Exact Tests of Hardy-Weinberg Equilibrium and Homogeneity of Disequilibrium across Strata

Daniel J. Schaid, Anthony J. Batzler, Gregory D. Jenkins, and Michelle A. T. Hildebrandt

Detecting departures from Hardy-Weinberg equilibrium (HWE) of marker-genotype frequencies is a crucial first step in almost all human genetic analyses. When a sample is stratified by multiple ethnic groups, it is important to allow the marker-allele frequencies to differ over the strata. In this situation, it is common to test for HWE by using an exact test within each stratum and then using the minimum *P* value as a global test. This approach does not account for multiple testing, and, because it does not combine information over strata, it does not have optimal power. Several approximate methods to combine information over strata have been proposed, but most of them sum over strata a measure of departure from HWE; if the departures are in different directions, then summing can diminish the overall evidence of departure from HWE. An exact stratified test is more appealing because it uses the probability of genotype configurations across the strata as evidence for global departures from HWE. We developed an exact stratified test for HWE for diallelic markers, such as single-nucleotide polymorphisms (SNPs), and an exact test for homogeneity of Hardy-Weinberg disequilibrium. By applying our methods to data from Perlegen and HapMap—a combined total of more than five million SNP genotypes, with three to four strata and strata sizes ranging from 23 to 60 subjects—we illustrate that the exact stratified test provides more-robust and more-powerful results than those obtained by either the minimum of exact test *P* values over strata or approximate stratified tests that sum measures of departure from HWE. Hence, our new methods should be useful for samples composed of multiple ethnic groups.

Evaluating Hardy-Weinberg equilibrium (HWE) among marker-genotype proportions is basic to all studies of population genetic data. Some causes of departure from HWE are nonrandom mating, recent migrations, mutations, selection, undetected "silent" or deleted alleles in heterozygotes, and mixture of subpopulations that do not completely interbreed. Because HWE is expected to occur for most large, randomly mating populations, departures from HWE are often interpreted as genotype errors. Genotypes that significantly depart from HWE are often removed from analyses, although one should be cautious when stretches of markers in linkage disequilibrium depart from HWE.[1] Current large-scale efforts to discover SNPs and to characterize their frequencies and correlation structure across the genome, as well as across different populations, use relatively small numbers of subjects from different ethnic groups. Hinds et al.[2] characterized >1.6 million SNPs among samples from three ethnic groups, using 23–24 subjects per group. The HapMap project has genotyped >3.7 million SNPs in four ethnic groups, using 45–60 independent subjects per group.[3] When testing for HWE in these studies, researchers computed tests within ethnic group strata and used the smallest *P* value over all strata, to measure the quality of each SNP. A problem with this approach is that the sample size within each of the strata may not be sufficient to detect meaningful departures from HWE, in contrast to a test that combines the evidence for departure from HWE across all strata. Several methods have been proposed to combine information across strata, allowing for differences in allele frequencies, but none are exact tests. These proposed methods can lead to inflated type I error rates or loss of power. For this reason, we developed an efficient algorithm to compute exact tests for HWE that combine information across strata for diallelic markers, such as SNPs.

To appreciate the limitations of past work on methods of testing HWE across strata, we briefly review some of the key aspects, because some points provide a deeper understanding of the issues and some developments are useful for our exact methods. For notation, we use *A* and *B* to represent the rare and common alleles, respectively, of a locus, with respective allele frequencies p and $q = 1 - p$ ($p \leqslant q$). As explained by Weir,[4] the frequencies of the three genotypes can be expressed in terms of the allele frequencies and a measure of departure from HWE (coefficient of disequilibrium D):

$$P_{AA} = p^2 + D \ ,$$

$$P_{AB} = 2pq - 2D \ ,$$

and

$$P_{BB} = q^2 + D \ .$$

Departure from HWE is then provided by $D = P_{AA} - p^2$ or,

From the Divisions of Biostatistics (D.J.S.; A.J.B.; G.D.J.) and Clinical Pharmacology (M.A.T.H.), Mayo Clinic College of Medicine, Rochester, MN

Received July 28, 2006; accepted for publication October 17, 2006; electronically published November 3, 2006.

Address for correspondence and reprints: Dr. Daniel J. Schaid, Harwick 775, Section of Biostatistics, Mayo Clinic, 200 First Street SW, Rochester, MN 55905. E-mail: schaid@mayo.edu

Am. J. Hum. Genet. 2006;79:1071–1080. 0002-9297/2006/7906-0009$15.00

equivalently, by $D = (4P_{AA}P_{BB} - P_{AB}^2)/4$. This latter expression is more commonly used in the literature.

Haldane[5] was the first to develop a stratified test for HWE. He did this by recognizing that D is expected to be zero when HWE is true. When HWE holds, the allele counts are sufficient statistics, and so the probability of the genotype counts, conditional on the allele counts, allows one to compute the mean and variance of the parameter of interest. Let N_{AA}, N_{AB}, and N_{BB} denote the counts of the genotypes. To estimate D from a sample, it may be tempting to plug in the sample estimates, $\hat{P}_{AA} = N_{AA}/N$, $\hat{P}_{AB} = N_{AB}/N$, and $\hat{P}_{BB} = N_{BB}/N$. However, because genotype counts are negatively correlated, this would lead to a biased estimate for small samples; the bias diminishes as the total sample size N increases. As emphasized by Smith,[6] this may not be an important bias for large samples, but, when adding contributions across strata, each of small sample size, the bias can be amplified. Hence, Haldane used an unbiased estimate of D,

$$\hat{D} = \frac{4N_{AA}N_{BB} - N_{AB}(N_{AB} - 1)}{N(N-1)} \ ,$$

so that testing $D = 0$ can be based on the sample estimate $h = 4N_{AA}N_{BB} - N_{AB}(N_{AB} - 1)$. Haldane[5] derived an unbiased variance of h when HWE is true. In contrast, Smith[6] derived the variance of h for when there are departures from HWE, illustrating how the variance of h depends on the population parameters p, q, and D. However, he did not derive an unbiased variance estimate; one cannot simply plug sample estimates into the variance formulas.

