Hardy-Weinberg Proportions Methods Manual Version 0.1.2-(September 9, 2009):

Testing fit of genotype frequencies to Hardy-Weinberg proportions

Available online at: www.ImmPort.org

Glenys Thomson[1,*], Hazael Maldonado-Torres[2], Alex K. Lancaster[1,3], Jill A. Hollenbach[4], Lisa F. Barcellos[5], Steven J. Mack[4], Richard M. Single[6]

[1] Department of Integrative Biology, 3060 Valley Life Sciences Building MC #3140, University of California, Berkeley, CA 94720-3140, USA, e-mail: glenys@berkeley.edu

[2] Anthony Nolan Research Institute and Department of Haematology, UCL Cancer Institute Royal Free Campus, Hampstead, NW3 2QG, UK, h.maldonado@ucl.ac.uk

[3] Mailing address: Department of Biological Sciences, University of Massachusetts Lowell, Lowell MA 01854, USA, e-mail: alexl@cal.berkeley.edu

[4] Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr Way, Oakland, CA 94609, jhollenbach@chori.org, sjmack@chori.org

[5] Division of Epidemiology, School of Public Health, University of California, Berkeley, CA 94720-7356, USA, e-mail: barcello@genepi.berkeley.edu

[6] Department of Mathematics and Statistics, 306 Mansfield House, University of Vermont, Burlington, VT 05405, USA, richard.single@uvm.edu

I. Overview

A. Introduction

The concept of testing genotype frequencies for fit to Hardy-Weinberg proportions (HWP) is simple and straightforward, and the usual first step in data analyses. Observed deviations from HWP can tell us a lot about population and patient samples (non-random mating, admixture), the accuracy of the genotyping, and selection. Included in the broad topic of selection is differential risk to disease, which in patients can manifest in *overall* genotype-level deviations from HWP, deviations for *individual* or specific sets of genotypes, or for *overall heterozygosity*.

As with all data analyses, we encourage manual inspection and subsequent manipulation of the data by the researcher following *overall* and *individual* genotype-level analyses of the data. This can lead to important insights, and is currently the only way to optimize analyses. Using researcher based knowledge of the disease, the genetic system under study, the typing methods employed, the population sampled, and the results of association studies, analyses can be refined to target specific sets of alleles or haplotypes, and their genotypes, with respect to HWP.

Remaining research issues with testing fit to HWP involve the statistical power of different test statistics with respect to alternative hypotheses. Note that we do not consider all possible test statistics in our discussions below. Also, for highly polymorphic genes, in terms of the biological relevance of the tests, further research is needed on how best to handle rare alleles and genotypes in the analyses. While no computer software package at this time contains all the tests we outline below, we list in Section I.I three websites that together cover the spectrum of these tests.

Note that no references are given in this section; a complete listing of references is given in each detailed section following this overview of HWP and testing thereof.

B. Hardy-Weinberg proportions (HWP)

A population is said to be at Hardy-Weinberg equilibrium (HWE) for a particular locus (gene) when the observed genotype frequencies are not statistically different from the frequencies determined by the appropriate products of its allele frequencies; these expected genotype frequencies are termed Hardy-Weinberg proportions (HWP). The most familiar format considers a locus denoted A, with two distinct alleles denoted A_1 and A_2, with respective allele frequencies p_1 and p_2 ($p_1 + p_2 = 1$). The HWP in this case are:

A_1A_1: p_1^2, A_1A_2: $2p_1p_2$, A_2A_2: p_2^2.

For a locus with k distinct alleles and allele frequencies p_i ($\Sigma\, p_i = 1$), the expected HWP are:

p_i^2 for homozygotes (A_iA_i), and $2p_ip_j$ for heterozygotes (A_iA_j), where $i < j$, and $i, j = 1, 2, ..., k$.

These expectations are derived from Mendel's first law of independent segregation of the two parental alleles when applied to *population* level variation.

Note that one must be able to *distinguish* all heterozygous and homozygous individuals, i.e., the test cannot be applied to traits with a recessive or dominant mode of inheritance; instead, the trait must show codominance or incomplete dominance. With molecular typing, e.g., of the

highly polymorphic HLA loci, single nucleotide polymorphisms (SNPs), and microsatellite (MSAT) loci, traits are codominant, unless typing errors obscure the true genotype.

The two-allele example above is usually described in terms of alleles *A* and *a*, and genotypes *AA*, *Aa*, and *aa*; the notation A_1 and A_2 is used here instead since the use of upper and lower case letters is traditionally used to designate dominant versus recessive alleles respectively. Further, while the more familiar notation denotes the allele frequencies by p and q ($p + q = 1$), we use the notation p_1 and p_2 as it readily extends to multiple alleles.

C. Testing for fit to HWP and factors causing deviations from HWP

In a test of HWE (i.e., fit to HWP allowing for random sampling effects), allele frequencies are estimated from the *observed* genotype counts and then the *expected* genotype counts under HWP are estimated from these allele frequencies. The *observed* and *expected* genotype counts are then compared using an appropriate statistical test. Significant deviation of genotype counts from HWP can be due to a number of factors, including sampling of admixed, stratified, or some other form of blended populations, inbreeding or other forms of non-random mating, genotyping errors, and selection.

