Chapter 6

Testing Departure from Hardy–Weinberg Proportions

Jian Wang and Sanjay Shete

Abstract

The Hardy–Weinberg principle, one of the most important principles in population genetics, was originally developed for the study of allele frequency changes in a population over generations. It is now, however, widely used in studies of human diseases to detect inbreeding, populations stratification, and genotyping errors. For assessment of deviation from the Hardy–Weinberg proportions in data, the most popular approaches include the asymptotic Pearson's chi-square goodness-of-fit test and the exact test. The Pearson's chi-square goodness-of-fit test is simple and straightforward, but it is very sensitive to small sample size or rare allele frequency. The exact test of Hardy–Weinberg proportions is preferable in these situations. The exact test can be performed through complete enumeration of heterozygote genotypes or on the basis of the Markov chain Monte Carlo procedure. In this chapter, we describe the Hardy–Weinberg principle and the commonly used Hardy–Weinberg proportions tests and their applications, and we demonstrate how the chi-square test and exact test of Hardy–Weinberg proportions can be performed step-by-step using the popular software programs SAS, R, and PLINK, which have been widely used in genetic association studies, along with numerical examples. We also discuss recent approaches for testing Hardy–Weinberg proportions in case–control study designs that are better than traditional approaches for testing Hardy–Weinberg proportions in controls only. Finally, we note that deviation from the Hardy–Weinberg proportions in affected individuals can provide evidence for an association between genetic variants and diseases.

Key words: Hardy–Weinberg proportion, Exact test, Pearson's chi-square goodness-of-fit test, Genetic association study, Quality control, Genotyping error, R, SAS/Genetics, PLINK, Case–control genetic association study, Population stratification

1. Introduction

1.1. What Is the Hardy–Weinberg Proportion?

The Hardy–Weinberg principle, derived independently by Castle (1), Hardy (2), and Weinberg (3), is one of the most important principles in population genetics (4). The Hardy–Weinberg principle states that, in the absence of natural selection, mutation, migration, nonrandom mating, random genetic drift, gene flow, and meiotic drive, the genotypic frequencies and the allele frequencies of a population remain constant from one generation to the next, and furthermore, the genotypic frequencies can be expressed as a

Robert C. Elston et al. (eds.), *Statistical Human Genetics: Methods and Protocols*, Methods in Molecular Biology, vol. 850, DOI 10.1007/978-1-61779-555-8_6,

Table 1
Punnett square for inferring genotypic frequencies from allele frequencies under assumption of the Hardy–Weinberg principle

	A(p)	**a($1-p$)**
A(p)	AA(p^2)	Aa[$p(1-p)$]
a($1-p$)	Aa[$p(1-p)$]	aa[$(1-p)^2$]

simple function of allele frequencies (5). The Hardy–Weinberg principle is now more commonly used in human studies to detect inbreeding, population stratification, and genotyping errors.

Consider a simple case of two alleles, A and a, at a single locus. If the allele frequency of A is denoted as p, then the allele frequency of a is $(1-p)$. If the Hardy–Weinberg principle holds, the expected frequencies of the three possible genotypes, AA homozygotes, Aa heterozygotes, and aa homozygotes are the products of allele frequencies p^2, $2p(1-p)$, and $(1-p)^2$, respectively (Table 1). The expected genotypic frequencies are called Hardy–Weinberg proportions. Whether the observed genotypic frequencies conform to the expected frequencies in a study sample is the very first question in population genetics. The departure from the Hardy–Weinberg proportion is tested by comparing the differences between observed and expected genotypic frequencies. This test is commonly referred to as the Hardy–Weinberg equilibrium test, but it is more accurate to refer to it as the Hardy–Weinberg proportion test, because Hardy–Weinberg equilibrium refers to a state of equilibrium with unchanged allele frequencies and genotypic frequencies over generations, whereas the Hardy–Weinberg proportions are the genotypic frequencies achieved in one generation. Therefore, we consider the terminology "departure from Hardy–Weinberg proportion" in a sample the most appropriate for genetic association studies and use it throughout this chapter.

1.2. Why Test for Deviation from the Hardy–Weinberg Proportion?

Deviations from Hardy–Weinberg proportions can result from evolutionary forces such as inbreeding, assortative mating, and small population size. Inbreeding is mating between close relatives, which can cause a decrease in heterozygosity across the genome in the population, that is, an increase in the number of homozygous genotypes in the individuals (5). In a simple two-allele situation with inbreeding, the inbreeding coefficient F (6, 7) can be calculated as one minus the ratio of the observed number of heterozygotes and the expected number of heterozygotes under the assumption of Hardy–Weinberg proportions. If the observed and expected numbers of heterozygotes are the same in the population, F will be equal to zero. Therefore, in this case, the tests for the

deviation from Hardy–Weinberg proportions and for the inbreeding coefficient $F = 0$ are equivalent, and the deviation from Hardy–Weinberg proportions can indicate inbreeding in the population while a nonzero F statistic can indicate either an excess of heterozygotes (negative F statistic) or an excess of homozygotes (positive F statistic) compared to the expected Hardy–Weinberg proportions. Assortative mating, with a mate who has a similar (positive assortative mating) or dissimilar (negative assortative mating) phenotype, can also increase homozygosity for the genes associated with the phenotype. The relationship between the degree of assortative mating in parents, measured by using a weighted covariance, and the degree of the deviation from Hardy–Weinberg proportions in offspring has been presented in the studies of Price (8) and Shockley (9). Small population size can also increase homozygosity in the population (10). When a population is small, the allele frequencies can drift from generation to generation, a process known as genetic drift. Therefore, the Hardy–Weinberg principle can be violated due to the random change of genotypic frequencies resulting from genetic drift.

In addition to serving as an indicator of evolutionary forces, such as inbreeding, the test for deviation from Hardy–Weinberg proportions can also be applied in studies of population genetics to indicate population stratification, admixture, or cryptic relatedness. It has been shown that the unrecognized population structure and cryptic relatedness (unknown to the investigators) might inflate the false-positive rates in genetic association studies (11), and therefore, Hardy–Weinberg proportions need to be carefully investigated before undertaking genetic association studies. Cryptic relatedness occurs when apparently unrelated individuals in a sample actually have a close kinship relationship. The related individuals will increase the homozygosity in the sample, which can lead to deviations from Hardy–Weinberg proportions across the entire genome (12).

If a population is formed from multiple subpopulations, deviation from the Hardy–Weinberg proportions can be observed in the admixed population, even if all the subpopulations are in Hardy–Weinberg proportion (13–17). For example, consider two subpopulations, each having 1,000 individuals. Also, let us assume that the counts for three genotypes, AA, Aa, and aa, are 160, 480, and 360, respectively, in the first subpopulation and 10, 180, and 810, respectively, in the second subpopulation. Then, the A allele frequency is 0.4 and 0.1 in the two subpopulations, respectively. It can be seen that both subpopulations are in perfect Hardy–Weinberg proportion (P-value = 1.0 in both). However, when the two populations are combined, the observed counts of the three genotypes AA, Aa, and aa will be 170, 660, and 1,170, respectively, and the allele frequency of allele A is now 0.25. The expected counts of the three genotypes can be calculated as 125, 750, and 1,125, respectively. The chi-square test of departure

from Hardy–Weinberg proportions gives a highly significant P-value of 8×10^{-8}, which implies the admixed population deviates from Hardy–Weinberg proportions. The combined population can deviate from Hardy–Weinberg proportion even when the allele frequencies in the subpopulations are not too far apart. For example, if in the first subpopulation, the genotypic counts are 190, 480, and 330, respectively, giving a minor allele frequency of 0.43, and in the second subpopulation, the genotypic counts are 120, 420, and 460, respectively, giving a minor allele frequency of 0.33, both subpopulations are in Hardy–Weinberg proportion, with chi-square-based P-values of 0.5105 and 0.1124, respectively. However, the chi-square test of the combined population provides a significant P-value of 0.0442; therefore, the combined population is not in Hardy–Weinberg proportion.

