

## PROGRAM NOTE

# HW-QUICKCHECK: an easy-to-use computer program for checking genotypes for agreement with Hardy–Weinberg expectations

STEVEN T. KALINOWSKI

*Department of Ecology, Montana State University, Bozeman, Montana 59717, USA***Abstract**

**Deviations from Hardy–Weinberg expectations are frequently a sign of genotyping error. HW-QUICKCHECK is an easy-to-use computer program for detecting departures from Hardy–Weinberg equilibrium. HW-QUICKCHECK uses exact tests for all of its calculations. These tests include a global test for heterozygote excess/deficiency and genotype-specific tests.**

*Keywords:* error, exact test, genotype, Hardy–Weinberg, software

*Received 25 March 2006; revision accepted 3 May 2006*

Microsatellite loci are genotyped for many reasons – ranging from identifying individuals to defining species. Data analysis, however, almost always begins with a test for Hardy–Weinberg (HW) equilibrium. There are diverse biological reasons why a locus might not be in HW equilibrium (e.g. selection, inbreeding, population structure, migration), but most loci are probably close to HW equilibrium. Because most loci are close to HW equilibrium, departures from HW expectations are often a symptom of genotyping errors. Null alleles (e.g. Dakin & Avise 2004) and allelic dropout (e.g. Wattier *et al.* 1998), for example, can reduce the heterozygosity observed at a locus, and therefore cause samples to differ from HW expectations. Such problems are often locus-specific and can be difficult to detect. HW tests often are the first sign of their presence.

When systemic genotyping error is present, it is advantageous to discover this as soon as possible – so that errors can be corrected or so that unreliable loci can be dropped from analysis. Therefore, it is usually best *not* to wait until all samples have been genotyped before running HW tests. However, organizing genotypes into a text file that can be read by population genetics software takes time (the most precious commodity in any laboratory), and it is all too easy to postpone HW tests until after the laboratory phase of a project is complete. Data analysis then begins with an unpleasant surprise. This has happened to me more often than I would like to admit.

In order to prevent this problem, I have written a computer program, HW-QUICKCHECK, that simplifies HW testing. HW-QUICKCHECK is a user-friendly WINDOWS application designed for the single purpose of running HW tests. Data are input by copying genotypes from a spreadsheet (or similar document) and pasting them into a text box. Results can be read immediately and copied to another document if desired. There are no input files to format or output files to clutter a hard drive.

In addition to being user-friendly, HW-QUICKCHECK uses a nice set of exact tests to compare observed genotypes with HW expectations. HW tests are familiar to all molecular ecologists, but a detailed discussion of these tests is worthwhile because they have a few subtle complexities that are seldom acknowledged. The main goal of this note therefore is to describe the HW tests implemented by HW-QUICKCHECK.

I use the following notation. Let  $N$  represent the total number of genotypes in a sample; let  $N_i$  represent the number of times allele  $i$  is present in a sample; and let  $m$  represent the number of alleles in a sample, i.e.  $\sum_{i=1}^m N_i = 2N$ . Let  $N_{ii}$  and  $N_{ij}$  represent the number of times genotypes  $a_i a_i$  and  $a_i a_j$ , respectively, are observed,  $\sum_{i=1}^m N_{ii} + \sum_{i<j} N_{ij} = N$ . Finally, let  $N_{\text{Het}}$  represent the number of heterozygotes observed in a sample,  $N_{\text{Het}} = \sum_{i<j} N_{ij}$ .

