standard two-sided (red), doubled one-sided (blue) and mid *p*-values (green).

better power when there is a larger degree of disequilibrium (e.g., θ=1 or θ=16). Note that the power graphics are not "symmetric" with respect to θ in the sense that with θ=16 power is generally lower than with θ=1 (both representing a 4-fold change) in particular for the lower allele frequencies. Note in this respect the particular low power for θ=8 and θ=16 for minor allele counts below 30. Most interestingly, Figure 3 shows that an exact test based on the mid p-value has a power that is always as large as or larger than the standard two-sided p-value. The latter statement about the mid p-value is universal: it is observed for all disequilibrium values, all minor allele frequencies and all sample sizes. The figures show, as expected, better power for the larger sample sizes. However, even for samples as large as 1000 individuals, low power is observed for low MAF and disequilibrium values of 2, 8 or 16. Power is, in particular, low for those values of θ in combination with a low MAF, because this combination of parameters results in low frequencies for both the heterozygous and the rare homozygous genotypes.

4 Practical implications

In this section we work out the practical implications of using the mid p-value instead of the standard two-sided p-value. For this purpose we use two data-sets. The first data-set is a genome-wide large selection of markers from the Hapmap project genotyped in a relatively small sample of subjects (n=45). The second data-set is a smaller dataset of markers used in an association study on colon cancer (Landi et al., 2005) with a larger sample size (377 cases and 329 controls).

4.1 Genome-wide HWP-analysis of HapMap data with the mid p-value

We used the full HapMap database (The International HapMap Consortium, 2003, 2005, 2007), using all chromosomes of the Han Chinese sample from Beijing (CHB) consisting of 45 unrelated individuals (phase II, NCBI build 35). This is an unfiltered database, and SNPs were filtered prior to HWP analysis as previously described in Graffelman (2010). We chose this database because it gives an impression of the degree of disequilibrium in a raw database, since testing for HWP is a data quality control requirement to identify markers with genotyping problems. Other databases available in the HapMap website have already filtered markers precisely by leaving out SNPs with a two-sided exact p-value below 0.001. For the HapMap project the choice of the type of p-value has consequences for the admission of markers to the project as is illustrated below.

All available markers were tested by a two-sided exact test, using all three types of p-values described. Results are summarized per chromosome in Table 2. This table shows the number of SNPs on each chromosome, and the percentage of SNPs for which HWP were rejected by exact tests with the three types of p-values, using three different nominal significance levels (0.001, 0.01 and 0.05). At the significance level α=0.001, the rejection rates of all tests are invariably much higher than 0.1%, around 1% for most chromosomes. Under a theoretical model of independent markers that are in equilibrium, with genotypes counts following a multinomial distribution, we would expect to find around 0.1% significant SNPs, but we find roughly 10 times as many. If we would however, use a significance level of 5% (α=0.05), then rejection rate for two-sided and mid p-values is somewhat lower than the theoretical level, between 2 and 4%. Thus, at the 5% level we are below the nominal rate, at the 1% level we exceed the nominal rate, and at the 0.1% level we exceed the nominal rate ten-fold. This indicates that the HapMap SNPs that are significant have a tendency to be *very* significant, indicating that disequilibrium probably arises because of technical genotyping problems and not merely by chance alone (Hosking et al., 2004). The HWP rejection rate at the 0.001 level is roughly the same for all autosomes (0.7 through 1.4%). For the X and Y chromosomes the rejection rates are lower, which is probably due to the fact that the test uses a smaller sample size (females and males only), and has less power to detect disequilibrium. The extent to which the p-values deviate from their theoretical null distribution can be judged in a Q-Q plot. The null p-value distribution is non-uniform as shown by Rohlfs and Weir (2008), and

Table 2 Rejection rates (in %) for two-sided HWP exact tests.