Most commonly, the Hardy–Weinberg proportion test is used as a quality control tool for identifying errors in genotyping before analysis (5, 18–28). Many genotyping errors can cause deviation from Hardy–Weinberg proportions. For example, a mistaken allele due to DNA contamination and allelic dropout due to low quantity or quality of DNA (29) might cause an increase in homozygotes in individuals, and therefore, cause deviations from Hardy–Weinberg proportions. Genotyping errors will result in inflated type I and type II error rates for genetic association studies (30). The Hardy–Weinberg proportion test is considered an essential procedure in genetic case–control association studies (19, 21, 31–33). However, the Hardy–Weinberg proportion test has very low power for detecting genotyping errors, especially when the genotyping error rate is low and the minor allele frequency is not rare. This is because when the genotyping error rates are small, the observed genotype counts will not be significantly different from the expected genotype counts under Hardy–Weinberg proportions and, therefore, any test that attempts to detect such errors based on proportion testing will have very little power (26, 34).

For example, suppose in a sample of 1,000 individuals without genotyping error the observed counts for the three genotypes AA, Aa, and aa are 85, 418, and 497, respectively (Fig. 1). Without genotyping error (panel A), the genetic variant is in Hardy–Weinberg proportion (P-value of Hardy–Weinberg proportion exact test $= 0.8790$). For the purpose of demonstration, we assumed three genotyping error models. In the first error model (panel B) (27), the genotyping error is that both homozygous genotypes (AA and aa) are miscoded as the heterozygous genotype (Aa) with equal probability (i.e., AA $\rightarrow$ Aa and aa $\rightarrow$ Aa). In genotyping error models two and three (panels C and D), we considered the miscoding only from rare homozygotes to heterozygotes and from heterozygotes to rare homozygotes (i.e., AA $\rightarrow$ Aa or Aa $\rightarrow$ AA). In Fig. 1, we demonstrate the variations in P-values of the Hardy–Weinberg proportion test with respect to increased miscoding probabilities of 1, 2.5, and 5%. The P-values were obtained with the use of the exact test of

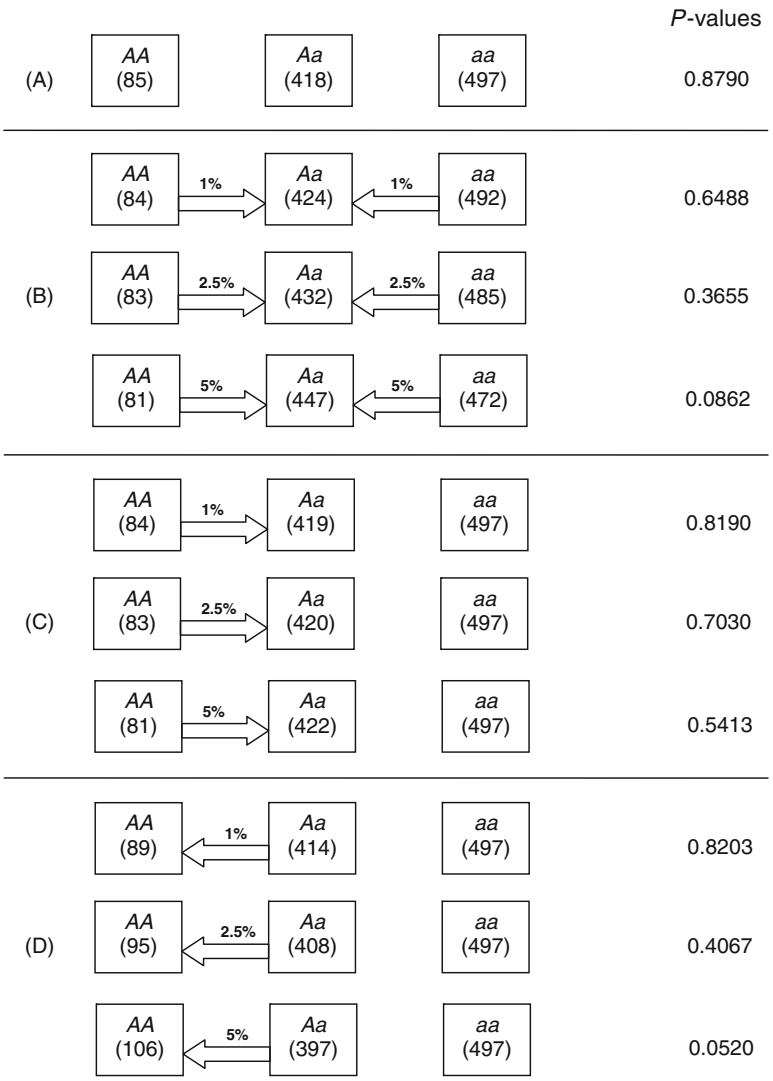

Fig. 1. The Hardy–Weinberg proportion test has poor power to detect genotyping errors when the genotyping error rates are low. *P* -values were obtained from Hardy–Weinberg proportion exact tests. (A) Model without genotyping error; (B) Genotyping error model with some homozygous individuals miscoded as heterozygotes; (C) Genotyping error model with some rare homozygous individuals miscoded as heterozygotes; (D) Genotyping error model with some heterozygous individuals miscoded as rare homozygotes.

Hardy–Weinberg proportions (5). Given a significance level of 5%, all the *P*-values of Hardy–Weinberg proportion tests are nonsignificant, implying that the sample is in Hardy–Weinberg proportion in all the scenarios of all the error models. Even when the overall probability of miscoding is 2.9% in the first error model, the observed genotypic counts are 81, 447, and 472 for AA, Aa, and aa, respectively. The Hardy–Weinberg proportion exact test gives a *P*-value of 0.0862, and thus, the test cannot identify the genotyping errors.

Table 2
Observed and expected genotypic counts for a diallelic locus in a sample with *n* individuals

Genotype	AA	Aa	aa
Observed counts	n_{AA}	n_{Aa}	n_{aa}
Expected counts	$n\hat{p}_A^2$	$2n\hat{p}_A(1-\hat{p}_A)$	$n(1-\hat{p}_A)^2$

$\hat{p}_A$: the estimated A allele frequency from the data
n_{AA}, n_{Aa}, and n_{aa}: observed genotypic counts for three genotypes

With recent advancements in genotyping techniques, the genotyping error rates are quite small (i.e., 0.01% using human GWAS SNP—Illumina). Therefore, the Hardy–Weinberg proportion test will not be a powerful tool for detecting genotype errors. However, the current sequencing technologies have high error rates, which will lead to the higher probability of errors in calling individual genotypes, particularly for rare and novel variants. The relationship between genotyping error and the Hardy–Weinberg proportion test has been studied in the literature (26, 34–39).

In genetic association studies, the genetic variants that deviate from Hardy–Weinberg proportions are usually considered to be genotyping errors and are removed from further analysis. However, such conclusions should be reached with great caution because a departure from Hardy–Weinberg proportions can also be evidence of an association between genetic variants and the disease of interest (12, 16, 17, 27, 28, 33, 40–46).

1.3. How to Test for Deviation from the Hardy–Weinberg Proportion?

To test the deviations from Hardy–Weinberg proportions in a population, the null hypothesis, H_0, is that there is no significant difference between the observed and the expected genotypic counts under Hardy–Weinberg proportions; the alternative hypothesis, H_a, is that there is a significant difference between the observed and expected genotype counts. The commonly used approaches for Hardy–Weinberg tests include the asymptotic Pearson's chi-square goodness-of-fit test and the exact test.