Next, we need a probability model to calculate expected genotype counts and sample probabilities. The HW principle states that the genotype frequencies in a large population are a simple function of the allele frequencies in the population (e.g. Hedrick 2005). If, for example, the frequency

of allele  $i$  in the population is equal to  $p$ , genotype  $a_i a_i$  should occur with frequency  $p^2$ . A problem arises when we attempt to test whether this is true; we seldom know the parametric allele frequencies in a population, and, therefore cannot test whether genotypes in a sample agree with these frequencies. This difficulty is avoided by testing whether alleles observed in a sample are randomly paired into genotypes (e.g. Levene 1949; Haldane 1954; Guo & Thompson 1992). Let the random variable  $\mathbf{X}$  represent the number of times each possible genotype is observed when a set of  $m$  alleles having counts  $(N_1, N_2 \dots N_m)$  is randomly paired into genotypes. For example,  $X_{ii}$  represents the number of times  $a_i a_i$  is observed and  $X_{ij}$  represents the number of times  $a_i a_j$  is observed. We assume that the genotypes observed in our sample are the outcome of such random pairing. Given this model, the probability of any specific set of genotype counts  $(x_{11}, x_{12}, x_{22} \dots x_{mm})$  is equal to

$$P(x_{11}, x_{12}, x_{22} \dots x_{mm} | N_1, N_2 \dots N_m) = \frac{N! \prod_{i=1}^m N_i! \binom{2N_{\text{Het}}}{2N}}{(2N)! \left( \prod_{i=1}^m x_{ii}! \right) \left( \prod_{i < j} x_{ij}! \right)} \quad (1)$$

(Levene 1949).

Expected genotype counts for our model of random pairing of alleles can be obtained from Equation 1 (Levene 1949). If a sample of alleles having  $N_i$  copies of allele  $i$  is randomly paired into genotypes, the expected number times the homozygous genotype  $a_i a_i$  is observed,  $E(X_{ii})$ , is equal to

$$E(X_{ii}) = \frac{N_i(N_i - 1)}{4N - 2} \quad (2)$$

(Levene 1949). Note that this expression (Equation 2) differs slightly from the more familiar expression  $(N_i/N)^2$ . A simple example illustrates why Equation 2 is more accurate and more useful. Assume that allele  $i$  is observed once in a sample ( $N_i = 1$ ). If there is only one copy of allele  $i$  in a sample, there will be no  $a_i a_i$  homozygotes in the sample. The expected number of  $a_i a_i$  homozygotes therefore should equal zero. Note that Equation 2 satisfies this criterion but  $(N_i/N)^2$  does not. The difference between each expression is slight, but using Equation 2 makes it easier to compare observed and expected genotype counts.

Equation 2 shows the number of times a homozygous genotype is expected to be observed in a sample having allele counts  $(N_1, N_2 \dots N_m)$ . The corresponding expectation for heterozygous genotypes is

$$E(X_{ij}) = \frac{2N_i N_j}{4N - 2} \quad (3)$$

(Levene 1949).

In addition to calculating expected counts for each genotype, HW-QUICKCHECK performs two types of HW tests — a global test and genotype specific tests. The global test compares the entire sample with HW expectations. Genotype specific tests compare the observed count of each specific genotype with HW expectations. The global test will probably have more power to detect departures from HW expectations, but genotype specific tests may be useful for identifying specific genotypes that might contain errors.

There are two basic types of tests available for HW tests: parametric and exact. Parametric tests, such as the chi-squared test, are reliable when sample sizes are large and when there are not too many alleles in a sample, but can give misleading results when sample sizes are small or when many alleles are present (Guo & Thompson 1992; Wigginton *et al.* 2005). Exact tests calculate the probability of each possible combination of alleles and use these probabilities to compute how likely or unlikely the observed sample was. Exact tests are not biased by small sample size or by large number of alleles, and therefore are preferred to parametric alternatives. HW-QUICKCHECK uses exact tests for all of its testing.