Chr.	#SNPs	Doubled one-sided			Two-sided			Mid		
		0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
1	337746	2.86	1.70	1.20	3.59	1.90	1.30	4.56	2.34	1.40
2	350757	2.05	0.93	0.58	2.77	1.13	0.66	3.74	1.51	0.72
3	277362	2.27	1.17	0.73	3.05	1.38	0.83	3.94	1.76	0.90
4	264172	2.58	1.34	0.91	3.44	1.58	0.99	4.43	2.04	1.07
5	268099	2.34	1.09	0.68	3.10	1.32	0.75	4.12	1.76	0.83
6	299796	3.16	1.74	1.17	4.03	1.99	1.28	5.07	2.53	1.39
7	232061	2.78	1.57	1.14	3.60	1.77	1.23	4.56	2.21	1.33
8	233975	2.27	1.08	0.65	3.05	1.28	0.72	4.02	1.69	0.79
9	197762	2.18	1.04	0.64	2.99	1.24	0.71	3.98	1.63	0.79
10	232330	2.80	1.51	1.00	3.54	1.75	1.11	4.49	2.20	1.23
11	226012	2.35	1.20	0.78	3.11	1.40	0.87	4.05	1.87	0.94
12	209138	2.46	1.21	0.76	3.25	1.41	0.86	4.15	1.87	0.94
13	173180	2.65	1.40	0.93	3.42	1.65	1.02	4.34	2.07	1.11
14	134267	2.44	1.20	0.78	3.18	1.39	0.88	4.09	1.83	0.96
15	116203	2.74	1.34	0.88	3.54	1.58	0.98	4.49	2.06	1.06
16	118981	2.52	1.25	0.85	3.22	1.48	0.94	4.11	1.91	1.01
17	97890	2.06	1.06	0.70	2.74	1.23	0.75	3.61	1.59	0.81
18	128811	2.25	1.14	0.76	2.95	1.34	0.82	3.85	1.72	0.91
19	62972	2.54	1.29	0.94	3.45	1.50	0.99	4.36	1.91	1.08
20	135235	2.02	1.08	0.73	2.60	1.22	0.80	3.38	1.56	0.88
21	57676	2.51	1.30	0.86	3.54	1.51	0.94	4.52	1.94	1.02
22	63978	1.85	0.88	0.61	2.45	1.01	0.67	3.29	1.30	0.73
X	12973	0.82	0.30	0.23	1.15	0.40	0.24	1.85	0.50	0.26
Y	1804	0.72	0.11	0.00	1.05	0.17	0.06	2.33	0.22	0.06

Rejection rates for different exact tests with doubled one-sided, standard two-sided and mid *p*-values per chromosome for the HapMap CHB population of 45 individuals. Three different nominal significance levels are considered (α=0.05, 0.01 and 0.001).

is different for the three types of *p*-values considered. A Q-Q plot of the mid *p*-values and a ternary plot of the first 5000 SNPs on chromosome 1 are shown in Figure 4. Markers simulated under HWP and conditioned to have the same sample size and allele frequency as the observed markers are also shown in the figure. The Q-Q plot shows that the lower tail of the *p*-value distribution is very different from what is expected under the null, and that many markers tend to be too significant in comparison with the null. This is in agreement with our findings in Table 2. The ternary plot also shows that the observed significant markers tend to have larger deviations from HWP, some characterized by heterozygote excess and others by heterozygote dearth. For the HapMap project, using the mid *p*-value instead of the current two-sided *p*-value implies that for most autosomes about 0.1% of the markers currently in the project would not be admitted (see Table 2, columns 8 and 11). Some chromosomes show a high degree of similarity in their rejection rates, in particular chromosomes 8 and 9, and chromosomes 11, 12 and 14.

4.2 HWP-analysis of markers for colon cancer

Landi et al. (2005) report on the analysis of a set of markers scored for 706 individuals (377 cases and 329 controls) possibly involved in colon cancer. These SNPs were genotyped with early microarray technology and might be prone to more genotyping error in comparison with more recent technology. In Table 3 we report HWP test results for 77 markers using the three variations of the exact test previously described. Table 3 illustrates the consequences of adopting the mid *p*-value for HWP tests, and Figure 5 shows HWP Q-Q plots based on the mid *p*-value. The mid *p*-value is, as expected, always the smallest *p*-value. There are many non-