1.3.1. Chi-Square Goodness-of-Fit Test

Pearson's chi-square goodness-of-fit test is the most commonly used approach for testing the departure from Hardy–Weinberg proportions (5, 16). If we consider a sample with n individuals, and denote the observed genotypic counts of AA, Aa, and aa at a single locus as n_{AA}, n_{Aa}, and n_{aa}, respectively (see Table 2), the test statistic of Pearson's chi-square goodness-of-fit test is given as (5):

$$\begin{aligned}\chi^2 &= \sum_{\text{genotypes}} \frac{(\text{Observed counts}-\text{Expected counts})^2}{\text{Expected counts}} \\ &= \frac{\left(n_{AA}-n\hat{p}_A^2\right)^2}{n\hat{p}_A^2} + \frac{\left[n_{Aa}-2n\hat{p}_A(1-\hat{p}_A)\right]^2}{2n\hat{p}_A(1-\hat{p}_A)} + \frac{\left[n_{aa}-n(1-\hat{p}_A)^2\right]^2}{n(1-\hat{p}_A)^2},\end{aligned}$$

where $\hat{p}_A$ is the A allele frequency estimated from the sample data, and $\hat{p}_A = \frac{2n_{AA}+n_{Aa}}{2n}$. The χ^2 test statistic asymptotically follows a chi-square distribution with one degree of freedom. For a multi-allele locus with m alleles, the degrees of freedom is calculated as the number $\binom{m}{2}$ (the number of independent parameters under the alternate hypothesis minus the number of independent parameters under the null hypothesis). In Fig. 1, panel A, when there are no genotyping errors, $n_{AA} = 85$, $n_{Aa} = 418$, and $n_{aa} = 497$, giving a total sample size of $n = 1{,}000$. The estimated allele frequency of A can be evaluated as $\hat{p}_A = 0.294$. Therefore, the expected counts are $1{,}000 \times \hat{p}_A^2 = 86.44$, $1{,}000 \times 2\hat{p}_A(1-\hat{p}_A) = 415.13$, and $1{,}000 \times (1-\hat{p}_A)^2 = 498.44$ for genotypes AA, Aa, and aa, respectively. Using the formula above, we can obtain the following value of the χ^2 statistics:

$$\chi^2 = \frac{(85-86.44)^2}{86.44} + \frac{(418-415.13)^2}{415.13} + \frac{(497-498.44)^2}{498.44} = 0.048.$$

Compared to the chi-square distribution with one degree of freedom, the P-value is 0.8268, which is not statistically significant at a significance level of 5%. Therefore, we do not reject the null hypothesis and can assume that this locus is in Hardy–Weinberg proportion. Also, since the genotypic counts are discrete, the Yates continuity correction of 0.5 can be used (5, 47):

$$\chi^2 = \sum_{\text{genotypes}} \frac{(|\text{Observed counts} - \text{Expected counts}| - 0.5)^2}{\text{Expected counts}}.$$

In this scenario, the P-value obtained here using the asymptotic chi-square test is 0.8733. It needs to be noted that the test statistic χ^2 follows a chi-square distribution asymptotically when the sample size is large. This asymptotic assumption of a chi-square distribution could fail when the sample size is too small or there are not enough genotype counts per cell. A locus with a rare minor allele could also have an impact on the performance of Pearson's chi-square test, even if the total sample size is large, because the expected counts of possible genotypes can still be low or close to zero owing to rare allele frequency, and therefore, can greatly inflate the test statistics. It has been suggested that the asymptotic Pearson's chi-square test for Hardy–Weinberg proportions should not be used if the expected count of a particular genotype is less than some specified number, which is typically five (5, 16). In this situation, the exact test is preferable (5).

1.3.2. Hardy–Weinberg Exact Test

In the exact approach, a test is performed by computing probabilities under the null hypothesis of all possible genotype combinations that have the same allele frequency and total sample size as the

observed sample. Then, the sum of all probabilities of events less or equally probable to the observed event probability is the exact P-value, and the null hypothesis is rejected if it is smaller than a prespecified significance level (5, 48). Consider the notation from Table 2 with n diploid individuals. For the genotypes at a single locus, the conditional probability of observed genotypic counts n_{AA}, n_{Aa}, and n_{aa}, given the observed allele frequencies, can be expressed in terms of the probability of heterozygote counts n_{Aa} conditional on observed counts of the A allele under the assumption of Hardy–Weinberg proportions and sample size. The conditional probability is given as follows (5, 49, 50):

$$\Pr(n_{Aa}|n,\ n_A) = \frac{n!n_A!n_a!2^{n_{Aa}}}{[(n_A - n_{Aa})/2]!n_{Aa}![n - (n_A + n_{Aa})/2]!(2n)!},$$

where $n_A = 2n_{AA} + n_{Aa}$ is the observed count for allele A, n is the sample size, and $n_a = 2n - n_A$. One can evaluate the conditional probabilities for all possible genotypes consistent with the observed data and order the counts of heterozygotes n_{Aa} according to these probabilities. The summation of the conditional probabilities that are less than or equal to the conditional probability of observed genotypes is then calculated as the P-value of the exact test (49, 51, 52).

The exact test is more desirable than Pearson's chi-square test because it is valid for any sample size and minor allele frequency. The exact test can provide an exact P-value for the test of Hardy–Weinberg proportions if one completely enumerates all possible genotypes, as described in the study by Louis and Dempster (53). However, the number of possible genotypes given the same sample size and allele frequencies increases exponentially with the number of alleles (54). Therefore, in practice, it might not be feasible to perform complete enumeration for large samples involving multiple alleles. Efficient algorithms have been proposed to improve the efficiency of the full enumeration algorithm (17, 51, 55, 56). In a recent study (51), Engels presented a new algorithm for full enumeration using recursion, and improved the efficiency by about two orders of magnitude. However, even using the recursion algorithm, complete enumeration is still not practical in some situations. Engels showed that the total number of possible genotypes is 2×10^{56} for the data from the human *Rh* locus (51). In this situation, complete enumeration would certainly be computationally inefficient.

Alternative approaches to full enumeration that are based on permutation or resampling for testing Hardy–Weinberg proportions have been extensively developed (17, 54, 57, 58). The conventional Monte Carlo test of Hardy–Weinberg proportions was first proposed by Guo and Thompson (54). For the Monte Carlo test, one can randomly generate a large number of independent possible genotypes based on the observed allele counts and sample size. Guo and Thompson (54) also adapted the Markov chain

algorithm to the Monte Carlo test by using the Markov chain to approximate the distribution of the test statistic. It has been shown that, when the sample size is relatively large, the Markov chain Monte Carlo (MCMC) algorithm is faster than the direct Monte Carlo algorithm (54). Other improvements to the Monte Carlo- or MCMC-based tests of Hardy–Weinberg proportions have been proposed (57–59). The MCMC-based tests are usually referred to as "exact tests" in the literature and software. However, it should be noted that these approaches are actually not "exact" because they do not enumerate the entire space of possible genotypes. However, compared to complete enumeration, the MCMC-based tests perform favorably and offer enormous improvement in computational time; therefore, they have been extensively applied when complete enumeration is not feasible.

Other approaches for testing Hardy–Weinberg proportions have been proposed, including unconditional exact tests, likelihood ratio tests, a confidence-limit-based approach, and Bayesian approaches (5, 60–65). Some considered the Hardy–Weinberg proportion test from a different point of view and proposed an equivalence test (66). In practice, however, the most popular tests of Hardy–Weinberg proportions remain Pearson's chi-square goodness-of-fit test and the exact tests (complete enumeration or MCMC-based). Although the derivations and examples in this chapter are based on two alleles, both tests can be extended to multiple alleles (5). Many commonly used programs and software (some available at no cost) can perform these two approaches. In Subheading 2, we demonstrate step-by-step how these two tests can be performed using popular software programs that have been widely used in genetic association studies, SAS (67), R (68), and PLINK (69, 70), along with numerical examples. The chi-square test has wider usage than the exact test because it is simpler and more straightforward. However, as we show in Subheading 2, the chi-square test is very sensitive to small expected counts in one cell and, therefore, provides more liberal P-values when the allele is rare or the sample size is small. Therefore, we recommend using the exact test when assessing Hardy–Weinberg proportions.

In the next section, we discuss recent approaches for testing Hardy–Weinberg proportions in case–control study designs that are better than traditional approaches for testing Hardy–Weinberg proportions in controls only. We also note that deviation from the Hardy–Weinberg proportions in affected individuals can provide evidence for an association between genetic variants and diseases. Some investigators have used the Hardy–Weinberg proportion test in cases to identify disease susceptibility genetic loci while others have combined this information with the standard genetic association tests. We would like to discuss these recent approaches briefly in this chapter. But no practical guideline for performing these approaches is provided. For readers who are interested, please

refer to the related papers for the details. In this chapter, we focus on demonstrating how to use the software programs SAS, R, and PLINK to perform the traditional tests of Hardy–Weinberg proportions (i.e., chi-square and exact tests).

1.4. Hardy–Weinberg Proportion in Case–Control Genetic Association Studies

Case–control genetic association studies with unrelated individuals, such as genome-wide association studies, have become a popular and powerful approach for identifying genetic variants associated with complex diseases. The test for the departure from Hardy–Weinberg proportion plays important roles in case–control genetic association studies. The most common usage is to assess the Hardy–Weinberg proportion in control subjects as a quality control measure for identifying genotyping errors. The relationship between genotyping errors and the Hardy–Weinberg proportion test has been studied and discussed in previous studies (26, 34–39). These studies suggest that the Hardy–Weinberg proportion test in controls has very low power for detecting genotyping errors, especially when the genotyping error rate is low and the minor allele frequency is not rare. However, the Hardy–Weinberg proportion test is still considered an essential and routine quality control tool in genetic association studies (19, 21, 31–33).