All HW tests, whether parametric or exact, require a test statistic to measure how much a sample differs from HW expectations. There are many test statistics to choose from, and the choice affects the statistical power of the test (Rousset & Raymond 1995). I am not aware of any research identifying the most powerful test statistic for identifying genotyping errors, but if a null allele is present or allelic dropout is occurring, there should be an excess of all homozygotes and a deficiency of all heterozygotes. This scenario is mathematically equivalent to regular inbreeding, and Raymond & Rousset (1995) have shown that the  $U$ -statistic and the maximum-likelihood estimator of the inbreeding coefficient,  $\hat{f}_A$ , have the highest power for detecting inbreeding. Nonetheless, I have chosen the number of heterozygotes in a sample,  $N_{\text{Het}}$ , as a test statistic for HW-QUICKCHECK's global tests. This test statistic has only slightly less power than the other two alternatives, but is much easier to interpret. Knowing that a sample has fewer heterozygotes than expected, for example, is more useful for identifying genotyping errors than knowing that a sample has a high  $U$  score. In addition, all the genotype specific tests (described below) are based on observed and expected counts of homozygotes and heterozygotes.

Exact tests can be calculated by complete enumeration or by randomization (Guo & Thompson 1992). Complete enumeration is preferable when computationally feasible, but becomes infeasible when loci have more than a few alleles. In this case, Monte Carlo randomization is used to estimate the probability distribution for the number of heterozygotes observed in random pairings of alleles (Guo & Thompson 1992). This is carried out by randomly pairing

the alleles in a sample into genotypes, and recording the total number of homozygotes obtained. This is repeated thousands of times. The  $P$  value for a one-tailed test of heterozygote deficiency, for example, is equal to the proportion of times these random pairings of alleles have  $N_{Het}$  or less heterozygotes. HW-QUICKCHECK uses Monte Carlo randomizations for its global tests. Seventeen thousand randomizations are used for each test. This number provides a  $P$  value that is accurate enough for most purposes (Guo & Thompson 1992) without requiring more than a couple seconds of computing time for a desktop computer.

In addition to global tests, HW-QUICKCHECK performs genotype specific tests. Two sets of formulae are needed for such tests: one for homozygous genotypes, and one for heterozygous genotypes.

Haldane (1954) developed the test for homozygotes (see Weir 1996; p. 99 for a recent description). This test views a sample as containing two alleles — allele  $i$  and allele  $k$  (all other alleles). There are therefore three genotype classes in the sample —  $a_i a_i$ ,  $a_i a_k$  and  $a_k a_k$ . Let  $N_k$  represent the number of copies of allele  $k$  observed in the sample ( $N_i + N_k = 2N$ ). As above, let the random variable  $X_{ii}$  represent the number of  $a_i a_i$  genotypes observed in a sample in which alleles having counts ( $N_i, N_k$ ) are randomly paired into genotypes. If a sample has  $x_{ii}$   $a_i a_i$  homozygotes, it will have  $(N_i - 2x_{ii})a_i a_k$  heterozygotes and  $(N - N_i + x_{ii})a_k a_k$  homozygotes. The probability that  $X_{ii}$  has any specific value,  $x_{ii}$ , is specified by Equation 1

$$P(X_{ii} = x_{ii} | N_i, N_k) = \frac{N!N_i!N_k!2^{N_k}}{x_{ii}!(N_i - 2x_{ii})!(N - N_i + x_{ii})!(2N)!} \quad (4)$$

(Levene 1949; Haldane 1954). Equation 4 is useful for constructing one-tailed tests for homozygote excess or deficiency. For example, the probability of observing  $N_{ii}$  or more homozygotes in a sample having allele counts  $\{N_i, N_k\}$  is equal to

$$P(X_{ii} \geq N_{ii} | N_i, N) = \sum_{x_{ii}=N_{ii}}^{Int(\frac{N_i}{2})} P(X_{ii} = x_{ii} | N_i, N_k) \quad (5)$$

where  $Int()$  is the integer function (e.g.  $Int(7/2) = 3$ ) and  $Int(N_i/2)$  is the maximum number of  $a_i a_i$  homozygotes in a sample having  $N_i$  copies of allele  $i$ .