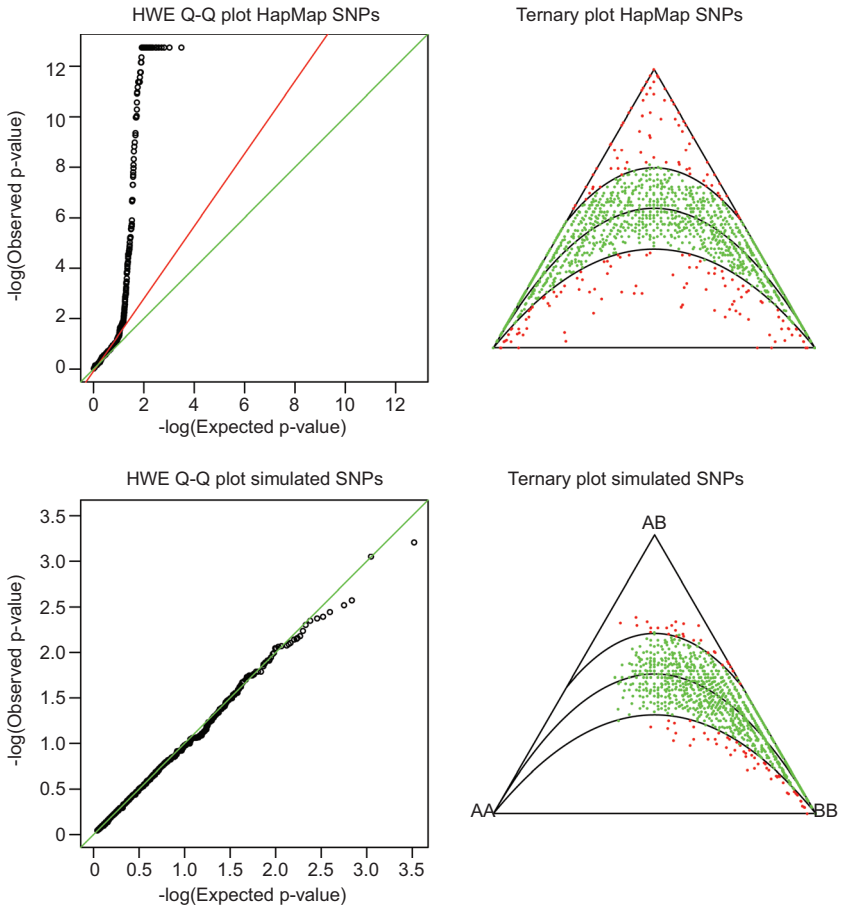

Figure 4 Q-Q plots of mid *p*-values and ternary plots for the HapMap CHB population of 45 individuals. Top panels represent observed markers, bottom panels simulated markers matched in allele frequency and sample size. Red lines in the Q-Q plots are lines through the first and third quartile, green lines are the reference lines with slope 1 and intercept 0. Curves in the ternary plot indicate the HWP parabola and the limits of the HWP acceptance region.

significant markers for which a two-sided *p*-value of 1 is observed. The mid *p*-value is typically much smaller for these samples. We note that markers UGT1A7.173 and NAT2.58 are the most significant ones. For controls, 10 significant markers are observed when the two-sided *p*-value is used, whereas 11 are found with the mid *p*-value (α=0.05). For cases these numbers are 15 and 18 for two-sided and mid *p*-value, respectively. If cases and controls are not distinguished, 19 significant results are observed for both type of *p*-values. The markers that became significant precisely by using mid *p*-values are CYP1B1.156, CYP2A6.107, CYP2D6.115 (cases) and CYP2E1.3 (controls), and are marked in bold in Table 3. For some of these markers, the difference in *p*-values for cases and controls are particularly large, which signals that these markers do deserve special attention, since departure of HWP in cases is an indication of association of the marker with the disease (Lee, 2003).

The Q-Q plots in Figure 5 for the data (top two panels) were obtained by plotting the mid *p*-values against the quantiles of the theoretical null distribution of the mid *p*-value. The latter distribution is, like the theoretical null distribution of the two-sided *p*-value (Rohlfs and Weir, 2008), not uniform. The Q-Q plots reveal that the *p*-value distributions of both cases and controls are not in agreement with HWP. The cases show larger deviations from HWP than the controls. Figure 5 also shows Q-Q plots of SNPs that were simulated under HWP with each SNP in the simulation being matched to an observed SNP in terms of sample size and allele frequency. These simulated SNPs are seen to follow the null distribution indicated by the green line, and serve as a reference against which the top two panels can be compared. The ternary plots in Figure 5 show that there is heterozygote dearth for both controls and cases, and that the extent of heterozygote dearth is

Table 3 Two-sided, doubled one-sided and mid p-values of exact test for HWP for 77 markers possibly involved in colon cancer.