In general, the Hardy–Weinberg proportion test assumes that the genotypes are sampled from the general population, and therefore, the expected genotype counts in the test should be evaluated from the general population. In a case–control genetic association study, when the Hardy–Weinberg proportion test is performed in control subjects, the observed genotypic counts in controls are compared against the expected genotypic counts in controls. This strategy might work if the disease under consideration is rare, where the controls might well represent the general population. However, when the disease is common in the population, it could be problematic to use only controls when evaluating the expected genotypic counts from the general population, as cases would account for a relatively large portion of the general population. This might lead to artificial departure from Hardy–Weinberg proportions, especially for the markers associated with the disease, and to discarding important SNPs that could potentially be causal SNPs associated with the disease. It has been shown that the type I errors can be inflated dramatically for Hardy–Weinberg proportion tests on the disease-associated markers (23, 26). Moreover, if the genotyping is problematic, it might likely have an impact on both case and control subjects (71). Therefore, the Hardy–Weinberg proportions should be tested in the entire study population rather than only in control subjects. Recently, several new approaches have been proposed for assessing Hardy–Weinberg proportions using both cases and controls for the case–control genetic association (23, 26, 71).

1.4.1. Hardy–Weinberg Proportion Test for Case–Control Study

Likelihood-Based Approach

Li and Li (23) proposed an approach for assessing Hardy–Weinberg proportions based on a general likelihood ratio framework and applied the approach to both case–control and family-based study designs (see Note 1). They considered a di-allelic locus with the three genotypes aa, Aa, and AA as $g = (0, 1, 2)$. The genotype frequencies were denoted as P_0, P_1, and P_2, respectively, where $P_0 = 1 - P_1 - P_2$. The disease status D was defined as a binary variable with 1 representing cases and 0 representing controls. Let the penetrance of the disease conditional on genotypes be $f_g = \Pr(D = 1 | g)$, then the prevalence of the disease can be written as $K = f_0 P_0 + f_1 P_1 + f_2 P_2$. Therefore, given n unrelated cases and m unrelated controls, the likelihood of the sample is given as:

$$L = \frac{1}{K^n (1 - K)^m} \prod_{g=0}^{2} f_g^{n_g} (1 - f_g)^{m_g} P_g^{n_g + m_g},$$

where n_g and m_g are the numbers of genotypes in cases and controls, respectively.

Under the null hypothesis of Hardy–Weinberg proportion, $P_0 = (1 - p)^2$, $P_1 = 2p(1 - p)$, and $P_2 = p^2$, where p is the allele frequency of allele A. The likelihood ratio test compares the likelihood that is maximized under the alternative hypothesis (departure from Hardy–Weinberg proportions) with the likelihood that is maximized under the null hypothesis (in Hardy–Weinberg proportion). The likelihood ratio statistic follows an asymptotic chi-square distribution with one degree of freedom under the null hypothesis.

The likelihood ratio test of population Hardy–Weinberg equilibrium proposed by Yu et al. (71) is similar to the one proposed by Li and Li. In this approach, they fit models by minimizing the deviation function, comparing the observed and expected numbers of genotypes in cases and controls.

Mixture Hardy–Weinberg Proportion Exact Test

Wang and Shete (26) proposed a mixture Hardy–Weinberg proportion (mHWP) exact test, where a mixture sample that mimics the general population is created and employed. The individuals in the mixture sample are randomly selected from the original cases and controls, and the number of cases in the mixture sample is proportional to the prevalence of the disease. Consider a case–control study with n_0 controls and n_1 cases. Let n_m be the sample size of the mixture sample, and let K be the estimated prevalence of disease. One could choose $n_m = \min(\lfloor n_1/K \rfloor, \lfloor n_0/(1 - K) \rfloor)$ to achieve the largest possible mixture sample size and then randomly select $\lfloor n_m \times K \rfloor$ individuals from the cases and $\lfloor n_m \times (1 - K) \rfloor$ individuals from the controls. The exact P-value of the Hardy–Weinberg proportion test can be evaluated using the mixture sample. The procedure is repeated L times to allow for variability in the mixture sampling, and L exact P-values are obtained (see Note 2).

The empirical distribution-based nonparametric density can be constructed based on L mixture sample exact P-values. The maximum likelihood estimator of this empirical distribution is estimated as the final P-value for mHWP in the general population.

When the marker is not associated with the disease, both likelihood-based and mHWP approaches perform similarly to the traditional test using controls only. When the marker is associated with the disease, the traditional test using only controls inflates the type I errors dramatically when minor allele frequencies and prevalence of disease increase (23, 26). However, the likelihood-based approaches and mHWP exact test can still control type I errors well for the disease-associated markers and, therefore, significantly outperform the traditional approach. If genotyping errors are absent, the mHWP exact test provides a conservative approach for assessing Hardy–Weinberg proportions relative to the likelihood-based approaches. Therefore, the mHWP exact test is more likely to retain causal SNPs for future analyses after the Hardy–Weinberg testing. When the genotyping error rates are higher, the genotyping error can generate extreme deviation from Hardy–Weinberg proportions and, therefore, all approaches can have high power to detect genotyping errors. However, when the genotyping error rates are low, likelihood-based approaches and the mHWP exact test are not very sensitive for detecting genotyping errors. Therefore, one may also consider a strategy of keeping all SNPs for the association study, performing the Hardy–Weinberg test only among significant markers.

1.4.2. Hardy–Weinberg Proportion Test for Genetic Association Studies

Researchers also suggest that deviation from Hardy–Weinberg proportions among case subjects can provide additional evidence for an association between genetic markers and the disease of interest (28, 33, 40–42). There is increasing interest in using deviation from Hardy–Weinberg proportions in patients as a tool in genetic association studies for identifying disease-susceptibility loci. Feder et al. (40) proposed to investigate the deviation from Hardy–Weinberg proportions among cases to fine map disease-susceptibility loci for an autosomal recessive disorder. In their paper, the degree of the deviation from Hardy–Weinberg proportions was measured by using the F parameter (known as the inbreeding coefficient), which compares the observed homozygosity and expected homozygosity under Hardy–Weinberg proportions. The markers with higher F values are considered to be closer to the disease susceptibility locus. Their method has been reviewed and extended by subsequent researchers (13, 41–43). Wittke-Thompson et al. (28) examined the directions of the difference between population and expected genotypic frequencies in cases and controls, respectively, and developed a chi-square test for determining whether the observed data in a case–control study are consistent with a genetic disease model.

Other researchers have proposed to combine the information of the departure from Hardy–Weinberg proportions and the commonly used measures of association tests (i.e., logistic regression, allelic association test, and the Cochran-Armitage trend test) to create new statistical genetic association tests (33, 44, 46). The set-association method proposed by Hoh et al. (46) employs the information of departure from Hardy–Weinberg proportions in both cases and controls. They first use the departure from Hardy–Weinberg proportions in controls as a trimming tool to eliminate markers with unusually high statistical values, which might indicate genotyping errors. Then, the departure from Hardy–Weinberg proportions in cases is combined with the allelic association test, through the product of these two test statistics, to form the new test statistic. The significance of the test is obtained by permutation of the disease status. Song and Elston (44) addressed the weakness of the approach proposed by Hoh et al. and developed a similar approach, called weighted average statistic, for fine-mapping disease susceptibility loci. This approach combines the Cochran-Armitage trend test statistic and the Hardy–Weinberg disequilibrium (HWD) trend test statistic, which is proposed in this study to examine the difference between the HWD coefficients in cases and controls. The linear function of the two test statistics using appropriate weights is used to form the new test statistic. The weighted average statistic for identifying disease susceptibility loci has better power than the adjusted Cochran-Armitage trend test, the HWD trend test and the product of these two tests, for all genetic disease models investigated in the study.