Statistical tests to detect significant deviations from HWP have low power, and for most genetic markers in relatively well-defined populations significant deviation from HWE is *not* common. When significant deviations from HWP are seen, genotyping errors or potential population stratification are the first consideration rather than the operation of natural selection. Testing fit to HWP is a crucial first step in quality control (QC) in verifying the integrity of genotype data, especially for highly polymorphic genotype data.

The largest body of evidence for selection via deviations from HWP comes from disease studies; deviations from HWP may be observed for a causative genetic variant, e.g., HLA genes, as well as all markers in linkage disequilibrium (LD) with the causative variant. Thus, a marker should not be removed from an analysis due to lack of fit to HWP in patients *unless* there is also a corresponding lack of fit in controls along with further investigation of the individual genotypes contributing to an overall deviation from HWP.

D. Overall genotype-level tests of HWP: asymptotic, exact, and approximation methods

Tests of the significance of *overall* genotype-level deviations from HWP combine information across all genotypes at a locus. Different methods can be categorized into three main groups:

asymptotic (standard) tests, e.g., the Chi-square goodness-of-fit test,

exact (complete enumeration - the “gold standard”) tests, and

approximation (resampling) tests, e.g., Monte Carlo (MC), Markov Chain Monte Carlo (MCMC).

The Chi-square test historically was the standard approach for testing fit to HWP at the *overall* genotype-level, while more recently the MC and MCMC resampling tests have been used. It is well documented that *asymptotic* tests can sometimes lead to false acceptance or rejection of the null hypothesis when asymptotic distributional assumptions for statistics are not met. This issue arises in particular when the sample size is small and/or the expected genotype counts are small or close to zero (i.e., there are sparse cells).

Tables of all possible genotypes for highly polymorphic loci in particular may have many sparse cells. Notably, the classical HLA class I A, B, and C and the class II DRB1, DQA1, DQB1, DPA1, and DPB1 loci are extremely polymorphic. Some of these loci have more than 1,000 described alleles; and it is not uncommon to observe 30 to 40 or more distinct alleles at a given locus in many populations worldwide, leading to many sparse cells. In these cases, many alleles must be lumped together in order to apply the Chi-square test. Alternatively, resampling MC or MCMC tests can be applied.

An *exact* (complete enumeration) algorithm has recently been developed for an arbitrary number of distinct alleles (k) and an arbitrary sample size (n), and we refer to this as the "gold standard." Its applicability to large samples with large numbers of alleles is currently limited by the available computer resources. The resampling MC and MCMC tests perform very favorably when compared to the *exact* test, and always outperform the *asymptotic* Chi-square test. It has been shown that in a few cases the MCMC method may fail to approximate to the exact p-value, hence in cases where the *exact* test cannot be performed due to large sample size and numbers of alleles the MC test is in general preferred.

E. Individual genotype-level tests of HWP

Identifying *individual* genotype deviations from HWP is also of interest for both QC of the data and in population and disease studies. For example, if a particular genotype has been typed incorrectly, it may show significant deviations from HWP, either excess or deficiency, in laboratories using that protocol. Also, some specific genotypes may be subject to selection at the population level, and study of *individual* genotype deviations from HWP may provide increased power to detect that selection. With admixture, the genotypes with the greatest differences in allele frequencies between the contributing populations are expected to show more deviation from HWP. In disease studies, specific genotypes may show significant differential risk, which may be detected via deviations from HWP. For example, in patients with type 1 diabetes (T1D) there is an excess of HLA DR3/DR4 individuals over the respective homozygotes (DR3 and DR4 as used here are abbreviations for the well-known HLA DRB1-DQB1 haplotypes associated with T1D).

The relative magnitude of contributions to significant deviation from HWP for each *individual* genotype was previously assessed by considering the p-value for the *asymptotic* 1 degree of freedom (df) Chi-square test for individual genotypes; however, such a test is always *invalid* (and tends to be conservative) as the individual contributions are not independent. Appropriate *asymptotic* tests for individual heterozygote and homozygote cases have been developed.

Individual genotype-level tests have recently been incorporated into the resampling MC and MCMC *overall* genotype-level tests, as well as the *exact* test. We have also shown that the p-value for *individual* genotype-level testing can always be calculated via an *exact* test applied to the appropriate 2x2 genotype table for homozygotes (alleles A_i and A_x, where x denotes all non-A_i alleles, and genotypes A_iA_i, A_iA_x, A_xA_x), and for heterozygotes (and also for the respective homozygotes) the appropriate 3x3 table (alleles A_i, A_j, and A_y, where y denotes all non-A_i, non-A_j alleles, and genotypes A_iA_i, A_iA_j, A_jA_j, A_iA_y, A_jA_y, A_yA_y).

F. Overall heterozygote excess/deficiency, and all heterozygotes for a specific allele

As well as individual genotype deviations from HWP, specific sets of genotypes may overall show deviations from HWP. This may occur in MSAT genotyping as a result of preferential amplification or allelic dropout (alleles that cannot be distinguished from background noise). Similarly, with null MSAT alleles (alleles that cannot be amplified due to sequence variation under PCR primers), there could be a significant excess of some homozygotes. Conversely, stuttering with MSAT typing can result in homozygotes being scored as heterozygotes. SNP data is also subject to deviations from HWP due to the typing procedure, and SNP data showing lack of fit to HWP are routinely removed from analyses; the cause is usually not examined, although it is often attributable to allelic dropout due to sequence variation in the primer region.