Marker	Controls						Cases					
	AA	AB	BB	Two-sided	Doubled one-sided	Mid	AA	AB	BB	Two-sided	Doubled one-sided	Mid
CYP2E1.1	222	78	5	0.6478	0.7418	0.5656	266	78	4	0.8030	0.7384	0.7154
CYP2E1.2	0	15	277	1.0000	1.0000	0.5842	1	16	313	0.2119	0.4238	0.1156
CYP2E1.3	**245**	**8**	**1**	**0.0863**	**0.1726**	**0.0444**	**296**	**15**	**0**	**1.0000**	**1.0000**	**0.5794**
CYP2E1.5	10	90	195	1.0000	1.0000	0.9237	13	92	235	0.3420	0.3727	0.3014
CYP2E1.6	263	41	0	0.3787	0.4709	0.2609	314	28	1	0.4846	0.9692	0.3061
SULT1A1.26	30	101	171	0.0126	0.0182	0.0101	33	120	195	0.0283	0.0361	0.0239
SULT1A1.27	299	3	0	1.0000	1.0000	0.5025	339	8	0	1.0000	1.0000	0.5200
COMPT.37	65	139	94	0.3491	0.3589	0.3217	86	158	95	0.2318	0.2475	0.2123
COMPT.38	84	161	57	0.2455	0.2475	0.2249	113	168	61	1.0000	1.0000	0.9560
COMPT.40	93	149	58	1.0000	1.0000	0.9536	93	174	80	1.0000	1.0000	0.9572
COMPT.42	82	153	61	0.5595	0.5993	0.5216	108	169	66	1.0000	1.0000	0.9563
NAT1.47	273	12	0	1.0000	1.0000	0.5557	319	10	0	1.0000	1.0000	0.5335
NAT1.48	265	8	0	1.0000	1.0000	0.5253	314	11	0	1.0000	1.0000	0.5412
NAT1.49	0	9	224	1.0000	1.0000	0.5378	0	9	249	1.0000	1.0000	0.5342
NAT1.50	0	10	278	1.0000	1.0000	0.5382	0	14	323	1.0000	1.0000	0.5644
NAT1.51	282	6	0	1.0000	1.0000	0.5130	333	6	0	1.0000	1.0000	0.5110
NAT1.52	0	14	287	1.0000	1.0000	0.5716	3	8	330	0.0001	0.0003	0.0001
NAT1.54	10	55	143	0.1551	0.1910	0.1277	20	59	170	0.0002	0.0002	0.0002
NAT2.56	0	1	304	1.0000	1.0000	0.5000	0	0	349	1.0000	1.0000	0.5000
NAT2.57	134	109	33	0.1638	0.1767	0.1456	142	101	37	0.0081	0.0106	0.0066
NAT2.58	47	94	87	0.0284	0.0315	0.0244	68	78	107	0.0000	0.0000	0.0000
NAT2.59	20	122	146	0.4685	0.5283	0.4249	24	133	175	1.0000	0.9922	0.9454
NAT2.60	94	138	46	0.8046	0.8150	0.7584	98	153	70	0.5000	0.5337	0.4658
NAT2.61	0	13	286	1.0000	1.0000	0.5624	0	14	326	1.0000	1.0000	0.5638
NAT2.62	105	128	40	0.8986	0.9982	0.8479	127	124	51	0.0364	0.0409	0.0317
ADH2.64	2	47	247	1.0000	1.0000	0.8541	2	52	281	1.0000	1.0000	0.8577
GSTM3.68	13	85	188	0.4478	0.4852	0.3984	13	77	251	0.0334	0.0514	0.0255

(**Table 3** Continued)