Both approaches discussed above use the Hardy–Weinberg proportion information from both cases and controls, and no covariate is considered. Alternatively, Wang and Shete (33) developed a new test statistic for genetic association studies that incorporates evidence about deviation from Hardy–Weinberg proportions only in cases into the regression-based models. With the use of regression models, this approach can easily include covariates in the analysis. The mean-based and median-based tail-strength measures (72) were proposed to combine P-values from two different hypothesis tests: the likelihood ratio test for association and the Hardy–Weinberg proportion exact test in cases. The significance of the new test can be assessed through analytic formulas as well as a resampling procedure. This approach showed a significant increase in power for genetic association studies and good control of type I errors with the additive genetic model. Wang and Shete (73) further pointed out that the analytic formulas for evaluating P-values might cause inflated type I errors for recessive and dominant genetic models, owing to the assumptions underlying the development of asymptotic null distribution; therefore, they recommended using the resampling-based approach to assess the significance of the new statistics. The computer program "CSig" performs the proposed association test through analytic formulas and is available at http://www.epigenetic.org/software.php.

2. Methods

To demonstrate difference in *P*-values obtained using the Pearson's chi-square goodness-of-fit test and the exact test for testing the departure from Hardy–Weinberg proportions, we considered di-allelic SNPs at 18 different genetic loci in a sample of 1,000 individuals. The three genotypic counts for all the SNPs are listed in Table 3. The minor allele frequencies of the SNPs vary from ~1% (rare variants) to ~50% (common variants). We utilized three software programs (SAS, R, and PLINK) to evaluate the *P*-values of Hardy–Weinberg proportion tests for each SNP based on asymptotic and exact approaches.

2.1. SAS/Genetic Software

The tests for the departure from Hardy–Weinberg proportions can be performed by using SAS/Genetics software (67). Although the Pearson's chi-square and Fisher's exact tests might be conducted using statistical procedures in SAS (e.g., PROC FREQ), the procedure ALLELE in the SAS/Genetics software is specially developed for analyzing genetic data, and it provides statistical tests for Hardy–Weinberg proportions based on these two commonly used approaches.

To examine the departure from Hardy–Weinberg proportions of the markers listed in Table 3, we first create the input data file in SAS format as below, based on the genotypic counts of the 18 markers.

```
. . . . . .
a a a a a a A a A A a a A a a a A a A a A a A a A a A A A a a a A a A A
a a A a a a a a A A a a A a a a a a A a A a A a A a A a A a A a A A A A
A A a a A A A a a a a a A a a a A a A a A a a a A a A a A A A A A a A a
. . . . . .
```

The input data include 36 columns, with the first two columns representing the set of two alleles for the first SNP, the third and fourth columns representing the set of alleles for the second SNP, and so on. There are 1,000 rows of data, each representing one individual. The following code reads the input data and conducts Hardy–Weinberg proportion tests for the 18 markers:

```
data markers;
infile 'markers';
input (a1-a36) ($);

proc allele data=markers nofreq perms=100000 seed=12345;
    var a1-a36;
run;
```

Table 3
Genotypic counts, estimated allele frequencies, and *P*-values obtained based on Pearson's chi-square goodness-of-fit test and exact test of Hardy–Weinberg proportions by using three different software programs: SAS, R, and PLINK

				SAS		R		PLINK	
AA	Aa	aa	$\hat{p}_A$	p_chisq*	p_exact†	p_chisq*	p_exact‡	p_chisq*	p_exact‡
3	15	982	0.0105	1.425501E-18	1.400000E-04	1.425501E-18	9.822870E-05	1.43E-18	9.82E-05
1	19	980	0.0105	6.767192E-03	1.006500E-01	6.767192E-03	1.006557E-01	0.006767	0.1007
0	20	980	0.0100	7.494065E-01	1.000000E+00	7.494065E-01	1.000000E+00	0.7494	1
20	50	930	0.0450	6.149249E-40	0.000000E+00	6.149249E-40	1.214089E-17	6.15E-40	1.21E-17
3	97	900	0.0515	8.218825E-01	7.437800E-01	8.218825E-01	7.425484E-01	0.8219	0.7425
3	95	902	0.0505	7.667648E-01	7.340400E-01	7.667648E-01	7.358119E-01	0.7668	0.7358
5	185	810	0.0975	1.053538E-01	1.468900E-01	1.053538E-01	1.471012E-01	0.1054	0.1471
15	155	830	0.0925	1.520538E-02	2.115000E-02	1.520538E-02	2.191775E-02	0.01521	0.02192
10	180	810	0.1000	1.000000E+00	1.000000E+00	1.000000E+00	1.000000E+00	1	1
30	360	610	0.2100	7.195676E-03	7.350000E-03	7.195676E-03	7.435283E-03	0.007196	0.007435
50	250	700	0.1750	2.198160E-05	1.000000E-04	2.198160E-05	6.528187E-05	2.20E-05	6.53E-05
40	320	640	0.2000	1.000000E+00	1.000000E+00	1.000000E+00	1.000000E+00	1	1
60	500	440	0.3100	9.450207E-08	0.000000E+00	9.450207E-08	5.864779E-08	9.45E-08	5.87E-08
100	420	480	0.3100	5.642284E-01	5.584000E-01	5.642284E-01	5.549472E-01	0.5642	0.5549
90	420	490	0.3000	1.000000E+00	1.000000E+00	1.000000E+00	1.000000E+00	1	1
300	400	300	0.5000	2.539629E-10	0.000000E+00	2.539629E-10	2.299897E-10	2.54E-10	2.30E-10
200	500	300	0.4500	7.494065E-01	8.003600E-01	7.494065E-01	7.982999E-01	0.7494	0.7983
250	500	250	0.5000	1.000000E+00	1.000000E+00	1.000000E+00	1.000000E+00	1	1

Note: the *P*-values obtained from R and PLINK are in their default format
$\hat{p}_A$: the estimated A allele frequency from the data
* *P*-values obtained using Pearson's chi-square goodness-of-fit test without continuity correction
† *P*-values obtained using permutation-based Hardy–Weinberg proportion test
‡ *P*-values obtained using exact Hardy–Weinberg proportion test

Alternatively, the input data can be read in a format of columns of genotypes instead of columns of alleles. Using this format, there is only one column for each marker. One can use different characters or strings as delimiters (i.e., "/" or "-"), or even no delimiter, to separate the two alleles for each marker (see below).

```
. . . . . .
a/a a/a a/a A/a A/A a/a A/a a/a A/a A/a A/a A/a A/a A/A A/a a/a A/a A/A
a/a A/a a/a a/a A/A a/a A/a a/a a/a A/a A/a A/a A/a A/a A/a A/a A/A A/A
A/A a/a A/A A/a a/a a/a A/a a/a A/a A/a A/a A/a a/a A/a A/a A/a A/A A/a A/a
. . . . . .
```

When this alternative format is used, there will be only 18 variables and the options GENOCOL and DELIMITER= should be included. The DELIMITER= option can be omitted in the following example since "/" is the default.

```
data markers;
infile 'markers';
input (a1-a18) ($);

proc allele data=markers nofreq perms=100000 seed=12345 genocol
delimiter='/';
    var a1-a18;
run;
```

The ALLELE procedure

Marker summary

								Test for HWE	
Locus	Number of individuals	Number of alleles	PIC	Heterozygosity	Allelic diversity	Chi-square	DF	Pr > ChiSq	Prob exact
M1	1,000	2	0.0206	0.0150	0.0208	77.3589	1	1.425501E-18	1.400000E-04
M2	1,000	2	0.0206	0.0190	0.0208	7.3337	1	6.767192E-03	1.006500E-01
M3	1,000	2	0.0196	0.0200	0.0198	0.1020	1	7.494065E-01	1.000000E+00
M4	1,000	2	0.0823	0.0500	0.0859	174.9468	1	6.149249E-40	0.000000E+00

The SAS code above generates a marker summary table (part of the table is shown).