HLA genotyping depends on the hybridization of PCR primers, and DNA probes in some cases, which may fail if novel sequence variants are present under the primer or probe. In the case of a novel variant under a primer, an allele can go undetected, resulting in an excess of homozygous genotypes. In the case of a novel variant under a probe, both alleles in a genotype may be incorrectly assigned (resulting in an increased number of rare alleles, and therefore an increase in sparse cells).

Given the central role HLA molecules play in disease association, heterozygosity *per se* may confer a selective advantage, as seen with progression to AIDS. It is thus of interest to study deviations from HWP for *all* homozygotes, *all* heterozygotes, as well as heterozygotes for each specific allele, e.g., A_1A_x, where $x = 2, 3, .., k$. In the latter case, these values are obtained from the appropriate 2x2 genotype table discussed above. Selection at the population level, and disease risk, may act at any of these levels.

As mentioned above, we do not cover all possible tests of HWP in our discussion below. Some of these additional tests may be more powerful depending on the alternative hypothesis being considered, for example, admixture and other forms of population stratification, as well as inbreeding, which lead to an expected heterozygote deficiency.

G. Rare alleles and genotype classes

As noted above, *individual* genotype-level *p*-values can always be calculated using an *exact* test. However, when the expected values are sufficiently small, the test *may not* be biologically meaningful, leading to a potential for over-interpretation of associated *p*-values. This results from the fact that genotype and allele counts are restricted to integer values, while expected values for genotypes involving rare alleles may be much less than 1. Comparing, for example, an observed count of 1 to an expected of 0.10 is in general not appropriate. Hence, we propose an *arbitrary* cut-off so that *p*-values for individual genotypes will only be reported when the expected genotype count under HWP is ≥ 2. This is strictly speaking a rule of thumb and the cut-off used by different researchers may vary. Our recommendation of using a cut-off of two may be revised upon further numerical research. Further, given that researchers have noted cases with low expected but quite a large observed genotype count, and that this may serve as a signal of admixture, researchers may wish to note such observations.

H. Application to population and patient data sets

As emphasized above, genotyping errors are always the first possibility to consider in quality control (QC) of HLA, and other, data. While some authors argue against using lack of fit to HWP to identify genotyping errors for some marker types, nevertheless with HLA data in

population studies it has been validated as useful for that purpose, although probably with low power. Deviations from HWP may be due to issues with a primer or probe. Non-random sampling of individuals, for example from bone marrow transplantation registries, can also lead to significant deviations from HWP. Admixture is usually identified by the presence of alleles or haplotypes which are recognized as probably coming from a different population; excluding potentially admixed individuals may improve the fit to HWP. Evidence for selection based on HWP has been found in a number of HLA studies.

For disease studies, deviations from HWP can indicate the presence of a disease gene. Further, the form of deviations can indicate specific genotype effects on disease risk. Fit to HWP with a patient sample is expected only for a marker in LD with a strictly recessive disease model.

I. Software available for tests of HWP

We give three websites that together cover the spectrum of tests we discuss (below we refer to these as PyPop, Cactus, and HW-QUICKCHECK respectively):

(1) www.pypop.org and www.ImmPort.org (Python for Population Genomics – PyPop, current release version 0.7.0.) (Lancaster et al. 2003, 2007a, 2007b; Lancaster 2006)

PyPop is affiliated with ImmPort.org, the Immunology Database and Analysis Portal. The ImmPort system provides advanced information technology support in the production, analysis, archiving, and exchange of scientific data.

(2) www.cactus-project.org/hardy-weinberg (Maldonado Torres 2009)

(3) www.montana.edu/kalinowski/ (HW-QUICKCHECK) (Kalinowski 2006).

Many other websites also cover one or more of the tests listed below.

Overall genotype-level tests of HWP

asymptotic: the Chi-square goodness-of-fit test - PyPop

The default setting for the Chi-square test appropriately lumps all genotype classes with $E_i < 5$; in cases with large numbers of rare alleles there may be no *df* left for statistical testing of the data.

exact: the complete enumeration algorithm for the general case - Cactus, PyPop in a later release

This algorithm applies to any number of alleles (k) and sample size (n), its applicability to large samples with large numbers of alleles is currently limited by the available computer resources.

approximation: resampling Monte Carlo (MC) and Monte Carlo Markov Chain (MCMC):

MC - PyPop, MCMC - PyPop and Cactus

The MC test in PyPop is based on the code of the method described in Guo and Thompson (1992). The MCMC test in PyPop is based on a modified version of the code from Guo and Thompson (1992) that allows for large numbers of distinct alleles. In both cases (MC and MCMC) the overall genotype-level *p*-value is approximated by running the algorithm for a fixed number of steps and recording the number of randomly generated samples that have a conditional probability less than or equal to that from the

observed sample. The *p*-value is then computed by dividing the number of such samples found by the total number of samples generated. In the MCMC case, the algorithm is run for *B* dememorisation steps ("burn-in") before statistics are collected.