Marker	Controls						Cases					
	AA	AB	BB	Two-sided	Doubled one-sided	Mid	AA	AB	BB	Two-sided	Doubled one-sided	Mid
GSTP1.69	142	115	34	0.1733	0.1890	0.1545	151	136	26	0.5929	0.6661	0.5468
GSTP1.70	271	32	2	0.2768	0.5537	0.1761	311	35	0	1.0000	0.8088	0.7978
GSTT2.72	1	19	283	0.3014	0.6029	0.1719	1	21	323	0.3155	0.6310	0.1815
GSTA2.73	111	125	54	0.0870	0.0957	0.0771	119	153	49	1.0000	1.0000	0.9533
GSTA4.103	2	26	275	0.1568	0.3136	0.0927	2	24	321	0.0980	0.1961	0.0556
GSTA4.104	86	89	37	0.1102	0.1249	0.0961	83	82	46	0.0044	0.0056	0.0036
GSTA4.105	59	160	81	0.2458	0.2667	0.2240	71	158	93	0.8230	0.8666	0.7801
CYP2C19.108	7	73	225	0.6431	0.8358	0.5612	7	80	253	0.8194	0.9447	0.7347
MDM2.113	241	38	0	0.6241	0.5176	0.4947	265	67	1	0.2295	0.2121	0.1862
CYP2C9.129	221	71	5	1.0000	1.0000	0.9024	242	84	13	0.1092	0.1516	0.0891
CYP2C9.130	259	45	0	0.3884	0.3445	0.3022	292	55	2	1.0000	1.0000	0.8622
CYP1A1.132	4	65	228	1.0000	1.0000	0.8932	3	66	278	1.0000	0.9633	0.8875
CYP1A1.133	292	0	0	1.0000	1.0000	0.5000	337	0	0	1.0000	1.0000	0.5000
CYP1A1.134	280	24	0	1.0000	1.0000	0.6882	319	28	1	0.4795	0.9589	0.3020
CYP1A1.135	0	38	263	0.6171	0.5752	0.4733	0	23	325	1.0000	1.0000	0.6566
CYP1A1.136	2	46	253	1.0000	1.0000	0.8529	1	61	283	0.3360	0.3940	0.2596
CYP1A1.137	291	9	0	1.0000	1.0000	0.5295	338	6	1	0.0402	0.0804	0.0203
CYP1A1.138	41	155	104	0.1847	0.2092	0.1658	41	153	145	1.0000	1.0000	0.9525
CYP1A1.139	234	61	4	1.0000	1.0000	0.8896	277	64	2	0.5548	0.6548	0.4551
CYP1A1.140	0	2	294	1.0000	1.0000	0.5008	0	1	345	1.0000	1.0000	0.5000
CYP1A2.141	0	11	278	1.0000	1.0000	0.5462	1	10	311	0.0991	0.1982	0.0513
CYP1A2.143	0	9	264	1.0000	1.0000	0.5324	1	13	301	0.1568	0.3137	0.0833
CYP1A2.144	147	132	24	0.4942	0.5612	0.4514	133	168	46	0.5685	0.6338	0.5300
CYP1A2.145	52	135	116	0.2319	0.2798	0.2086	71	164	106	0.6613	0.6650	0.6237

(**Table 3** Continued)