It provides chi-square test statistic values without continuity correction, degree of freedom, *P*-values based on the chi-square test, and *P*-values based on the exact test obtained using a permutation approach. In addition, it provides population genetic measures such as the polymorphism information content, heterozygosity, and allelic diversity. By default, the ALLELE procedure performs a chi-square goodness-of-fit test for Hardy–Weinberg proportions and reports the asymptotic *P*-values. When the PERMS=*number* option is included in the procedure, the Monte Carlo permutation test of Hardy–Weinberg proportions based on "*number*" permutations is performed, and the *P*-value thus

obtained is provided. One can also use the EXACT=*number* option instead of PERMS=*number* to perform the same permutation-based exact test. The exact test conducted here is based on the approaches proposed by Guo and Thompson (54). In this permutation procedure, the alleles are randomly permuted to form new genotypes. For each permutation, the conditional probability of genotypic counts given the allele frequency and sample size is evaluated. The *P*-value is obtained as the proportion of permutations where the conditional probabilities are less than or equal to the observed probability. It is recommended that 10,000 or more permutations be used for accuracy. Increasing the number of permutations will provide more accurate *P*-values, but the execution time will be longer (see Note 3). The SEED= option is used to define the random seed for the random number generator for permuting the alleles. It should be a nonnegative integer. If this option is omitted, the computer clock will be used (see Note 3). The exact *P*-values reported in Table 3 were based on 100,000 permutations. The ALLELE procedure can also deal with markers with multiple alleles. If the option NOFREQ is omitted, two more tables of allele frequencies and genotype frequencies will be generated. All the analyses performed in this section were conducted using SAS/Genetics 9.2 (67).

2.2. R Software

R is a free software environment for statistical computing and graphics (68). Several functions or packages have been developed for the purpose of testing Hardy–Weinberg proportions. We first focus on the functions available in the population genetic package, "genetics" (74), and also introduce several other packages developed specifically for the Hardy–Weinberg proportion exact test.

The "genetics" package has two functions: HWE.chisq and HWE.exact, for testing the departure from Hardy–Weinberg proportions based on the chi-square and exact tests, respectively. The syntaxes for the two tests are as follows:

```
HWE.chisq(x)
HWE.exact(x)
```

where x is the genotype data in object class "genotype," which can be obtained by using the genotype function also available in this package. By default, HWE.chisq provides chi-square test statistics without continuity correction and simulated *P*-values based on 10,000 iterations. If one wants the asymptotic *P*-value based on continuity-corrected chi-square statistics, then one has to use the option simulate.p.value=FALSE. The function HWE.exact provides exact Hardy–Weinberg *P*-values. The algorithm for the exact test used by this function is based on the approach proposed by

Emigh (49). This function only works for genotypes with two alleles. The following code performs the asymptotic and exact Hardy–Weinberg proportion tests within the "genetics" package.

1. Install and load the "genetics" package:

   > install.packages("genetics")

   > library("genetics")

2. Read the data for the 18 genetic markers (see Note 4), and create genotype object using function genotype. Since the genotype function only works for a single marker, we use a marker in Table 3 as an example with genotypic counts for AA, Aa, and aa of 5, 185, and 810, respectively.

   > allmarker<-read.table('markers')

   > onemarker<-allmarker[,7]

   > genodata<-genotype(onemarker, sep="/")

3. Conduct asymptotic chi-square test for Hardy–Weinberg proportions, and the asymptotic *P*-values can be obtained in three ways (see Note 5):

   > # to obtain chi-square test statistics without continuity correction and *P*-value based on simulation one needs to run default setting of the function

   > t_chisq<-HWE.chisq(genodata)

   > # to obtain asymptotic chi-square test statistics with continuity correction and associated *P*-value one needs to run the following script

   > t_chisq<-HWE.chisq(genodata, simulate.p.value=FALSE)

   > # to perform asymptotic chi-square test and associated *P*-value without continuity correction

   > t_chisq<-HWE.chisq(genodata, simulate.p.value=FALSE, correct=FALSE)

4. Conduct exact test for Hardy–Weinberg proportions:

   > t_exact<-HWE.exact(genodata)

For this marker, the asymptotic chi-square *P*-value obtained is 0.1107, based on simulation iterations. Alternatively, if one chooses not to use the simulation to compute *P*-values (simulate.p.value=FALSE), the function computes the test statistic using Yates' continuity correction and uses the asymptotic chi-square distribution to evaluate the *P*-value. The *P*-value obtained in this way is 0.1499. We can further choose not to use Yates' continuity correction (correction=FALSE), which results in a *P*-value of 0.1054. The exact *P*-value obtained from the HWE.exact function is 0.1471 in this example.

Several other R functions are available for the exact test of Hardy–Weinberg proportions, such as the function HWExact in the "GWASExactHW" package (75) and function hwexact in the "hwde" package (76). Compared to the HWE.exact function in the "genetics" package, these two functions directly deal with the genotypic counts. Both functions were adapted from code by Wigginton et al. (17). The approach proposed by Wigginton et al. (17) uses the recurrent relationships from the previous study of Guo and Thompson (54) and performs the exact test for SNPs in a computationally efficient manner for diallelic loci. Again considering genotypic counts for AA, Aa, and aa of 5, 185, and 810, respectively, the following codes can be used to obtain the exact *P*-values based on those counts.

1. Use the function HWExact in the "GWASExactHW" package:

   > genocounts<-data.frame(nAA=5,nAa=185,naa=810)

   > p_exact<-HWExact(genocounts)

2. Use the function hwexact in the "hwde" package:

   > p_exact<-hwexact(5,185,810)

Both functions provide, in this example, an exact *P*-value of 0.1471, which is the same as that obtained using the HWE.exact function in the "genetics" package. The function hwexact is simpler to perform than the function HWExact. However, by using the data.frame function, the function HWExact can deal with a large number of markers simultaneously (i.e., n_{AA}, n_{Aa}, and n_{aa} can be defined as arrays) and, therefore, is more favorable for large-scale genome-wide association studies.

It should be noted that all the R packages/functions discussed so far for the exact test of Hardy–Weinberg proportions only work for markers with two alleles. The Hardy–Weinberg proportion exact test for markers with more than two alleles can be conducted using the function hwe.hardy in the genetic analysis package "gap" (77). This function was adapted from the code by Guo (54). Interestingly, for markers with only two alleles, this function cannot be applied. All the analyses performed in this section were conducted using R version 2.10.1 (68).

2.3. PLINK Software

PLINK (69, 70) is a free software program providing a computationally efficient way of performing statistical analyses for large-scale genome-wide association studies. As a basic summary statistic, the Hardy–Weinberg proportion test can be conducted by using one command line with the option --hardy in PLINK, using the pedigree and map files (see Note 6). Both pedigree (PED) and map (MAP) files are required as the standard input files for PLINK. We first briefly introduce the formats of PED and MAP files and then describe how the Hardy–Weinberg test is conducted.

PLINK has detailed guidelines for the formats of PED and MAP files. These files are in the standard "linkage format," and all the formats and coding must conform to these guidelines. The PED file stores all the data for all the variables of all the individuals. The columns refer to variables, and the rows refer to individuals. The first six columns are mandatory: Family ID, Individual ID, Father ID, Mother ID, Sex, and Phenotype. The combination of Family ID and Individual ID needs to be unique to identify an individual. Sex is coded as 1 = male, 2 = female, and other = unknown. The Phenotype can be a quantitative trait or a case–control status. If the Phenotype is a cases–controls status, it is coded as 1 = controls, 2 = cases, and 0/−9 = missing. Starting from column seven, genotype data can be defined with two columns representing one marker. The genotypes can be coded using any numbers (1, 2, 3, and 4) or characters (A, B, C, and D) except 0, as 0 represents missing genotypes by default. The MAP files store additional information for markers, with each row describing one single marker. By default, the MAP file has exactly four columns (default settings): Chromosome (1–22, X, Y, or 0 if unplaced), rs# or SNP identifier, Genetic distance (morgans), and Base-pair position (bp units). For the detailed guidelines, the reader can refer to the online PLINK manual (http://pngu.mgh.harvard.edu/~purcell/plink/).

Once the example.ped and example.map files are created, the Hardy–Weinberg proportion tests for all markers can be performed using the following command line:

```
plink --ped example.ped --map example.map --hardy
```

By default, this command conducts the exact test of Hardy–Weinberg proportions, described and implemented by Wigginton et al. (17). To perform the asymptotic chi-square test, one can use the option --hardy2 instead:

```
plink --ped example.ped --map example.map --hardy2
```

Two files are created from this command (1) plink.log file captures all the information that should appear on the console, including information about commands used, as well as information about markers included, individuals used, cases and controls, male and female, and the missing genotypes and individuals, etc.; (2) plink.hwe provides the Hardy–Weinberg proportion test *P*-value. The first line in the plink.hwe file includes headers. The last column gives the Hardy–Weinberg *P*-value: asymptotic or exact. For each marker, there are three rows with respect to Hardy–Weinberg tests in three different samples (i.e., all data [ALL], cases only [AFF], and controls only [UNAFF]) (see Note 7).

Part of the resulting plink.hwe file based on the option --hardy is shown below.