To combine the h values over strata, Haldane first standardized each stratum's h by its SE, $h_k/\sqrt{\mathrm{Var}(h_k)}$, and then summed these standardized terms over the K strata to compute the combined statistic $T_{\mathrm{Haldane}} = \frac{1}{\sqrt{K}} \sum h_k/\sqrt{\mathrm{Var}(h_k)}$, which has an approximate standard normal distribution when HWE is true. A problem with this approach is first standardizing and then summing. A more powerful approach would be to first sum and then standardize, much like the way the Mantel-Haenszel test is constructed for testing a common odds ratio over strata[7] or the way NPL statistics can be optimally combined across pedigrees.[8] Hence, we propose the statistic $T = (\sum h_k)/\sqrt{\sum \mathrm{Var}(h_k)}$, which also has an approximate standard normal distribution. Positive values of T imply an excess of homozygotes, and negative values an excess of heterozygotes, making it simple to interpret significant departures from HWE.

In contrast to Haldane's method, Smith[6] computed a weighted sum of the h_k values, using weights proportional to the inverse of the variance. However, his derivations were a bit odd, because he assumed that the allele frequencies are constant over strata, which is counter to what we wish to assume.

The methods by Haldane and Smith are appropriate if the h_k values are all in the same direction (positive or negative) over strata, but they can cancel each other if this is not the case, which will weaken power. For this reason, others have assumed that the ratio $\theta = P_{AB}^2/4P_{AA}P_{BB}$ is constant over strata, much like the assumption of a constant odds ratio across stratified 2×2 tables in epidemiological studies.[9–11] Nonetheless, the resulting test for HWE across strata, derived by Olson,[9,10] is also based on a weighted sum of h_k values. Nam[11] derived score statistics based on likelihoods that depend on the θ parameter and showed that his combined tests for HWE had properties similar to the test proposed by Olson. Hence, the variety of proposed tests for HWE that combine information across strata are all based on the stratum-specific h_k values, with merely slightly different ways of weighting the contribution from each stratum. Simulations (not shown) suggest that the type I error rate and power of the different methods are similar, and our simple T statistic would provide a powerful test for HWE when the sample sizes of the strata are not too small and departures from HWE are all in the same direction.

Because summing h_k values over strata can cancel each other when they differ in sign, Troendle and Yu[12] proposed a statistic that is analogous to summing $h_k^2/\mathrm{Var}(h_k)$ over strata. The resulting statistic has a χ^2 distribution with K df. Although this method can have greater power when the h_k values differ in sign, it is likely to have weak power in general, because of the many df. An alternative approach is to test whether the θ_k values significantly differ over the strata, because significant heterogeneity implies departure from HWE. To compute this type of statistic under the null hypothesis of homogeneity (yet allowing departure from HWE), one needs to estimate a common θ parameter. Using estimating equations, Olson and Foley[10] derived a consistent estimator for θ, whereas Nam[11] used an iterative maximum-likelihood method. Both of these approaches, however, can run into undefined parameter estimates; Olson's θ is undefined when there are no AA homozygotes across all strata (or no BB homozygotes); similar problems occur for the maximum-likelihood estimator. In these cases, the test for homogeneity breaks down.

Because of the above complications with large-sample statistical tests for HWE across strata or for homogeneity of departures from HWE across strata, exact methods are appealing. Instead of summing a measure of departure from HWE, an exact test evaluates the combined evidence over strata by considering the probability of genotype configurations when the null hypothesis is true; extreme departures in different directions are rare under HWE, giving a small P value, yet a sum of measures of departure from HWE could, in fact, completely miss this situation. Exact tests also avoid numerical problems (e.g., division by zero), and they provide appropriate control of the type I error rates. To date, only Olson and Foley[10] considered exact methods. However, because their methods allowed for an arbitrary number of alleles, exact computations were not feasible. Rather, they needed to rely on Markov Chain–Monte Carlo methods. Because of the broad use of SNPs, we present efficient computational methods to compute exact tests both for HWE over strata and for homogeneity of departures of HWE over strata. We demonstrate our methods by applying them to the SNP genotype data from Perlegen[2] and HapMap.[3] The results illustrate the advantages of our exact stratified test for HWE over the minimum exact-test P value or other approximate methods. Furthermore, simulations confirm our empirical findings.

Methods

Exact Stratified Test for HWE

To derive an exact stratified test for HWE, we use well-known methods for computing the probability of a sample of genotypes when HWE is true. In this case, Fisher[4] showed that the allele counts N_A and N_B are sufficient statistics and that the probability of genotype counts, conditional on allele counts, can be expressed as

$$P(N_{AA},N_{AB},N_{BB}\,|\,N_A,N_B) = \frac{N!N_A!N_B!2^{N_{AB}}}{N_{AA}!N_{AB}!N_{BB}!(2N)!} \ . \qquad (1)$$

Because N, N_A, and N_B are all fixed, the only random genotype is the number of heterozygotes, so expression (1) can be written as

$$P(N_{AB}=x) = \frac{N!N_A!N_B!}{(2N)!} \times \frac{2^x}{[(N_A-x)/2]!x![N-(N_A+x)/2]!} \ . \qquad (2)$$