Individual genotype-level tests of HWP

asymptotic: the Chi-square goodness-of-fit 1 *df* test - PyPop, valid asymptotic tests (Chen statistics) - PyPop in a later release

The Chi-square goodness-of-fit 1 *df* test is always invalid and conservative, and is useful only as a relative measure. Appropriate *asymptotic* test statistics for heterozygotes and homozygotes are available (Section I.E).

exact: using the full set and appropriate 3x3 subsets - Cactus, PyPop in a later release

For a sample with *k* observed alleles, the *exact* test for individual heterozygotes and homozygotes can always be calculated via the reduction of the observed genotype distribution into a set of $k(k-1)/2$ tables of the appropriate 3x3 genotype sub distributions, even when computing power limits analysis of the full distribution.

approximation: resampling MC and MCMC tests - PyPop, HW-QUICKCHECK

Recently, *individual* genotype-level tests have been added to the MC and MCMC tests in PyPop, however, the suggested cut-off for rare genotypes has not yet been implemented.

Overall heterozygote excess/deficiency

exact: using the full set - Cactus and PyPop in later releases

approximation: resampling MC test - HW-QUICKCHECK, PyPop in a later release

All heterozygotes for a specific allele

exact: using the full set, and the *k* subsets of 2x2 tables - Cactus and PyPop in later releases

approximation: resampling MC and MCMC tests - PyPop in a later release

Rare alleles and genotype classes

Later releases of both Cactus and PyPop will include cut-offs for rare alleles and genotypes so that the *p*-values will be indicated with an asterisk when the test is not biologically reasonable.

II. Hardy-Weinberg proportions (HWP)

A. Introduction

Consider the three examples of population level genotype frequencies shown in Table II.1: the *MN* gene encodes blood group antigens; the *Δ32* variant of the *CCR5* gene protects against progression to AIDS; and cystic fibrosis which is a recessive trait (*cc* individuals are affected and *Cc* individuals are carriers: these heterozygotes can be detected using molecular typing methods).

One might ask: Why are these particular genotype frequencies observed? Is there some relationship between allele frequencies and the genotype frequencies? And why are the heterozygote frequencies so much larger than the homozygous frequencies for the *Δ32/Δ32* and

cc genotypes? These are explained by the fact that these three cases have observed genotype counts very close to those expected under HWP.

Table II.1: Comparison of observed and expected genotype counts under HWP*

(1) MN blood group system in an Egyptian population: $p_1 = 0.5225$, $p_2 = 0.4775$

Genotypes:	*MM*	*MN*	*NN*	
Observed numbers	278	489	233	Total: 1,000
Expected HWP	273	499	228	Total: 1,000

(2) CCR5 *Δ32* variant in a French population: $p_1 = 0.89$, $p_2 = 0.11$

Genotypes:	*A/A*	*A/Δ32*	*Δ32/Δ32*	
Observed numbers	795	190	15	Total: 1,000
Expected HWP	792.1	195.8	12.1	Total: 1,000

(3) Cystic fibrosis in a European population: $p_1 = 0.9796$, $p_2 = 0.0204$

Genotypes:	*CC*	*Cc*	*cc*	
Observed numbers	9,596	400	4	Total: 10,000
Expected HWP	9,596.2	399.6	4.2	Total: 10,000

* The data are modified from real examples to give total sample sizes that are easy to visualize and use in calculations.

In each of these three examples, the genotype frequencies are very close to (and in fact not statistically different from) those expected under HWP. To estimate the HWP in each case, first the allele frequencies, denoted p_1 and p_2 ($p_1 + p_2 = 1$), are estimated by the method of gene (allele) counting. The HWP expectations are then determined by the appropriate products of the allele frequencies:

A_1A_1: p_1^2, A_1A_2: $2p_1p_2$, A_2A_2: p_2^2;

and the expected genotype *counts* are obtained by multiplying the HWP by the sample size (n). (See Appendix A for the method of gene (allele) counting and the derivation of HWP). Note that the trait under study must show codominance or incomplete dominance so that all heterozygote and homozygote individuals can be distinguished. Observed genotype counts are then compared to those expected under HWP using an appropriate statistical test (discussed later).

The extension to multiple alleles is straightforward, for a locus with k alleles and allele frequencies p_i, the expected HWP are:

$A_iA_i : p_i^2$ for homozygotes, $A_iA_j : 2p_ip_j$- for heterozygotes, where $i < j$, and $i, j = 1, 2, ..., k$.

For a 2 allele X-linked trait (with obvious extensions for multiple alleles), the expected HWP are:

females: $A_1A_1 : p_1^2$, $A_1A_2 : 2p_1p_2$, $A_2A_2 : p_2^2$, males: $A_1 : p_1$, $A_2 : p_2$,

and statistical testing for HWP can only be carried out in females (see Appendix A).

B. Assumptions in calculation of HWP

The effect of Mendelian segregation on the genetic structure of a population is the central consideration for the calculation of HWP. Because we want to consider the effect of this factor acting in isolation from all other possible factors, we assume:

(a) random mating, e.g., no inbreeding

(b) migration, admixture, etc. rates and mutation rates are negligible,

(c) no selection (including disease risk in patient samples),

(d) segregation according to Mendelian rules, and

(e) a very large population (in theory, an infinitely large population).

These five assumptions are known to those familiar with classical population genetics. However, given the variety of molecular typing methods currently in use, and the possibility of error with all of them, we *add here* a sixth assumption:

(f) no genotyping errors.

Students often asked why such stringent criteria must be considered, given that no real population will exactly satisfy all of them. This question is often followed by the suggestion that modern computers might be employed to consider all the variables simultaneously. Nevertheless, a very stringent and simplistic model is used in order to understand the effect of Mendelian segregation on short and long-term population effects, independent of other evolutionary parameters. At the same time, the power of HWP tests is low; deviations from HWP will only be detected if the effect of any one of these factors is very large. For practical purposes, moderate individual violations of these assumptions do not cause significant deviations from HWP at many loci.