CHR	SNP	TEST	A1	A2	GENO	O(HET)	E(HET)	P
1	rs123456	ALL	2	1	3/15/982	0.015	0.02078	9.82E-05
1	rs123456	AFF	2	1	0/0/0	nan	nan	NA
1	rs123456	UNAFF	2	1	3/15/982	0.015	0.02078	9.82E-05
1	rs234567	ALL	2	1	1/19/980	0.019	0.02078	0.1007
1	rs234567	AFF	2	1	0/0/0	nan	nan	NA
1	rs234567	UNAFF	2	1	1/19/980	0.019	0.02078	0.1007

Since we assumed all the individuals were controls, no Hardy–Weinberg test P-value was calculated for the cases (label AFF) for all markers, and the results obtained using all data (ALL) and controls (UNAFF) were exactly the same. For example, for the first SNP with genotype counts 3/15/982, the exact P-value is 9.82×10^{-5}, which is the same as the one obtained in R using different exact test functions (see Table 3). The asymptotic P-value based on the chi-square test can also be calculated. PLINK does not provide values of chi-square test statistics. Both the asymptotic and exact P-values from PLINK are listed in Table 3. The Hardy–Weinberg proportion tests available in PLINK can only deal with markers of two alleles. All the analyses performed in this section were conducted using PLINK version 1.07 (69, 70).

In Table 3, R and PLINK provide similar P-values for both asymptotic and exact tests of Hardy–Weinberg proportions. By default, SAS will provide results with only four decimals. For the purpose of comparisons, we used a scientific notation for the P-values obtained from SAS. The exact P-values from SAS are slightly different from those obtained from R or PLINK because the P-values are permutation-based. Even when the allele is common, asymptotic and exact P-values could be different. For example, when the genotypic counts are 15, 155, and 830 for AA, Aa, and aa, the minor allele frequency is 0.0925. The asymptotic and exact P-values are 0.0152 and 0.0219, respectively. In this situation, the asymptotic P-value is liberal. Furthermore, when the allele is rare, the asymptotic chi-square test is very sensitive and provides more liberal P-values than the exact test. For example, when the genotypic counts are 1, 19, and 980 for AA, Aa, and aa, the minor allele frequency is 0.0105. The asymptotic P-value is 0.0068, which is statistically significant at the 5% level and implies that this marker deviates from the Hardy–Weinberg proportions, but the exact P-value is 0.1007, which is not significant at the 5% level and suggests that this marker is in Hardy–Weinberg proportion. With 18 markers in the data, the computation time for the analysis is SAS>R>PLINK if the exact test is performed (the asymptotic

chi-square test is always quick). Using 100,000 permutations, SAS needs approximately 4–5 min to complete the analysis, R package "genetics" needs about 2–3 s, and PLINK only needs about 0.03 s. Therefore, the time it takes to perform the exact Hardy–Weinberg proportion test for SNPs in a candidate region or at the genome-wide level is less than a day. Given the liberal nature of the asymptotic-based chi-square test, we recommend that the exact test be performed routinely.

2.4. Other Software

Many other software/programs are also useful for testing the departure from Hardy–Weinberg proportion:

SNP-HWE (http://www.sph.umich.edu/csg/abecasis/Exact/) (17)

HWtest (http://www.mathworks.com/matlabcentral/fileexchange/14425-hwtest) (78)

Haploview (http://www.broadinstitute.org/mpg/haploview/) (17, 79)

TFPGA (http://www.marksgeneticsoftware.net/)

3. Notes

1. Li and Leal (80) studied the departure from Hardy–Weinberg equilibrium in a family-based study (i.e., parental and unaffected sibling genotype data). They found that the pattern of departure from Hardy–Weinberg equilibrium is different in different groups of individuals, such as the parent group, affected proband group, and unaffected sibling group.
2. The number of mixture samples L can be decided by conducting simulations. For example, given a data set, one can use different numbers of L to evaluate the empirical distribution of P-values and the maximum likelihood estimator. If the empirical distribution and the value of the maximum likelihood estimator are approaching stability when L is greater than some number, one can use this number or a greater number in the analysis.
3. We tried two different numbers of permutations, PERMS = 10,000 and 100,000. In both cases, the exact P-values show some variation if we conduct the ALLELE procedure multiple times without the SEED = option. If the multiple tests are conducted using a fixed random seed number, the exact same results can be replicated. The variations of exact P-values are larger with PERMS = 10,000 than with PERMS = 100,000. These variations might not have a significant impact on the conclusions from the exact test (Hardy–Weinberg proportion test is significant or nonsignificant), but we still recommend more permutations for accurate results, if it is feasible.

4. The input data are in a format of columns of genotype pairs. One can use different delimiters to separate two alleles, such as "A/a" and "A-a," or use no delimiter between two alleles, such as "Aa." When creating genotypes using genotype function, the delimiter needs to be specified in the function with the option sep="" (by default, sep="/"). One can also use 0, 1, or 2 to represent genotypes aa, Aa, or AA, and then use function as.genotype.allele.count to convert them to genotype pairs A/A, A/a, and a/a. If only the genotypic counts are available, one can also create the genotype data and then apply the genotype function:

   > genocounts<-c(5, 185, 810)

   > data<-c(rep("A/A",genocounts[1]), rep("a/A",genocounts[2]), rep("a/a",genocounts[3]))

   > genodata<-genotype(data)

5. When using the HWP.chisq function based on the asymptotic chi-square distribution (simulate.p.value=FALSE), a warning message might appear regarding the validity of chi-squared approximation:Warning messages:1: In chisq.test(tab, ...) : Chi-squared approximation may be incorrect. This is probably due to a small expected count in one cell. To check the expected counts under the null, one can use results$expected, where the "results" is the variable saving all the outcomes.
6. PLINK is a command-line program with no GUI interface. All the command lines need to be written at the command prompt (e.g., DOS window or Unix terminal). The basic syntax of PLINK is as follows:

   plink --ped file.ped --map file.map --option

   The options --ped and --map indicate the input pedigree and map data files. The --option specifies the analysis or methods to be applied. All the results are saved in files with different extensions according to the analyses performed.
7. The example data used the case–control status as the phenotype. For a quantitative trait, each SNP only has one row, labeled as ALL(QT). By default, only founders are considered in the Hardy–Weinberg proportion analysis. Instead, the option --nonfounders can be used to indicate that all individuals will be included to perform an approximate test.

References

1. Castle WE (1903) The laws of Galton and Mendel and some laws governing race improvement by selection. Proc Amer Acad Arts Sci **35**:233–242
2. Hardy GH (1908) Mendelian proportions in a mixed population. Science **28**:49–50
3. Weinberg W (1908) On the demonstration of heredity in man. In: Boyer SH (ed) Papers on human genetics. Prentice Hall, Englewood Cliffs, NJ
4. Crow JF (1988) Eighty years ago: the beginnings of population genetics. Genetics **119**:473–476