Marker	Controls						Cases					
	AA	AB	BB	Two-sided	Doubled one-sided	Mid	AA	AB	BB	Two-sided	Doubled one-sided	Mid
ALDH2.146	198	94	11	1.0000	1.0000	0.9263	223	105	12	1.0000	1.0000	0.9298
ALDH2.147	8	72	208	0.5006	0.6773	0.4296	9	84	230	0.6759	0.8048	0.6028
ALDH2.149	302	0	0	1.0000	1.0000	0.5000	341	1	0	1.0000	1.0000	0.5000
ALDH2.150	8	86	209	1.0000	1.0000	0.9190	9	98	232	0.8491	0.9146	0.7751
CYP3A4.155	3	51	241	0.7421	1.0000	0.6193	2	70	268	0.3998	0.4171	0.3304
CYP1B1.156	**158**	**80**	**9**	**1.0000**	**0.9747**	**0.9218**	**167**	**101**	**27**	**0.0504**	**0.0636**	**0.0425**
CYP1B1.159	97	138	59	0.4764	0.4916	0.4422	121	149	63	0.1748	0.1861	0.1586
CYP1B1.161	187	88	16	0.2089	0.2510	0.1801	210	107	15	0.7390	0.8653	0.6768
MTHFR.168	107	131	59	0.1209	0.1251	0.1087	124	156	61	0.3693	0.3738	0.3423
MTHFR.169	163	118	21	1.0000	1.0000	0.9412	183	138	24	0.8912	0.9051	0.8383
TPMT.170	0	6	292	1.0000	1.0000	0.5125	0	7	335	1.0000	1.0000	0.5153
TPMT.171	1	16	250	0.2564	0.5128	0.1427	3	24	277	0.0283	0.0567	0.0158
TPMT.172	281	23	0	1.0000	1.0000	0.6755	310	36	2	0.3110	0.6220	0.2020
UGT1A7.173	160	19	28	0.0000	0.0000	0.0000	186	20	19	0.0000	0.0000	0.0000
CYP2A6.107	**0**	**19**	**280**	**1.0000**	**1.0000**	**0.6278**	**2**	**21**	**317**	**0.0670**	**0.1340**	**0.0370**
CYP2A6.164	22	134	111	0.0395	0.0495	0.0333	26	145	129	0.1164	0.1311	0.1025
CYP2A6.166	4	16	238	0.0009	0.0019	0.0005	1	26	257	0.5032	1.0000	0.3206
CYP2D6.115	**1**	**19**	**246**	**0.3368**	**0.6736**	**0.1953**	**3**	**29**	**283**	**0.0601**	**0.1201**	**0.0351**
CYP2D6.116	159	63	16	0.0133	0.0168	0.0105	190	74	14	0.0705	0.0978	0.0567
CYP2D6.117	15	51	181	0.0003	0.0006	0.0002	10	58	218	0.0236	0.0427	0.0164
CYP2D6.119	183	57	13	0.0094	0.0122	0.0073	208	66	9	0.2345	0.2730	0.1978
CYP2D6.120	18	63	201	0.0003	0.0006	0.0002	15	71	237	0.0050	0.0078	0.0037
CYP2D6.120B	90	117	57	0.1053	0.1273	0.0924	98	123	70	0.0130	0.0156	0.0110
CYP2D6.121	203	1	0	1.0000	1.0000	0.5000	182	0	0	1.0000	1.0000	0.5000
CYP2D6.123	127	87	42	0.0002	0.0003	0.0002	115	103	50	0.0028	0.0038	0.0023
CYP2D6.124	102	143	57	0.6376	0.6462	0.5979	122	156	65	0.2241	0.2661	0.2032