As mentioned above, when significant deviations from HWP are seen, genotyping errors are the first consideration. Note however that there is debate about the exclusion or not of data due to lack of fit to HWP for certain types of genetic polymorphisms (see e.g., Zou and Donner 2006, Teo et al. 2007). Significant deviation from HWP has been observed due to a number of factors, including sampling of admixed, stratified, or some other form of blended populations, inbreeding or other non-random mating (see e.g., Bourgain et al. 2004), and selection (see e.g., Hedrick 2003, Meyer and Thomson 2001, Section IX, and Appendix C).

Finally, it should be noted that the assumption of an infinite population allows for no sampling variance at the population level. This is equivalent to assuming one gets 50% heads in an infinite series of coin tosses. The statistical tests described below however account for the fact that the sample size is *finite*.

C. Relationship between HWP and allele frequencies

In Table II.2, HWP are given for some representative values of the allele frequency p_1 of allele A_1, and hence p_2 of allele A_2 ($p_1 + p_2 = 1$). These show that for relatively rare alleles, the frequency of heterozygotes is much larger than for the homozygote (see examples 2 (the *Δ32* variant of the *CCR5* gene) and 3 (cystic fibrosis) in Table II.1).

Table II.2: HWP for a range of allele frequencies and the *gene diversity (H)* index

A_1	A_2	A_1A_1	A_1A_2	A_2A_2
p_1	p_2	p_1^2	$2p_1p_2$	p_2^2
.01	.99	.0001	.0198	.9801
.05	.95	.0125	.095	.9025
.1	.9	.01	.18	.81
.2	.8	.04	.32	.64
.3	.7	.09	.42	.49
.4	.6	.16	.48	.36
.5	.5	.25	.50	.25

Note from Table II.2 that for a two-allele system, the maximum heterozygosity under HWP occurs when the two alleles have equal frequency of 1/2. For a given set of allele frequencies, the expected proportion of heterozygotes under HWP is the probability that two randomly chosen allele copies will be different, and is referred to as the gene diversity (H) index. In summarizing the amount of genetic variation at a polymorphic locus, we usually give the number of observed distinct alleles (k), the gene diversity index (H), and the sample size (n). Note that for multiple alleles, the maximum possible gene diversity index for a given number of alleles (k) observed in a sample, occurs when these alleles all have equal frequency of $1/k$, in which case $H = (k\text{-}1)/k = 1 - 1/k$, and the observed homozygosity under HWP $(F) = 1/k$.

For recessive traits, a test of fit to HWP can not be carried out. However, we can use the fact that most genes show fit to HWP, to *estimate* the allele frequency, and use this to *estimate* the frequency of heterozygotes (carriers) of the recessive allele (see Appendix A).

III. *Asymptotic* (standard) Chi-square *overall* test of HWP

The basis of the *asymptotic* (standard) goodness-of-fit Chi-square test of HWP is to calculate a standardized measure of deviation from HWP for each individual genotype and then add together these individual contributions ,$(O\text{-}E)^2/E$ where O is the observed genotype count and E is the expected count under HWP.The sum of these individual values gives the Chi-square test statistic value, with a specified degrees of freedom (*df*); and the relevant *p*-value under the null hypothesis is then determined. It is usual to require $E \geq 5$ for all genotypes, which may necessitate appropriate lumping of some alleles or genotypes. (See Appendix B for details on calculating the Chi-square test statistic, including calculation of *df* and lumping of classes, and accepting or rejecting the null hypothesis and *p*-values in the latter case.)

It is well documented that the *asymptotic* tests can sometimes lead to false acceptance or rejection of the null hypothesis, in particular when the sample size is small and/or the expected genotype frequencies are small or close to zero (Cochran 1954; Emigh, 1980; Louis and Dempster 1987; Guo and Thompson 1992; Wigginton et al. 2005). Other proposed asymptotic tests include the Freeman-Tukey test (Freeman et al. 1950), the conditional Chi-square test (Li

1955), the Mantel-Li test (Mantel and Li 1974), the likelihood-ratio test (Elston and Forthofer 1977; Hernandez and Weir 1989), and the Kullback-Leibler test (Ebrahimi and Bilgili 2007). Different corrections for small sample sizes have been proposed (Yates 1934, 1963; Emigh 1980; Elston and Forthofer 1977; Smith 1986); however, they do not always greatly improve the results from the traditional *asymptotic* tests (Emigh 1980).

IV. *Exact* (complete enumeration) *overall* tests of HWP

Levene (1949) described the conditional sampling distribution, which gives the exact probabilities for all possible samples of genotypes where the sample size (n) and the allele frequencies p_i, $i = 1, 2, ..., k$, are held constant, and z is the total number of heterozygotes observed in the sample.