An exact P value is computed by summing values from equation (2) over all values of x that generate probabilities equal to or smaller than do the observed number of genotypes, N_{AB}. As emphasized by Wiggington et al.,[13] when N_A is odd, the possible values of x are 1, 3, …, N_A, and, when N_A is even, the possible values of x are 0, 2, …, N_A. Furthermore, equation (2) can be computed efficiently by recursion:

$$P(N_{AB}=x+2) = P(N_{AB}=x) \frac{4[(N_A-x)/2][N-(N_A+x)/2]}{(x+2)(x+1)} \ .$$

Now, to extend these ideas to strata, let $\tilde{N}_{AB}$ be the vector of observed counts of AB heterozygotes for the different strata, and let $\tilde{x}$ be a vector containing a configuration of possible values of heterozygotes for the different strata. The probability of $\tilde{x}$ under HWE is the product of expression (2) over the K strata, $P(\tilde{x}) = \prod P(x_k)$. This allows us to compute an exact stratified P value by

$$P \text{ value} = \sum_{\tilde{x}\in S} P(\tilde{x}) \ ,$$

where S is the set of $\tilde{x}$ configurations that have probabilities equal to or less than that of the observed configuration: $S = \{\tilde{x}:P(\tilde{x}) \leqslant P(\tilde{N}_{AB})\}$. If m_k is the number of possible values of x in stratum k, then the number of possible $\tilde{x}$ configurations is $\prod m_k$, which can be a very large number.

A naive approach to compute the exact P value is to evaluate all possible configurations, which is inefficient. Rather, we first compute $P(x)$ for all possible values of x within each stratum. This avoids having to recompute $P(x)$ many times. Using recursion makes this fast, and using log-probabilities avoids numerical imprecision. Because it is of critical importance, we order the log-probabilities such that we can stop the summation for the P values as soon as possible. To do this, we sort the log-probabilities into increasing order within each stratum, using quick sort. Then, we begin to evaluate different possible $\tilde{x}$ configurations by summing, across strata, the log-probabilities for values of x in the $\tilde{x}$ vector. If this sum is less than or equal to the log-probability of the observed data, then we exponentiate it and add it to the running sum for the P value. The prior sorting of the log-probabilities within each stratum allows us to skip over computations that would generate $P(\tilde{x}) > P(\tilde{N}_{AB})$ and, hence, would not contribute to the sum for the P value. This is explained in greater detail by use of an example in appendix A. An advantage of our approach is that, when small P values are used as a quality-control (QC) filter (e.g., $P < .001$ as used elsewhere[2,3]), we can stop computations early when the computed P value exceeds a specified threshold.

Exact Test of Homogeneity of Disequilibrium

Olson and Foley[10] derived a test of homogeneity of disequilibrium across strata, on the basis of the assumption that $\theta = P^2_{AB}/4P_{AA}P_{BB}$ is constant over strata. They showed that the sufficient statistics for this test are the allele counts within strata, as well as the total (across strata) genotype counts. As in the exact test for HWE, when conditioning on the sufficient statistics, we need to focus on only $\tilde{x}$ configurations of counts of heterozygotes across the strata. However, by additionally conditioning on the total genotype counts, we require the sum of the elements of the $\tilde{x}$ vector to equal the observed total number of heterozygotes. This leads to fewer possible $\tilde{x}$ vectors than those possible when testing HWE over strata.

Under the assumption of constant θ, the probability of an $\tilde{x}$ configuration is

$$P(\tilde{x}) = \frac{Q(\tilde{x})}{\sum_{\tilde{x}^*\in G} Q(\tilde{x}^*)} \ , \qquad (3)$$

where G is the set of possible $\tilde{x}$ configurations (each summing to the total number of observed heterozygotes) and

$$Q(\tilde{x}) = \prod_k \frac{N_k!}{N_{AA,k}!x_k!N_{BB,k}!} = \prod_k \frac{N_k!}{[(N_{A,k}-x_k)/2]!x_k![(N_{B,k}-x_k)/2]!} \ .$$

An exact P value is the sum of the configuration probabilities that are equal to or less than the observed configuration (denoted $\tilde{x}_{\text{obs}}$),

$$P \text{ value} = \sum_{\tilde{x}\in S} P(\tilde{x}) \ ,$$

where $S = \{\tilde{x}:P(\tilde{x}) \leqslant P(\tilde{x}_{\text{obs}}),\tilde{x} \in G\}$.

To efficiently compute the exact P value, we first enumerate possible x_k values for each stratum (for now, ignoring the constraint that the x_k values must sum to the total number of ob-

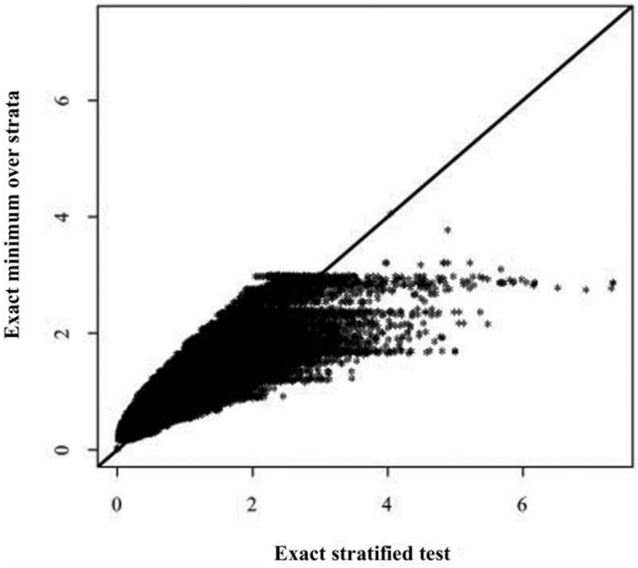

Figure 1. Perlegen data. Exact stratified test versus the minimum within-strata exact test; plot of $-\log_{10}(P \text{ value})$.

served heterozygotes) and use recursion to determine the contribution of the x_k values to $Q(\tilde{x})$. To see this, let $q(x_k)$ denote the ratio of factorials for stratum k (i.e., $Q(\tilde{x}) = \prod q(x_k)$). It is easy to verify that

$$q(x_k + 2) = \frac{q(x_k)[(N_{A,k} - x_k)/2][(N_{B,k} - x_k)/2]}{(x_k + 2)(x_k + 1)} \ .$$

By precomputing these $q(x_k)$ values, we merely need to look up their values as we determine $Q(\tilde{x})$ for different $\tilde{x}$ configurations. Some differences between our method to compute the P value for homogeneity versus our method to compute the exact P value for HWE combined over strata are that (1) we need to consider all possible configurations, because the sum of $Q(\tilde{x})$ over all configurations is used in the denominator of $P(\tilde{x})$ (expression [3]), and (2) the constraint that the elements of $\tilde{x}$ must sum to the total number of observed heterozygotes reduces the number of possible $\tilde{x}$ vectors. This is used to our advantage. Further details of this algorithm are explained by an example in appendix A.