5. Weir BS (1996) Genetic data analysis II: methods for discrete population genetic data. Sinauer Associates, Sunderland, Mass
6. Cockerham CC (1969) Variance of gene frequencies. Evolution **23**:72–84
7. Wright S (1951) The genetical structure of populations. Ann Eugen **15**:323–354
8. Price GR (1971) Extension of the Hardy-Weinberg law to assortative mating. Ann Hum Genet **34**:455–458
9. Shockley W (1973) Deviations from Hardy-Weinberg frequencies caused by assortative mating in hybrid populations. Proc Natl Acad Sci USA **70**:732–736
10. Templeton A (2006) Population genetics and microevolutionary theory. John Wiley & Sons, Hoboken, NJ
11. Voight BF, Pritchard JK (2005) Confounding from cryptic relatedness in case–control association studies. PLoS Genet 1:e32
12. Weinberg CR, Morris RW (2003) Invited commentary: Testing for Hardy-Weinberg disequilibrium using a genome single-nucleotide polymorphism scan based on cases only. Am J Epidemiol **158**:401–403
13. Deng HW, Chen WM, Recker RR (2000) QTL fine mapping by measuring and testing for Hardy-Weinberg and linkage disequilibrium at a series of linked marker loci in extreme samples of populations. Am J Hum Genet **66**:1027–1045
14. Deng HW, Chen WM, Recker RR (2001) Population admixture: detection by Hardy-Weinberg test and its quantitative effects on linkage-disequilibrium methods for localizing genes underlying complex traits. Genetics **157**:885–897
15. Grover VK, Cole DE, Hamilton DC (2010) Attributing Hardy-Weinberg disequilibrium to population stratification and genetic association in case–control studies. Ann Hum Genet **74**:77–87
16. Ryckman K, Williams SM (2008) Calculation and use of the Hardy-Weinberg model in association studies. Curr Protoc Hum Genet Chapter 1:Unit 1.18
17. Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet **76**:887–893
18. Attia J, Thakkinstian A, McElduff P et al (2010) Detecting genotyping error using measures of degree of Hardy-Weinberg disequilibrium. Stat Appl Genet Mol Biol **9**(1) :Article 5
19. Gomes I, Collins A, Lonjou C et al (1999) Hardy-Weinberg quality control. Ann Hum Genet **63**:535–538
20. Graffelman J, Camarena JM (2008) Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. Hum Hered **65**:77–84
21. Hosking L, Lumsden S, Lewis K et al (2004) Detection of genotyping errors by Hardy-Weinberg equilibrium testing. Eur J Hum Genet **12**:395–399
22. Laurie CC, Doheny KF, Mirel DB et al (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol **34**(6):591–602
23. Li M, Li C (2008) Assessing departure from Hardy-Weinberg equilibrium in the presence of disease association. Genet Epidemiol **32**:589–599
24. Schaid DJ, Batzler AJ, Jenkins GD et al (2006) Exact tests of Hardy-Weinberg equilibrium and homogeneity of disequilibrium across strata. Am J Hum Genet **79**:1071–1080
25. Tapper W, Collins A, Gibson J et al (2005) A map of the human genome in linkage disequilibrium units. Proc Natl Acad Sci USA **102**:11835–11839
26. Wang J, Shete S (2010) Using both cases and controls for testing hardy-weinberg proportions in a genetic association study. Hum Hered **69**:212–218
27. Weale ME (2010) Quality control for genome-wide association studies. Methods Mol Biol **628**:341–372
28. Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. Am J Hum Genet **76**:967–986
29. Pompanon F, Bonin A, Bellemain E et al (2005) Genotyping errors: causes, consequences and solutions. Nat Rev Genet **6**:847–859
30. Akey JM, Zhang K, Xiong M et al (2001) The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. Am J Hum Genet **68**:1447–1456
31. Weiss ST, Silverman EK, Palmer LJ (2001) Case–control association studies in pharmacogenetics. Pharmacogenomics J **1**:157–158
32. Xu J, Turner A, Little J et al (2002) Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? Hum Genet **111**:573–574
33. Wang J, Shete S (2008) A test for genetic association that incorporates information about deviation from Hardy-Weinberg proportions in cases. Am J Hum Genet **83**:53–63
34. Cox DG, Kraft P (2006) Quantification of the power of Hardy-Weinberg equilibrium testing

to detect genotyping error. Hum Hered **61**:10–14

35. Fardo DW, Becker KD, Bertram L et al (2009) Recovering unused information in genome-wide association studies: the benefit of analyzing SNPs out of Hardy-Weinberg equilibrium. Eur J Hum Genet. doi:10.1038/ejhg.2009.85
36. Leal SM (2005) Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. Genet Epidemiol **29**:204–214
37. Teo YY, Fry AE, Clark TG et al (2007) On the usage of HWE for identifying genotyping errors. Ann Hum Genet **71**:701–703
38. Zou GY, Donner A (2006) The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case–control data: a cautionary note. Ann Hum Genet **70**:923–933
39. Salanti G, Amountza G, Ntzani EE et al (2005) Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. Eur J Hum Genet **13**:840–848
40. Feder JN, Gnirke A, Thomas W et al (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet **13**:399–408
41. Jiang R, Dong J, Wang D et al (2001) Fine-scale mapping using Hardy-Weinberg disequilibrium. Ann Hum Genet **65**:207–219
42. Nielsen DM, Ehm MG, Weir BS (1998) Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet **63**:1531–1540
43. Lee WC (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
44. Song K, Elston RC (2006) A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case–control studies. Stat Med **25**:105–126
45. Won S, Elston RC (2008) The power of independent types of genetic information to detect association in a case–control study design. Genet Epidemiol **32**:731–756
46. Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case–control association studies. Genome Res **11**:2115–2119
47. Yates F (1934) Contingency tables involving small numbers and the X^2 test. J Roy Stat Soc Suppl **1**:217–235
48. Fisher RA (1935) The logic of inductive inference. J Roy Stat Soc **98**:39–54
49. Emigh T (1954) A comparison of tests for Hardy-Weinberg equilibrium. Biometrics **36**:627–642
50. Haldane JBS (1954) An exact test for randomness of mating. J Genet **52**:631–635
51. Engels WR (2009) Exact tests for Hardy-Weinberg proportions. Genetics **183**:1431–1441
52. Levene H (1949) On a matching problem arising in genetics. Ann Math Stat **20**:91–94
53. Louis EJ, Dempster ER (1987) An exact test for Hardy-Weinberg and multiple alleles. Biometrics **43**:805–811
54. Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics **48**:361–372
55. Aoki S (2003) Network algorithm for the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrical J **45**:471–490
56. Maurer HP, Melchinger AE, Frisch M (2007) An incomplete enumeration algorithm for an exact test of Hardy-Weinberg proportions with multiple alleles. Theor Appl Genet **115**:393–398
57. Huber M, Chen Y, Dinwoodie I et al (2006) Monte Carlo algorithms for Hardy-Weinberg proportions. Biometrics 62:49–53
58. Yuan A, Bonney GE (2003) Exact test of Hardy-Weinberg equilibrium by Markov chain Monte Carlo. Math Med Biol **20**:327–340
59. Lazzeroni LC, Lange K (1997) Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. Ann Stat **25**:138–168
60. Hernandez JL, Weir BS (1989) A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics **45**:53–70
61. Maiste PJ, Weir BS (2004) Optimal testing strategies for large, sparse multinomial models. Comput Stat Data An **46**:605–620
62. Montoya-Delgado LE, Irony TZ, de BPC et al (2001) An unconditional exact test for the Hardy-Weinberg equilibrium law: sample-space ordering using the Bayes factor. Genetics **158**:875–883
63. Shoemaker J, Painter I, Weir BS (1998) A Bayesian characterization of Hardy-Weinberg disequilibrium. Genetics **149**:2079–2088
64. Wakefield J (2010) Bayesian methods for examining Hardy-Weinberg equilibrium. Biometrics **66**:257–265
65. Wellek S, Goddard KA, Ziegler A (2010) A confidence-limit-based approach to the assessment of Hardy-Weinberg equilibrium. Biom J **52**:253–270
66. Goddard KA, Ziegler A, Wellek S (2009) Adapting the logical basis of tests for Hardy-Weinberg Equilibrium to the real needs of

association studies in human and medical genetics. Genet Epidemiol **33**:569–580

67. SAS Institute Inc. (2008) SAS/Genetics^{TM} 9.2 User's Guide. SAS Institute Inc., Cary, NC
68. R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
69. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81**:559–575
70. Purcell S (2009) PLINK (v1.07).
71. Yu C, Zhang S, Zhou C et al (2009) A likelihood ratio test of population Hardy-Weinberg equilibrium for case–control studies. Genet Epidemiol **33**:275–280
72. Taylor J, Tibshirani R (2006) A tail strength measure for assessing the overall univariate significance in a dataset. Biostatistics **7**:167–181
73. Wang J, Shete S (2009) Is the tail-strength measure more powerful in tests of genetic association? response. Am J Hum Genet **84**:298–300
74. Warnes G, Gorjanc G, Leisch F et al (2008) genetics: Population Genetics.
75. Painter I (2010) GWASExactHW: Exact Hardy-Weinburg testing for Genome Wide Association Studies.
76. Maindonald JH, Johnson R (2009) hwde: Models and tests for departure from Hardy-Weinberg equilibrium and independence between loci.
77. Zhao JH (2007) gap: Genetic analysis package. J Stat Softw **23**(8):1–18
78. Cardillo G (2007) HWtest: a routine to test if a locus is in Hardy Weinberg equilibrium (exact test).
79. Barrett JC, Fry B, Maller J et al (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics **21**:263–265
80. Li B, Leal SM (2009) Deviations from hardy-weinberg equilibrium in parental and unaffected sibling genotype data. Hum Hered **67**:104–115