Table IV.1

A_1	n_{11}			
A_2	n_{21}	n_{22}		
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
A_k	n_{k1}	n_{k2}	$\cdots$	n_{kk}
	A_1	A_2	$\cdots$	A_k

$$\Pr(\mathbf{f}) = \frac{n! \prod n_i!}{(2n)!} \cdot \frac{2^z}{\prod_{j \le i} n_{ij}!} \qquad \text{(eq. IV.1)}$$

An exact test for HWE based on Levene's conditional sampling distribution was developed by Louis and Dempster (1987). Their test utilized an algorithm for generating all possible tables of genotypes (based on observed allele counts) when the sample size and allele frequencies are held constant in accordance with the exact distribution. Louis and Dempster published individual algorithms for samples with two, three, and four distinct alleles (k =2, 3, and 4), and they indicated how the algorithms could be appropriately extended for samples with a higher number of distinct alleles. The p-value is given by the cumulative conditional probability of obtaining a table of genotypes (with sample size, number of alleles, and allele frequencies equal to the observed sample) with a conditional probability less than or equal to that of the genotypes in the observed sample (Levene 1949; Emigh 1980). This test provides the exact p-value for the observed sample and it does not require input parameters that may affect the result (c.f., the MCMC method, below). However, the number of possible tables of genotypes grows factorially as either the sample size or the number of distinct alleles (k) increases, reducing

the feasibility of this test when *n* and *k* are large.

Maldonado Torres (2009) recently developed a complete enumeration algorithm which efficiently generates all possible tables of genotypes for the exact distribution of Levene (1949) for a given number of distinct alleles (k) with allele frequencies (p_i) and sample size (n). This algorithm is used to perform a true *exact* test. In practice, however, its applicability on large samples and/or samples with a large number of distinct alleles is currently limited by the available computational resources, in particular the processor's speed. On a Pentium III class machine with 1 Gigabyte of system memory, a realistic upper limit was found to be around 100 individuals and 12 alleles. Approaches to circumvent this problem are being investigated and they are mainly related to the intelligent generation of the subset of needed samples from the set of all possible samples. An implementation of this exact test is available as a standalone program at http://www.cactus-project.org/hardy-weinberg/ and will soon be available in the Hardy-Weinberg module of PyPop.

V. *Approximation* (resampling) MC and MCMC *overall* tests of HWP

Approximation methods to complete enumeration of all possible samples of genotypes were developed for data sets with larger numbers of distinct alleles, where the *asymptotic* Chi-square test may be particularly problematic and *exact* tests were not possible. These methods generally use the Monte Carlo (MC) simulation method to approximate to the exact *p*-value and, therefore, represent an acceptable alternative to the *exact* test. Guo and Thompson (1992) developed the first conventional MC test of HWP based on Levene's conditional sampling distribution. Alternative tests to the conventional MC test have been proposed (Guo and Thompson 1992; Yuan and Bonney 2003; Huber et al. 2006). Guo and Thompson (1992) also proposed a Monte Carlo test that uses a finite and irreducible Markov Chain (MCMC) (Metropolis et al. 1953) to randomly generate the tables of genotypes. In each of these MC-based tests, the *p*-value is given by the fraction of randomly generated tables of genotypes with a conditional probability less than or equal to the conditional probability of the observed sample. These *approximation* tests are often erroneously referred to as "exact" tests; however, since they do not perform a true exhaustive search for all possible samples we refer to them as *approximation* or resampling tests.

With the exact test as the "gold standard" it was possible to compare not only *asymptotic* tests to resampling MC and MCMC tests as had been done previously (e.g., Emigh 1980; Louis and Dempster 1987; Guo and Thompson 1992; Chen and Thomson 1999; Chen et al. 2005; Wigginton et al. 2005), but also the resampling tests to the *exact* test results (within the scope of the size limitations imposed by the available computational resources). Both the MC and MCMC perform very favorably when compared to the *exact* test, and always outperform the *asymptotic* Chi-square test (Lancaster 2006, Maldonado-Torres et al. 2010). Either can be used in place of full enumeration when it is computationally more practical. Guo and Thompson (1992) reported that the MCMC algorithm is faster than the MC when the sample size is moderate or large; we have also found this to be the case (Lancaster 2006).

The MCMC method may fail to approximate to the exact *p*-value in a few cases (Lancaster 2006, Maldonado-Torres et al. 2010). In particular, some example tables of genotypes for which the constituent allele frequencies were skewed presented some difficulties for the MCMC method. A skewed allele frequency distribution may yield a low number of possible tables of genotypes. The effect of skewed allele frequencies may be exacerbated when combined

with a small sample size. In terms of the MCMC, a low number of possible tables of genotypes can result in the number of steps in the Markov chain (equal to the number of tables of genotypes randomly generated) being substantially larger than the number of all possible tables of genotypes. In certain cases, the latter makes the MCMC method prone to generating, or to sampling, a set of tables of genotypes too many times, affecting the approximation to the exact *p*-value.

Additionally, even though a skewed allele frequency distribution combined with a large sample size may yield a large number of possible tables, the number of observed distinct alleles and the presence of rare alleles can affect the performance of the MCMC. This is the result of the lack of possible switches (changes from one randomly generated table to the next) due to sparse cells (genotype counts of 0) that can cause the MCMC to get stuck in a state where it cannot move forward or where the choices are limited, therefore, affecting the approximation of the exact *p*-value.

Even though these scenarios may rarely happen in some genetic systems, they may be more common in others. Therefore, it appears more prudent in estimation of *overall* genotype-level *p*-values to use the MC algorithm in preference to the MCMC in all cases. The extra computational time required to use the MC over the MCMC is almost negligible with the speeds available in today's computational resources as compared to those available in 1992. Obviously, the *exact* test (complete enumeration) is the preferred method when computationally feasible.