Applications

We applied our exact methods to SNPs in the Perlegen[2] and HapMap[3] data sets. For the Perlegen data, we used 1,585,674 SNPs from all chromosomes. For the X chromosome, we used males and females for the pseudoautosomal regions and only females for other regions on X. The Perlegen data have a total of 71 subjects from three ethnic groups: 23 African Americans, 24 European Americans, and 24 Han Chinese. Furthermore, the Perlegen data were "cleaned" by a number of criteria, including exact tests for HWE within each of the ethnic groups. SNPs were given a poor quality score if the smallest P value across the three ethnic groups was <.001. Hence, the Perlegen data are useful to evaluate whether combining information across strata detects significant departure from HWE that was missed by using the minimum P value.

To provide a more complete comparison of using the minimum P value versus the exact stratified P value, we also applied our exact methods to SNPs in the HapMap data, using only the autosomes. These data have a total of 210 independent subjects from four ethnic groups: 60 Yoruba from Ibadan, Nigeria; 60 U.S. residents with northern and western European ancestry (CEPH samples); 45 Han Chinese; and 45 Japanese. Note that the offspring of "trios" were not used in our tests of HWE.

For the HapMap data, instead of using the cleaned data, we used the "redundant-unfiltered" genotype data. This allows us to evaluate our methods on data that did not have genotypes removed because of prior tests of HWE within strata. For this data, various QC flags were used to indicate reasons why SNPs failed the QC criteria, including an exact test for HWE within each of the strata; $P <$ .0001 in any of the strata was flagged as a failure. For our analyses, we did not eliminate SNPs that failed for this reason. Rather, we eliminated SNPs that failed the QC criteria for any reason not indicated by a HWE failure. For the duplicate samples, we coded a genotype as "missing" if the duplicates did not agree and then removed the duplicates for analyses. This resulted in the examination of 3,798,286 SNPs.

Results

Perlegen Application

To compare the results from different statistical tests, we compare the values of $-\log_{10}(P \text{ value})$, denoted lg$P$, so that small P values give large values of lg$P$. The contrast of using our exact stratified P value versus the minimum ex-

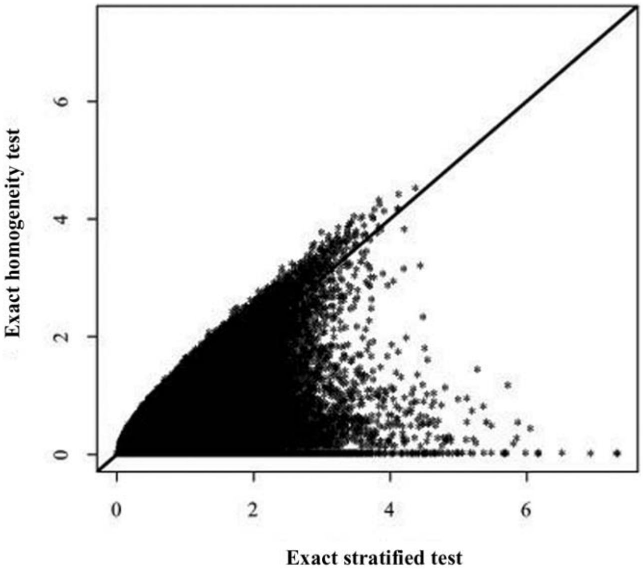

Figure 2. Perlegen data. Exact stratified test versus exact test of homogeneity; plot of $-\log_{10}(P \text{ value})$.

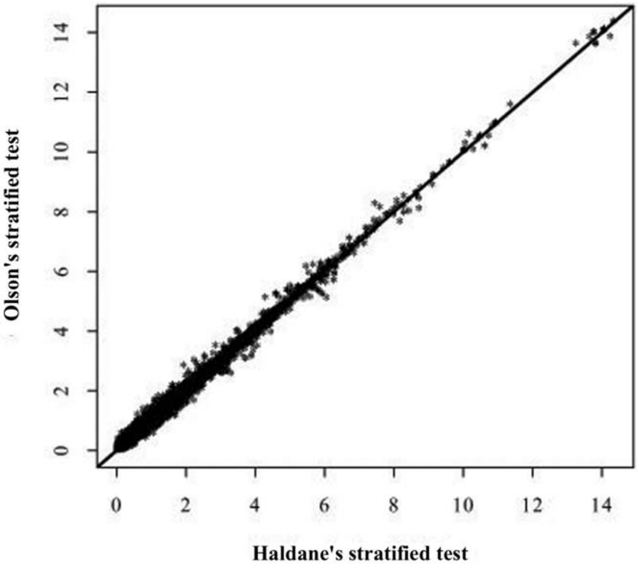

Figure 3. Perlegen data. Haldane's versus Olson's stratified tests; plot of $-\log_{10}(P \text{ value})$.

act P value for the three strata of the Perlegen data is illustrated in figure 1. This figure illustrates that a large number of SNPs have an exact stratified lg$P>4$ yet a minimum P value over strata with lg$P<4$ (a total of 300 SNPs). The majority of these SNPs (286) had an excess number of homozygotes compared with that expected if HWE were true. Note that, because the available SNP data from Perlegen were already cleaned of SNPs with departures from HWE based on the minimum P value, the minimum P values in figure 1 are truncated.