I have not been able to find a corresponding test for evaluating whether there is an excess or deficiency for a specific heterozygous genotype, but deriving one is not difficult. As above, the test is facilitated by condensing the data. In this case, three alleles are recognized —  $i, j$  and  $k$  (all alleles other than  $i$  or  $j$ ). There are therefore six genotype classes —  $a_i a_i, a_i a_j, a_i a_k, a_j a_j, a_j a_k, a_k a_k$ . Let the random variables  $X_{ii}, X_{jj}, X_{kk}, X_{ij}, X_{ik}, X_{jk}$  represent the number of times each possible genotype is observed in a sample after the alleles

in the sample are randomly paired into genotypes. If the alleles in a sample are randomly paired into a genotypes, the probability of observing any specific set of genotype counts,  $(x_{ii}, x_{jj}, x_{kk}, x_{ij}, x_{ik}, x_{jk})$ , is specified by Equation 1

$$P(x_{ii}, x_{jj}, x_{kk}, x_{ij}, x_{ik}, x_{jk} | N_i, N_j, N_k) = \frac{N!N_i!N_j!N_k!2^{(N_{ij}+N_{ik}+N_{jk})}}{(2N)!x_{ii}!x_{jj}!x_{kk}!x_{ij}!x_{ik}!x_{jk}!} \quad (6)$$

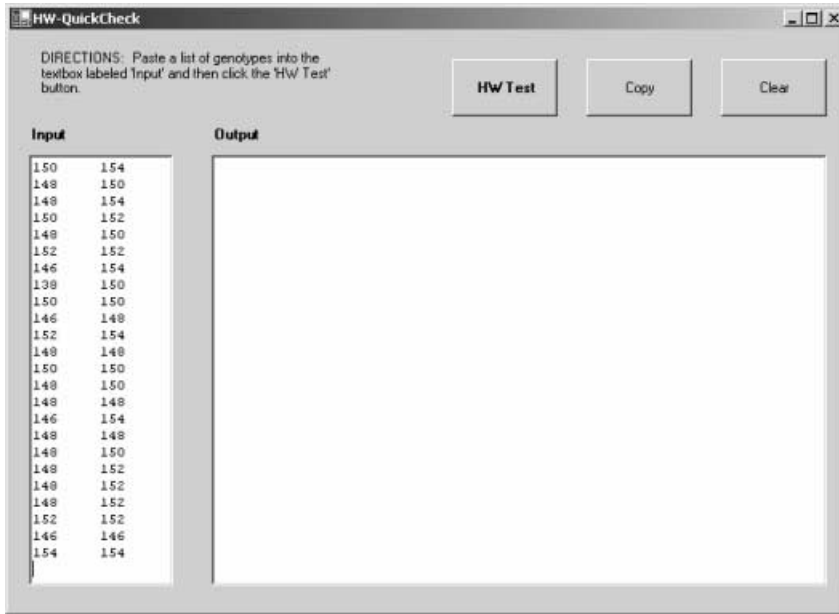
Note that Equation 6 specifies the probability of observing genotype counts  $(x_{ii}, x_{jj}, x_{kk}, x_{ij}, x_{ik}, x_{jk})$ . This is not equal to the probability of observing  $x_{ij}$   $a_i a_j$  heterozygotes. The probabilities are not equal because, in most cases, there will be many possible genotype configurations that have  $x_{ij}$   $a_i a_j$  heterozygotes. Calculating the probability of observing  $x_{ij}$  heterozygotes requires summing over all these possible samples. This summation is facilitated by noting that if  $x_{ij}, x_{ii}$  and  $x_{jj}$  are specified, each of the three other genotype counts can be calculated from the allele counts ( $x_{ik} = N_i - x_{ij} - 2x_{ii}$ ;  $x_{jk} = N_j - x_{ij} - 2x_{jj}$ ;  $x_{kk} = N - x_{ii} - x_{jj} - x_{ij} - x_{ik} - x_{jk}$ ). The probability that  $X_{ij}$  is equal to any specific value,  $x_{ij}$ , is equal to