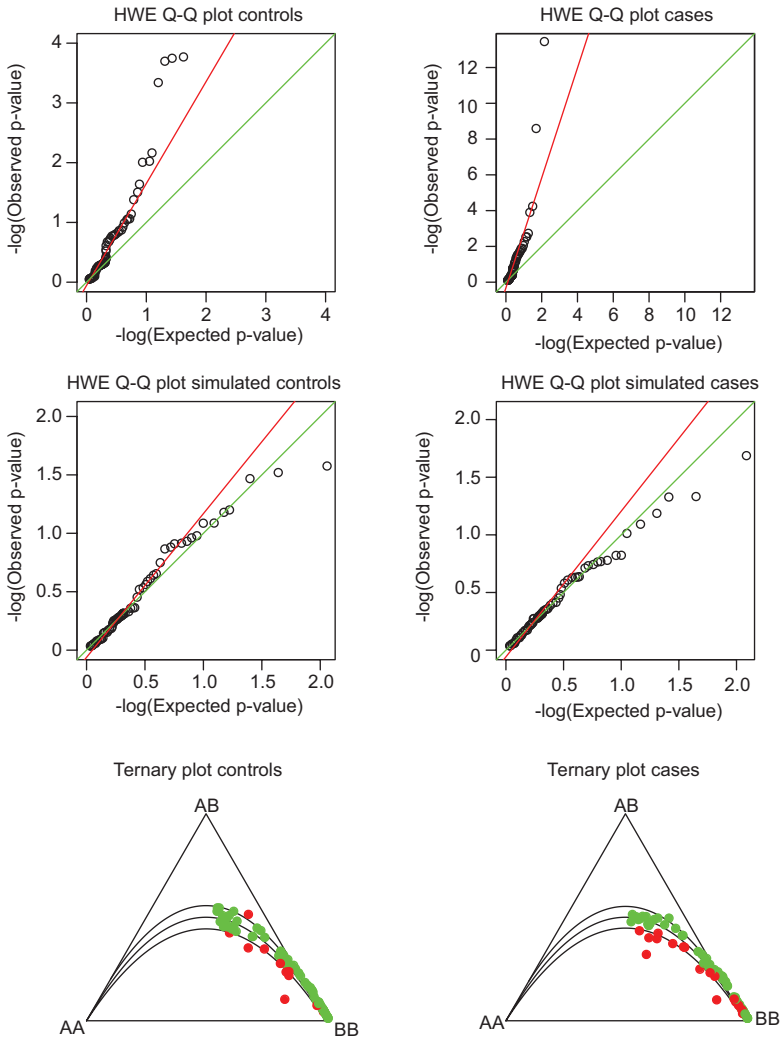

Figure 5 Q-Q plots of mid *p*-values and ternary plots for colon cancer data. Controls are represented in the left column, cases in the right column. Red lines in the Q-Q plots are lines through the first and third quartile, green lines are the reference lines with slope 1 and intercept 0. Curves in the ternary plot indicate the HWP parabola and the limits of the HWP acceptance region.

larger for cases. The results suggest that the observed degree of HW disequilibrium is partly due to genotyping error, and that cases show an additional amount of disequilibrium that may be explained by disease association.

5 Discussion

In genetic studies with large databases of SNPs, a two-sided test for HWP will usually be the most adequate test, because there are usually no a priori reasons to expect excess or dearth of heterozygotes. Figure 2 shows that the exact test based on the mid *p*-value is the most well behaved test because its rejection rate under the null hypothesis is closer to the nominal significance level. Moreover, under all possible combinations of MAF, sample size and degree of disequilibrium, tests based on the mid *p*-value were found to have better power. Based on these results, the mid *p*-value, hitherto largely unknown in genetics, is the preferred *p*-value for an exact test for Hardy-Weinberg proportions.

A change from the standard two-sided *p*-value to the mid *p*-value implies that more evidence against HWP will be found, because the mid *p*-value is smaller or equal to the two-sided *p*-value. If the observed sample is highly unlikely, then the mid and two-sided *p*-values will be virtually identical. However, if the observed sample is highly likely, then the mid *p*-value can be 50% off the two-sided *p*-value, that is $\frac{1}{2}p_{\text{TWO-SIDED}} \leq p_{\text{MID}} \leq p_{\text{TWO-SIDED}}$.

The colon cancer data set suggests that the HWP test with the mid *p*-value is a useful tool. For 3 of the 4 additional significant markers (see Table 3) there is no evidence against HWP in controls, but considerable evidence among cases. Moreover, for 1 of these markers (CYP1B1.156) a co-dominant test for disease association was found significant. This suggests that a HWP test with the mid *p*-value can detect potentially disease-related markers that would have gone unnoticed in a standard HWP test.

6 Software

The R package HardyWeinberg, version 1.5. available from http://www-eio.upc.es/~jan and http://www.r-project.org includes a routine for the exact test that is capable of computing all three *p*-values considered in this paper, routines for Q-Q plots that use the appropriate non-uniform reference distribution for each *p*-value, as well as functions for the computation of power and sample size.

Acknowledgements: This study was supported by grants ECO2011-28875 and CODA-RSS MTM2009-13272 of the Spanish Ministry of Education and Science. The colorectal cancer study was supported by the Instituto de Salud Carlos III grants FIS PI08/1635, FIS PI08/1359 and CIBERESP CB07/02/2005. The authors thank the referees for their comments which have helped to improve the paper.

References

Agresti, A. (2002): Categorical data analysis. New York: John Wiley & Sons, second edition.

Attia, J., A. Thakkinstian, P. McElduff, E. Milne, S. Dawson, R. J. Scott, N. de Klerk, B. Armstrong and J. Thompson (2010): "Detecting genotyping error using measures of degree of Hardy-Weinberg disequilibrium," Stat. Appl. Genet. Mol. Biol., 9, doi: 10.2202/1544-6115.1463.

Ayres, K. L. and D. J. Balding (1998): "Measuring departures from Hardy-Weinberg: a markov chain monte carlo method for estimating the inbreeding coefficient," Heredity, 80, 769–777.

Barnard, G. A. (1989): "On alleged gains in power from lower p-values," Stat. Med., 8, 1469–1477.