Because the test for homogeneity of departures from HWE could also be used as a test for HWE, we plot in figure 2 the exact P values for the exact stratified test for HWE and the test for homogeneity. This illustrates that the lg$P$ values for the exact stratified test tend to be larger than the lg$P$ values for the homogeneity test, which suggests that using the homogeneity test as a way to test for HWE is not likely to be as powerful as using the stratified test. Olson and Foley[10] also noted the weak power of the homogeneity test.

We also applied our adaptation of Haldane's stratified test for HWE, as well as Olson's[9] stratified test, and contrasted these with the exact stratified test. The results in figure 3 illustrate that Haldane's and Olson's stratified tests give nearly identical P values, suggesting that the different ways to weight the h values over strata has little impact in real applications. In figure 4, we compare the exact stratified test with our version of Haldane's stratified test. This figure illustrates that the exact test and Haldane's test can have large discrepancies. At one extreme, where the exact test gives large lg$P$ values and Haldane's test gives lg$P$ values near 0, the cause was typically that the h values differed in sign over strata, causing them to cancel each other, to result in Haldane's summary statistic to be near zero. In contrast, the exact test was able to detect these types of departure from HWE. At the other extreme, where the exact stratified test gave lg$P$ values near 0 and Haldane's test gave lg$P$ values near 1, the summary Haldane statistics were typically negative, implying too few homozygotes. In these cases, only one type of homozygote was observed over all strata, yet the observed and HWE expected genotype counts were quite close. Furthermore, Haldane's test tended to have more extreme values of lg$P$ than did the exact test; 109 instances with lg$P>12$. In all of these cases, the strata had no heterozygotes and either one or two rare homozygotes. These results empirically emphasize the inadequacy of the normal distribution for Haldane's stratified statistic when there are sparse data, leading to P values that are likely much too small.

HapMap Application

The HapMap data provide an unbiased comparison of using the minimum exact-test P value over strata versus using the exact stratified test because the "uncleaned" data were available. Figure 5 illustrates that the lg$P$ value for the exact stratified test tends to be larger, sometimes much larger, than that based on the minimum of exact test P values, implying a significant gain in power by using the exact stratified test. One should be cautious, however, when interpreting figure 5, because almost 3.8 million points are plotted, and so the density of points that represent acceptable SNPs (i.e., $P>.0001$) cannot be easily viewed. When this threshold was used, there were 2,095 SNPs significant by the minimum exact-test P value and not by the exact stratified test and 15,147 significant by the exact stratified test and not by the minimum P value

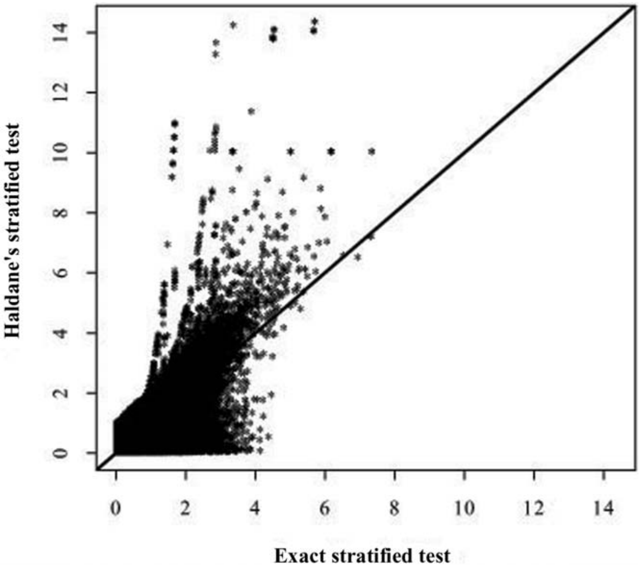

Figure 4. Perlegen data. Exact versus Haldane's stratified tests: plot of $-\log_{10}(P \text{ value})$.

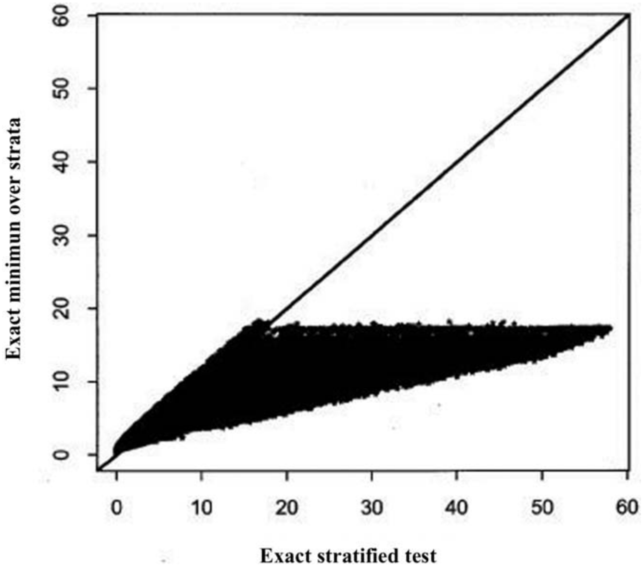

Figure 5. HapMap data. Exact stratified test versus the minimum within-strata exact test; plot of $-\log_{10}(P \text{ value})$.

over strata. This illustrates the greater sensitivity of the exact stratified test. Note that these results are for the redundant-unfiltered genotype data. To evaluate the quality of the cleaned data that most investigators use, we also applied our methods to the 3,751,020 cleaned autosomal SNPs. For these SNPs, there were 2,006 that are significant by our exact stratified test, suggesting that these SNPs have suspicious quality. Note that we did not correct for multiple testing when using the minimum P value over the four strata; had we done so, the minimum P value would increase, accentuating the greater power of the exact stratified test.