$$P(X_{ij} = x_{ij} | N_i, N_j, N_k) \sum_{x_{ii}} \sum_{x_{jj}} \left\{ \begin{array}{l} \frac{N!N_i!N_j!N_k!2^{(x_{ii}+x_{jj}+x_{kk})}}{(2N)!x_{ii}!x_{jj}!x_{kk}!x_{ij}!x_{ik}!x_{jk}!} \\ \text{Where} \\ x_{ik} = N_i - x_{ij} - 2x_{ii} \\ x_{jk} = N_j - x_{ij} - 2x_{jj} \\ x_{kk} = N - x_{ii} - x_{jj} - x_{ij} - x_{ik} - x_{jk} \end{array} \right\} \quad (7)$$

where summation is taken over all values for  $x_{ii}$  and  $x_{jj}$  that are possible given the allele counts in the sample. One-tailed tests for heterozygote excess or deficiency are calculated using Equation 7. For example, the  $P$  value for a one-tailed test of heterozygote deficiency is equal to

$$P(X_{ij} \leq N_{ij} | N_i, N_j, N_k) = \sum_{x_{ij}=0}^{N_{ij}} P(X_{ij} = x_{ij} | N_i, N_j, N_k) \quad (8)$$

Deciding whether to perform one- or two-tailed tests deserves careful consideration. One-tailed tests will be more powerful than two-tailed tests if the direction of the test is chosen correctly, but knowing a priori which direction genotypes are likely to depart from HW expectations can be difficult. Null alleles and allelic dropout will decrease heterozygosity. Therefore, a one-tailed test for heterozygote deficiency should be a powerful method for detecting null alleles or allelic dropout. However, microsatellite stuttering could cause homozygotes to be scored as heterozygotes. Such a problem would be missed in a one-tailed test for heterozygote deficiency; a one-tailed test for heterozygote excess would be the most effective method for detecting such a problem. HW-QUICKCHECK presents one-tailed probabilities for either genotype excess or deficiency — depending on what is observed. For example, if there are fewer homozygotes observed than expected, HW-QUICKCHECK presents the one-tailed probability



**Fig. 1** The user interface for HW-QUICKCHECK. The user has copied a list of genotypes from a spreadsheet and pasted them onto the HW-QUICKCHECK form.

for observing that few (or fewer) homozygotes. If there are more homozygotes observed than expected, HW-QUICKCHECK presents the probability of that many homozygotes (or more) being observed. This type of testing gives the user the most power to detect departures from HW expectations, but it must be interpreted correctly.

Interpretation of these one-tailed  $P$  values depends on whether the user has an a priori alternative hypothesis. If, for example, the user wanted to know a priori whether there is a deficiency of heterozygotes, all  $P$  values associated with heterozygote deficiencies can be interpreted in the usual manner. If this is carried out, however, heterozygote excesses will have to be ignored. If the user wants to perform a two-tailed test for either heterozygote excess or deficiency, the one-tailed  $P$  values reported by HW-QUICKCHECK should be doubled (Yates 1984; see discussion in Weir 1996; p. 100).

What should be avoided is scanning the  $P$  values with no a priori hypotheses in mind, finding a  $P < 0.05$ , and concluding that there is a statistically significant deviation from HW expectations. Such a practice would double the probability of a type I error (rejecting the null hypothesis when it is true). This happens because there is a 0.05 probability of observing a heterozygote deficiency having a one-tailed  $P$  value of  $\leq 0.05$  when the null hypothesis is true; and there also is a 0.05 probability of observing a heterozygote excess with a one-tailed  $P$  value  $\leq 0.05$  when the null hypothesis is true. Therefore, there is a 0.10 probability of observing either a heterozygote deficiency or excess having a one-tailed  $P$  value  $\leq 0.05$  — which is why one-tailed  $P$  values need to be doubled if the user wants a two-tailed test.

However, the practice of searching results for any deviations from HW expectations without an a priori

hypothesis is a good method for identifying possible problems to check with subsequent sets of samples. For example, a user may detect a suspicious excess of a certain type of homozygote in an early set of samples, and then watch subsequent sets of samples to see whether the same excess is observed.