VI. *Individual* genotype tests of HWP

A. Introduction

In addition to measuring *overall* genotype-level deviations from HWP, we are interested in identifying *individual* genotype-level deviations both for QC of the data, as well as for population and disease studies. The overall test may not always be significant even when individual genotype-level deviations are significant, and vice-versa. Note that the *individual* genotype testing approach is most useful if one has an a priori hypothesis about one or more genotypes in the sample, regarding their excess or deficit from HWP. Appropriate methods to statistically account for multiple comparisons are necessary if the methods below are applied to a large number of (or all) individual genotypes. By applying any of the methods discussed below to all *individual* genotypes, we face the problem of multiple tests, no matter whether we use *asymptotic*, *exact*, or *resampling* approaches. Given that genotype counts (O_i) are restricted to integer values,but the expected counts (E_i) can take small fractional values, conclusions made using these tests must be drawn carefully (see Section VIII below) to ensure that they reflect biological reality.

An incorrectly typed genotype may appear more or less often than expected by chance in population samples typed in a particular laboratory or using a particular protocol. For example, in the case of MSAT genotyping of heterozygous individuals, the shorter fragment size generally amplifies better than the larger fragment (in what is known as preferential amplification). In extreme cases, the longer allele may not be distinguished from background noise (resulting in allelic dropout), which leads to an over-representation of homozygotes for the shorter allele (Demers et al. 1995; Chen et al. 2005; Kalinowski 2006). Similarly, with null MSAT alleles (which are not detected as a result of variation under a PCR primer), there should be an excess of

all homozygotes and deficiency of all heterozygotes (Kalinowski 2006). Conversely, stuttering with MSAT typing can result in homozygotes being scored as heterozygotes. SNP data are also subject to typing-derived deviations from HWP, and SNP data showing lack of fit to HWP are routinely removed from analyses with no subsequent examination. For more details on genotyping errors and their causes see the review by Pompanon et al. (2005).

HLA genotyping depends on the hybridization of PCR primers, and in some cases of DNA probes, which can fail if novel sequence variants are present in the target sequence. When novel variants fall under primer targets, the allele in question can go unamplified, and therefore undetected, resulting in an excess of homozgous genotypes. When DNA probes are used, the identity of the alleles in a genotype is inferred from the pattern of probe hybridization; if a probe fails to hybridize as a result of novel sequence variation in one allele, one or both of the alleles in the genotype may be incorrectly identified. In cases where both alleles are incorrectly identified, the allele lacking novel sequence variation in that genotype will be correctly identified in other genotypes, so that novel sequences under DNA probes can result in an increased number of rare alleles, with the potential for a large number of sparse cells.

B. Asymptotic individual genotype tests of HWP

The relative magnitude of contributions to significant deviation from HWP for each *individual* genotype has been assessed by assuming that the contribution of each individual genotype to the chi-square statistic is a separate 1 d.f. statistic. However, this assumption is incorrect and the test is *always invalid* (and tends to be conservative) because the individual contributions are not independent. Additionally, as mentioned above, formal testing would require correction for multiple comparisons. However, quantiles from a Chi-square distribution with 1 *df* can provide insight as to the potential significance of the individual heterozygotes or homozygotes.

Chen and Thomson (1999) and Chen et al. (2005) developed an appropriate *asymptotic* test statistics for individual genotypes, heterozygous and homozygous cases respectively. They used a difference statistic (*O* - *E*) normalized by the correct variance of this individual genotype statistic under the null hypothesis of fit to HWP (see below). The corrected variance addresses a problem with the individual heterozygote genotype test of Hernandez and Weir (1989), which can generate negative variance values.

For an individual heterozygous genotype, the sample disequilibrium coefficient is: $d_{ik} = p_i p_k - \frac{1}{2} p_{ik}$, where $p_i,\ i = 1, 2, ..., m$ are the sample allele frequencies. Chen and Thomson (1999) showed that the correct variance for d_{ik} is given by

$$var(d_{ij}) = \frac{1}{2n} \; p_i p_j [(1 - p_i)(1 - p_i) + p_i p_j] + p_i^2 (p_{jj} - p_j^2) + p_j^2 (p_{ii} - p_i^2) \quad .$$

Thus, the single d.f. χ^2 test statistic, $\chi^2_{1d.f.} = \frac{d_{ik}^2}{var(d_{ik})}$, is used to test $H_0 : D_{ik} = 0$.

Similarly, for an individual homozygous genotype, $d_{ii} = p_i^2 - p_{ii}$ and Chen et al. (2005) showed that

$$var(d_{ii}) = \frac{1}{n} p_i^2 \; 1 - p_i \; ^2 .$$

In this case, $H_0 : D_{ik} = 0$ is tested using the test statistic $\chi^2_{1d.f.} = \frac{d_{ii}^2}{var(d_{ii})}$.

C. Exact and approximation individual genotype tests of HWP

Individual genotype-level tests have recently been added to the MC and MCMC tests in PyPop, and are also available in the exact test. We have demonstrated (Maldonado Torres 2009, Maldonado-Torres et al. 2010) that the individual genotype-level exact test can always be calculated via the reduction of the observed genotype distribution into a set of $k(k-1)/2$ tables of the appropriate 3x3 genotype sub distributions. Note that this variant of the *individual* exact test can *always* be applied even if the overall exact test cannot be performed due to computer limitations.