Like the Perlegen results displayed in figure 2, we found for the HapMap data that the lg$P$ values for the exact stratified test tended to be larger than the lg$P$ values for the homogeneity test, emphasizing that using the homogeneity test as a way to test for HWE is not likely to be useful (results not shown). Also, for the HapMap data, Haldane's and Olson's stratified tests gave similar results (results not shown).

Simulations

To demonstrate the need to use exact methods when there are sparse data (because of small strata sizes and rare alleles), we performed a limited set of simulations. For these, we evaluated the type I error rates of the exact stratified test, our version of Haldane's test, Olson's test, and the omnibus χ^2 statistic proposed by Troendle and Yu.[12] For all simulations, we used 25 subjects per stratum and either 3 or 5 strata. A total of 10,000 replicates were used for each simulation. The type I error rates presented in table 1 illustrate that the exact test is slightly conservative when the rare-allele frequency is $P = .05$ but gives the correct type I error rate when $P = .20$. In contrast, the other statistics have inflated type I error rates for $P = .05$, with the Troendle and Yu statistic[12] having grossly inflated type I error rates (likely from multiple df and sparse counts). Nonetheless, the asymptotic statistics gave approximately correct type I error rates for $P = .20$.

Simulations for power were conducted for only $P = .20$ because all tests have the correct type I error rate for this situation, again restricted to 25 subjects per stratum. Results for power are presented in table 2, for when the departure from HWE is in the same direction and magnitude across all strata. In this case, the Haldane and Olson statistics had the greatest power, as expected, because they were derived under the assumption of constant departure from HWE across strata. However, the decreases in power for the other tests were generally small. Furthermore, the power was approximately the same for the exact test and the χ^2 statistic of Troendle and Yu.[12] Results for power when the departure from HWE differed over strata are presented in table 3. In this case, the exact test and Troendle and Yu's statistic had similar power that was greater than that for Haldane's and Olson's tests.

Timing of Software

The time to compute the exact stratified test depends on the number of strata, K, and the number of rare alleles, $N_{A,k}$, within each stratum; the larger the values of K and $N_{A,k}$, the more time the tests require for computation. To evaluate the practical time limits for computing the exact stratified test, we varied the number of strata from 2 to 5 and the sample size per stratum from $N_k = 20$ to 100 (constant over strata). For all situations, we evaluated the

Table 1. Simulation Type I Error Rates

No. of Strata, Rare-Allele Frequency, and Statistic[a]	Nominal Type I Error Rate		
	$P = .05$	$P = .01$	$P = .001$
3:			
.05:			
Exact	.0243	.0044	.0004
Haldane/Olson	.0679	.0374	.0197
Troendle and Yu	.0681	.0602	.0277
.20:			
Exact	.0475	.0119	.0013
Haldane/Olson	.0428	.0088	.0014
Troendle and Yu	.0548	.0177	.0041
5:			
.05:			
Exact	.0317	.0071	.0009
Haldane/Olson	.0622	.0276	.0140
Troendle and Yu	.1046	.0986	.0362
.20:			
Exact	.0532	.0110	.0010
Haldane/Olson	.0474	.0089	.0010
Troendle and Yu	.0593	.0193	.0038

[a] Each stratum had 25 subjects. Haldane and Olson statistics gave identical results.

Table 2. Simulation Power When Departure from HWE is the Same for All Strata

No. of Strata, Fraction of Maximum HWD, and Statistic[a]	Nominal Type I Error Rate		
	$P = .05$	$P = .01$	$P = .001$
3:			
.2:			
Exact	.3635	.1669	.0462
Haldane/Olson	.4088	.2295	.0884
Troendle and Yu	.3563	.1866	.0687
.3:			
Exact	.6507	.4090	.1719
Haldane/Olson	.7005	.5084	.2797
Troendle and Yu	.6381	.4328	.2187
.5:			
Exact	.9572	.8778	.6885
Haldane/Olson	.9737	.9287	.8154
Troendle and Yu	.9550	.8891	.7392
5:			
.2:			
Exact	.4928	.2681	.0921
Haldane/Olson	.5763	.3623	.1682
Troendle and Yu	.4658	.2775	.1194
.3:			
Exact	.8096	.6085	.3377
Haldane/Olson	.8783	.7331	.5149
Troendle and Yu	.7836	.6066	.3758
.5:			
Exact	.9961	.9805	.9232
Haldane/Olson	.9984	.9932	.9720
Troendle and Yu	.9943	.9796	.9353

[a] HWD = departure from HWE, in terms of fraction (f) of maximum departure ($D_{\max} = p(1-p)$, where p is rare-allele frequency), so that the genotype frequencies are $P_{AA} = p^2 + D$, $P_{AB} = 2p(1-p) - 2D$, and $P_{AB} = (1-p)^2 + D$, where $D = fD_{\max}$. Rare-allele frequency was 0.20, with 25 subjects per stratum, and Haldane and Olson statistics gave identical results.

worst-case scenario by setting $N_{A,k}$ to its largest possible value, $N_{A,k} = N_k - 1$. All computations were performed on a Sun workstation (SUNW [Ultra-80]) with 4 GB RAM (random-access memory) and a 450-MHz processor. Timing results are given in seconds in table 4. For up to five strata and for sample sizes <50 per stratum, our software will compute within a few seconds (from ~0.001 to 10 s). In contrast to these worst-case scenarios, the average time per genotype for the HapMap data was 0.025 s. Although we illustrate computation times for up to 100 subjects per stratum, the exact test is likely not necessary for this situation, and the asymptotic tests should suffice, as long as the minor-allele frequencies are not too small.