My motivation for writing HW-QUICKCHECK was to help molecular ecologists detect genotyping errors as quickly as possible. However, I should emphasize that genotyping errors are not the only cause of deviations from HW expectations. There are diverse biology causes of non-random association of alleles (e.g. hidden population structure) that the user must consider. One informative test for whether deviations from HW at a locus are caused by technical or biological processes is whether the same pattern is observed at multiple loci. Hidden population structure, for example, should decrease heterozygosity at all loci. Null alleles, on the other hand, will only decrease heterozygosity only at loci having nulls.

HW-QUICKCHECK is available for free download from the author's website, [www.montana.edu/kalinowski](http://www.montana.edu/kalinowski). The zip file containing the program is only 11 k, so downloading HW-QUICKCHECK does not take long. HW-QUICKCHECK should work with recent versions of MICROSOFT WINDOWS (1998 or later). Genotypes are input into HW-QUICKCHECK by copying a list of genotypes from a spreadsheet, text file, MICROSOFT WORD document, or similar document (Fig. 1). Each line in the genotype list should correspond to the genotype of an individual. Alleles can be separated by spaces (e.g. '100 102'), tabs (e.g. '100 102'), or be written in GENEPop format (e.g. '100102') (Raymond & Rousset 1995). A user guide with instructions and sample data is available at the website cited above. Sample output is shown in Appendix.

## References

- Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. *Heredity*, **93**, 504–509.
- Guo SW, Thompson EA (1992) Performing the exact test of Hardy–Weinberg proportions for multiple alleles. *Biometrics*, **48**, 361–372.
- Haldane JBS (1954) An exact test for randomness of mating. *Journal of Genetics*, **52**, 631–635.
- Hedrick PW (2005) *Genetics of populations*. Jones & Bartlett, Sudbury, Massachusetts, USA.
- Levene H (1949) On a matching problem arising in genetics. *Annals of Mathematical Statistics*, **20**, 91–94.
- Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.
- Rousset F, Raymond M (1995) Testing heterozygote excess and deficiency. *Genetics*, **140**, 1413–1419.
- Wattier R, Engel CR, Saumitou-Laprade P, Valero M (1998) Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in *Gracilaria gracilis* (Rhodophyta). *Molecular Ecology*, **7**, 1569–4573.
- Weir BS (1996) *GENETIC DATA ANALYSIS II*. Sinauer Associates, Sunderland, Massachusetts, USA.
- Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy–Weinberg equilibrium. *American Journal of Human Genetics*, **76**, 887–883.
- Yates F (1984) Tests of significance for  $2 \times 2$  contingency tables. *Journal of Royal Statistical Society. Series A*, **147**, 426–463.

## Appendix

Sample output from HW-QUICKCHECK

---

### Summary

Sample size: 59  
 N alleles: 6  
 $H_O$ : 0.64  
 $H_E$ : 0.78

### Allele frequencies

138 0.01 ← Singleton  
 146 0.08  
 148 0.26  
 150 0.29  
 152 0.23  
 154 0.13

Global test	obs	exp	sign	P value
Homozygotes	21	13.1	>	
Heterozygotes	38	45.9	<	0.0103

Homozygotes	obs	exp	sign	P value
138/138	0	0.0	=	ns
146/146	1	0.4	+	ns
148/148	5	4.0	+	ns
150/150	6	4.8	+	ns
152/152	7	3.0	+	0.006
154/154	2	0.9	+	ns

Heterozygotes	obs	exp	sign	P value
138/146	0	0.1	-	ns
138/148	0	0.3	-	ns
138/150	1	0.3	+	ns
138/152	0	0.2	-	ns
138/154	0	0.1	-	ns
146/148	3	2.6	+	ns
146/150	2	2.9	-	ns
146/152	1	2.3	-	ns
146/154	2	1.3	+	ns
148/150	10	9.0	+	ns
148/152	5	7.2	-	ns
148/154	3	4.0	-	ns
150/152	5	7.8	-	ns
150/154	4	4.4	-	ns
152/154	2	3.5	-	ns

---