Berry, G. and P. Armitage (1995): "Mid-p confidence intervals: a brief review," J. Roy. Stat. Soc. D, 44, 417–423.

Chakraborty, R. and Y. Zhong (1994): "Statistical power of an exact test of Hardy-Weinberg proportions of genotypic data at a multiallelic locus," Hum. Hered., 44, 1–9.

Elston, R. C. and R. Forthofer (1977): "Testing for Hardy-Weinberg equilibrium in small samples," Biometrics, 33, 536–542.

Emigh, T. H. (1980): "A comparison of tests for Hardy-Weinberg equilibrium," Biometrics, 36, 627–642.

Graffelman, J. (2010): "The number of markers in the hapmap project: some notes on chi-square and exact tests for Hardy-Weinberg equilibrium," Am. J. Hum. Genet., 86, 813–818.

Graffelman, J. and J. Morales-Camarena (2008): "Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot," Hum. Hered., 65, 77–84.

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

Haldane, J. B. S. (1954): "An exact test for randomness of mating," J. Genet., 52, 631–635.

Hirji, K. F. (1991): "A comparison of exact, mid-p, and score tests for matched case-control studies," Biometrics, 47, 487–496.

Hosking, L., S. Lumsden, K. Lewis, A. Yeo, L. McCarthy, A. Bansal, J. Riley, I. Purvis, and C. Xu (2004): "Detection of genotyping errors by Hardy-Weinberg equilibrium testing," Eur. J. Hum. Genet., 12, 395–399.

Lancaster, H. O. (1961): "Significance tests in discrete distributions," J. Am. Stat. Assoc., 56, 223–234.

Landi, S., F. Gemignania, V. Moreno, L. Gioia-Patricola, A. Chabrier, E. Guino, M. Navarro, J. de Oca, F. Capella and F. Canzian (2005): "A comprehensive analysis of phase i and phase ii metabolism gene polymorphisms and risk of colorectal cancer," Pharmacogenetics and Genomics, 15, 535–546.

Lee, W. C. (2003): "Searching for disease-susceptibility loci by testing for Hardy-Weinberg disequilibrium in a gene bank of affected individuals," Am. J. Epidemiol., 158, 397–400.

Levene, H. (1949): "On a matching problem arising in genetics," Ann. Math. Stat., 20, 91–94.

Lindley, D. V. (1988): Statistical inference concerning Hardy-Weinberg equilibrium. In Bernardo, J. M., M. H. DeGroot, D. V. Lindley and A. F. M. Smith, (Eds.), Bayesian Statistics, 3, Oxford: Oxford University Press, pp. 307–326.

Okamoto, M. and G. Ishii (1961): "Test of independence in intraclass 2×2 tables," Biometrika, 48, 181–190.

Rohlfs, R. V. and B. S. Weir (2008): "Distributions of Hardy-Weinberg equilibrium test statistics," Genetics, 180, 1609–1616.

Salanti, G., G. Amountza, E. E. Ntzani and J. P. A. Ioannidis (2005): "Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power," Eur. J. Hum. Genet., 13, 840–848.

Shoemaker, J., I. Painter and B. S. Weir (1998): "A bayesian characterization of Hardy-Weinberg disequilibrium," Genetics, 149, 2079–2088.

Smith, C. A. B. (1986): "Chi-squared tests with small numbers," Ann. Hum. Genet., 50, 163–167.

The International HapMap Consortium (2003): "The international hapmap project," Nature, 426, 789–796.

The International HapMap Consortium (2005): "A haplotype map of the human genome," Nature, 437, 1299–1320.

The International HapMap Consortium (2007): "A second generation human haplotype map of over 3.1 million snps," Nature, 449, 851–861.

Wakefield, J. (2010): "Bayesian methods for examining Hardy-Weinberg equilibrium," Biometrics, 66, 257–265.

Weir, B. S. (1996): Genetic Data Analysis II. Massachusetts: Sinauer Associates.

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. (1984): "Tests of significance for 2×2 contingency tables," J. Roy. Stat. Soc A., 147, 426–463.

Yu, C., S. Zhang, C. Zhou, and S. Sile (2009): "A likelihood ratio test of population Hardy-Weinberg equilibrium of case-control studies," Genet. Epidemiol., 33, 275–280.