Discussion

Application of our exact stratified test for HWE to the Perlegen and HapMap data sets provides an empirical comparison of our new methods with the common approach that uses the minimum exact-test P value. Both data sets emphasize the greater power of the exact stratified test, which makes intuitive sense because it simultaneously evaluates HWE over all strata rather than independently testing each stratum. The exact stratified test also accounts for testing multiple strata; a Bonferroni correction would be needed to control the type I error rate when using the minimum P value over the strata.

Although a number of approximate stratified tests of HWE have been proposed, our applications illustrate that our version of Haldane's test gives results nearly identical to those of Olson's stratified test for HWE, suggesting that the different ways of weighting the contribution from each stratum do not have major influences on the tests. By comparing results from Haldane's stratified test with those from the exact stratified test, we found that using the standard normal distribution to approximate the distribution of Haldane's test can give exceptionally small P values when there are sparse genotype counts, which suggests that the normal distribution is not adequate and that the exact stratified test is more reliable. Finally, both the

Table 3. Simulation Power When Departure from HWE Differs over Strata

No. of Strata, Fraction of Maximum HWD, and Statistic[a]	Nominal Type I Error Rate		
	$P = .05$	$P = .01$	$P = .001$
3:			
.2:			
Exact	.2443	.0924	.0205
Haldane/Olson	.1930	.0785	.0206
Troendle and Yu	.2492	.1120	.0355
.3:			
Exact	.4470	.2250	.0694
Haldane/Olson	.3392	.1712	.0554
Troendle and Yu	.4517	.2589	.1060
.5:			
Exact	.8514	.6640	.3902
Haldane/Olson	.6824	.4778	.2490
Troendle and Yu	.8589	.7062	.4755
5:			
.2:			
Exact	.2939	.1232	.0305
Haldane/Olson	.2180	.0924	.0244
Troendle and Yu	.2905	.1425	.0492
.3:			
Exact	.5433	.3068	.1100
Haldane/Olson	.3899	.2052	.0725
Troendle and Yu	.5422	.3338	.1515
.5:			
Exact	.9339	.8131	.5712
Haldane/Olson	.7569	.5644	.3151
Troendle and Yu	.9325	.8350	.6426

[a] HWD = departure from HWE, in terms of fraction (f) of maximum departure ($D_{\max} = p(1-p)$, where p is rare-allele frequency), so that the genotype frequencies are $P_{AA} = p^2 + D$, $P_{AB} = 2p(1-p) - 2D$, and $P_{AB} = (1-p)^2 + D$, where $D = \text{sign } fD_{\max}$. For three strata, sign is positive for strata 1 and 3 and negative for stratum 2. For five strata, sign is positive for strata 3–5 and negative for strata 1–2. Rare-allele frequency was 0.20, with 25 subjects per stratum, and Haldane and Olson statistics gave identical results.

Table 4. Timing of hweStrata

Sample Size per Stratum and No. of Strata	Time (s)
20:	
2	.00
3	.01
4	.01
5	.11
50:	
2	.00
3	.02
4	.44
5	9.78
100:	
2	.01
3	.11
4	6.15
5	304

Perlegen and HapMap data illustrated that the exact stratified test for HWE is much more powerful to detect departures from HWE than the exact test for homogeneity, echoing the simulation results of Olson and Foley.[10]

Our simulation results confirmed that the exact stratified test provides the correct type I error rate, whereas the tests proposed by Haldane, Olson, and Troendle and Yu can have inflated type I error rates in the presence of sparse data (i.e., small strata and rare alleles). Furthermore, our simulations confirmed that the exact test provides the greatest power when the departure from HWE is in different directions across strata. Finally, the simulations suggest that when the strata sizes are not small and the frequency of the rare allele is at least 5%, the omnibus test of Troendle and Yu would be a good substitute for the exact stratified test.

Although our work was motivated by the relatively small ethnic groups within the Perlegen and HapMap data sets and by the potential for using these data sets for planning large-scale genome association studies in large "homogeneous" ethnic groups, our exact tests should prove useful for many genetic studies. Some examples are follow-up studies in multiple ethnic groups, each of which may not be large (note that association analyses would need to account for the different ethnic groups, such as a Mantel-Haenszel stratified analysis), population genetic studies in multiple ethnic groups from a geographic region, or studies in which apparently homogeneous ethnic groups can be clustered into smaller ethnic subsets on the basis of many measured markers.[14]

In conclusion, our simulations and the application of exact and approximate stratified tests for HWE to more than five million SNPs, with strata sizes ranging from 23 to 60 subjects, provide convincing results that the exact stratified test provides the most-robust and most-powerful results. Furthermore, efficient computational algorithms for SNP genotype data, which we developed in the C programming language, allow the exact stratified test to be computed within reasonable computing time for sample sizes on the order of the HapMap data (e.g., strata sizes ranging from 45 to 60 subjects over four strata). The C source code for our software, called "*hweStrata,*" is available from our Web site.

Acknowledgments

This work was supported by U.S. Public Health Service, National Institutes of Health, contract grants GM065450 and GM61388 (The Pharmacogenetics Research Network). The constructive criticisms from two anonymous reviewers helped to improve the evaluation and presentation of the proposed exact stratified test.

Appendix A

Illustration of Algorithm for Computing Exact Stratified Test for HWE

To illustrate how sorting the log-probabilities within each stratum leads to an efficient way to compute an exact *P* value, we present in table A1 a hypothetical example of genotypes for three strata. For this example, the total log-probability of the observed genotypes is −24.3649. Although there are 288 possible $\tilde{x}$ configurations for the possible number of heterozygotes over the strata, configurations that have log-probabilities larger than the observed value of −24.3649 do not contribute to the *P* value, so they can be skipped. The log-probabilities, sorted within each stratum, are illustrated in table A2. Note that summing the first log-probability across strata gives the log-probability of the most rare $\tilde{x}$ configuration. The indices for these log-probabilities are 1, 1, 1, with a resulting sum of −30.0905 (see table A3). Because this sum is smaller than the summed log-probabilities for the observed data, this configuration contributes to the *P* value. Increasing the index for the third stratum, up to its maximum value of 4, gives summed log-probabilities that are less than those for the observed data, so all these configurations contribute to the *P* value. When the index for the second stratum is increased to 2, we find that the indices 1, 2, 2, lead to a summed log-probability of −22.3117 (configuration 6 in table A3), which is larger than the observed value. Increasing the index for stratum 3 to 3 or 4 would only lead to larger summed log-probabilities, so these values can be "jumped" over. The next array of indices is 1, 3, 1, which contributes to the *P* value, but the following array, 1, 3, 2, is another "jump" array. The arrays that indicate the jumps are given in table A3. For this example, we need to evaluate only 13 configurations of the total 288 configurations.

Illustration of Algorithm for Computing Exact Test for Homogeneity

To illustrate how we enumerate $\tilde{x}$ configurations that meet the constraints (1) that each x_k is restricted to a range of values determined by $N_{A,k}$ (0–$N_{A,k}$ if $N_{A,k}$ is even and 1–$N_{A,k}$ otherwise) and (2) that the elements in $\tilde{x}$ sum to the total number of observed heterozygotes, we present a simple example in table A4. This table is a slightly modified version of the data in table A1. For this example, the observed total number of heterozygotes is seven, so the maximum number of heterozygotes $N_{A,k}$ cannot be achieved for any of the strata.

To enumerate possible $\tilde{x}$ configurations, we determine the minimum and maximum possible values of x for strata 1 and 2. The value for x_3 is $x_3 = 7 - x_1 - x_2$. We then decrease x_3 by increments of 2 and increase x_2 by increments of 2, keeping x_1 fixed. Once x_2 achieves its maximum value, we increment x_1 by 2, set x_2 to its minimum value, determine $x_3 = 7 - x_1 - x_2$, and then again decrement x_3 and increment x_2. This process continues until x_1 achieves its maximum value. This pattern is easily viewed below with the example data of possible $\tilde{x}$ configurations that sum to 7, the observed total number of heterozygotes.

x_1	x_2	x_3
0	0	7
0	2	5
0	4	3
0	6	1
2	0	5
2	2	3
2	4	1
4	0	3
4	2	1
6	0	1

Table A1. Demonstration Genotype Data for Exact Stratified Test of HWE

Stratum	N_{AA}	N_{AB}	N_{BB}	Log-Probability	N_A	No. of Possible Heterozygotes
1	10	2	13	−11.0767	22	12
2	5	0	8	−8.3254	10	6
3	3	0	5	−4.9628	6	4
Total	18	2	26	−24.3649	…	…

Table A2. Sorted Values of Log-Probabilities for Each Stratum

Stratum	Log-Probabilities
1	−16.8068, −11.0767, −9.1269, −7.3077, −5.7640, −4.6405, −3.5127, −2.8022, −2.0658, −1.6673, −1.3036, −1.1748
2	−8.3254, −3.9433, −2.8980, −1.7097, −1.1707, −.8343
3	−4.9628, −1.5616, −1.4971, −.5808

Table A3. Sum of Log-Probabilities over Strata Configurations

Configuration Number	Strata Index 1	2	3	Sum of Log-Probabilities	Presence of Jump[a]
1	1	1	1	−30.0950	No
2	1	1	2	−26.6938	No
3	1	1	3	−26.6292	No
4	1	1	4	−25.7129	No
5	1	2	1	−25.7129	No
6	1	2	2	−22.3117	Yes
7	1	3	1	−24.6676	No
8	1	3	2	−21.2664	Yes
9	1	4	1	−23.4794	Yes
10	2	1	1	−24.3649	No
11	2	1	2	−20.9637	Yes
12	2	2	1	−19.9828	Yes
13	3	1	1	−22.4151	Yes

[a] Indicated for arrays where a jump over indices was possible.

Table A4. Demonstration Genotype Data for Exact Test of Homogeneity of HWD

Stratum	N_{AA}	N_{AB}	N_{BB}	N_A	Range of Possible Heterozygotes
1	10	2	13	22	0–22
2	5	2	8	12	0–12
3	3	3	5	9	1–9
Total	18	7	26	…	…

Web Resources

Accession numbers and URLs for data presented herein are as follows:

HapMap, http://www.hapmap.org/
hweStrata software, http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm
Perlegen, http://genome.perlegen.com/

References

1. Weir BS, Hill WG, Cardon LR (2004) Allelic association patterns for a dense SNP map. Genet Epidemiol 27:442–450
2. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079
3. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
4. Weir B (1996) Genetic data analysis II. Sinauer Associates, Sunderland, MA
5. Haldane HBS (1954) An exact test for randomness of mating. J Genet 52:631–635
6. Smith C (1970) A note on testing the Hardy-Weinberg Law. Ann Hum Genet 33:377–383
7. Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from the retrospective study of disease. J Natl Cancer Inst 22:719–748
8. McPeek MS (1999) Optimal allele-sharing statistics for genetic

mapping using affected relatives. Genet Epidemiol 16:225–249

9. Olson JM (1993) Testing the Hardy-Weinberg law across strata. Ann Hum Genet 57:291–295
10. Olson JM, Foley M (1996) Testing for homogeneity of Hardy-Weinberg disequilibrium using data sampled from several populations. Biometrics 52:971–979
11. Nam JM (1997) Testing a genetic equilibrium across strata. Ann Hum Genet 61:163–170
12. Troendle JF, Yu KF (1994) A note on testing the Hardy-Weinberg law across strata. Ann Hum Genet 58:397–402
13. Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76:887–893
14. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909