A *p*-value is calculated for each individual genotype by defining an appropriate test statistic for the difference between observed and expected genotype counts. As in the overall test, the fraction of tables of genotypes in which this test statistic is less than or equal to the test statistic of the observed table of genotypes is computed for each genotype. Two test statistics are reported by PyPop: (1) the "diff statistic" is based on the difference between observed and expected counts, while, (2) the "Chen statistic" normalizes this difference based on the correct variance under the null hypothesis and gives the statistic for heterozygotes from Chen and Thomson (1999) and that for homozygotes from Chen et al. (2005).

VII. *Overall* heterozygote excess/deficiency, and *all* heterozygotes for a specific allele

In addition to individual genotype tests, tests of *all* homozygotes and *all* heterozygotes can be very useful for detecting genotyping errors during the QC of data (Kalinowski 2006). If for example null MSAT alleles are present in a sample, an MSAT allelic dropout has occurred, or a specific HLA allele was not amplified, there should be an excess of all homozygotes and corresponding deficiency of all heterozygotes; the opposite pattern would be observed in cases of MSAT stuttering, for example. The HW-QUICKCHECK program of Kalinowski (2006) performs tests of deviation from HWP using the MC approximation method for *overall* genotype level, *individual* genotype, and *overall* homozygosity and heterozygosity. These tests are assumed to be hypothesis-driven and hence report one-sided *p*-values.

It is of additional interest to consider all heterozygotes for a particular (common) allele (e.g. A_1A_X, where X denotes the combined set of all alleles excluding A_1) both for QC of data, and for the detection of selection or differential relative disease risk. This test, using the *k* subsets of the appropriate 2x2 tables, along with an *overall* test of heterozygote excess/deficiency, will be incorporated in an upcoming version of PyPop.

VIII. *Rare* alleles and genotype classes

It is not uncommon for HLA datasets to include several alleles with counts of 1 or 2 (and other such small numbers) and therefore many genotypes will have counts of 0 in these cases. Further, expected values for genotypes with these alleles will all be very small. Should rare alleles all be

combined into one class for both *overall* genotype-level and *individual* genotype-level analyses? Another alternative would be to remove from analysis all individuals with a rare allele.

Why do we raise this issue? The pertinent problem with *individual* genotype-level analyses is that while all allele counts are integers, the expected counts for specific genotypes containing rare alleles may be very much less than 1; this is especially true for highly polymorphic genes, like the HLA genes. However, biological reality dictates that observed genotype counts must be integers, so that statistical tests carried out in cases with low expected genotype counts are rendered meaningless, leading to spurious results and erroneous interpretations. While a strict theoretical guideline is not feasible, we devote attention to this issue, with some suggestions for dealing with rare alleles.

A fictitious example that illustrates problems in testing individual genotype deviations from HWP is as follows: in a sample of size $n = 100$ individuals, suppose there are 26 distinct alleles observed,

24 of which have a count of 8 (alleles A_1 - A_{24}, each with $p_i = 0.04$),

1 of which has a count of 7 (allele A_{25}, with $p_{25} = 0.035$), and

1 of which has a count of 1 (allele A_{26}, with $p_{26} = 0.005$).

The allele A_{26} must occur in a genotype with 1, and only 1, of the other 25 alleles, and under random mating will do so with probabilities 0.04/0.995 (= 0.0402) for each of alleles A_1 - A_{24}, and probability 0.035/0.995 (= 0.0352) for allele A_{25} (with correction to allow for the fact that no homozygous A_{26} individuals can occur). Thus, with a nominal *p*-value of 0.05, every observation is significantly different from random expectation (observed 1 and expected 0.0402 or 0.0352) *Biologically of course this makes no sense.*

Similarly, even if all alleles are not this rare, there may be problems. As another example consider a sample of size $n = 100$ individuals, with 13 distinct alleles observed,

1 of which has a count of 100 (allele A_1, with $p_1 = 0.5$),

1 of which has a count of 40 (allele A_2, with $p_2 = 0.2$),

9 of which have a count of 6 (alleles A_3 - A_{11}, each with $p_i = 0.03$),

1 of which has a count of 5 (allele A_{12}, with $p_{12} = 0.025$), and

1 of which has a count of 1 (allele A_{13}, with $p_{13} = 0.005$).

Again, as above, allele A_{13} must occur with 1, and only 1, of the other alleles. If it occurs with one of the 10 alleles A_3 - A_{12} it will show up as individually significant from random expectations, as above. However, this subset of alleles accounts for 30.2% of the possible genotypes that A_{13} can occur with.

Obviously this problem is not restricted to singleton alleles and heterozygotes thereof. There are many possible examples when the expected values are sufficiently small that the statistical test *may not* reflect biological reality. We have proposed an *arbitrary* cut-off (subject to change after further research) that *p*-values for individual genotypes with an expected genotype count under HWP that is <2 will be indicated with an asterisk, indicating that one must be very careful in not over interpreting the result. Given that researchers have noted cases with low expected but quite a large observed genotype count, and that this may serve as a signal of admixture, researchers may wish to note such observations.

Further research is required to determine whether very rare alleles should be combined as a lumped category in overall genotype-level tests of HWP. While such combining of rare classes is required for the Chi-square test, it is not required for the *resampling* MC and MCMC methods and the *exact* complete-enumeration test, which made these more appealing. However, there is insufficient understanding at this time of the effect on power of lumping, or not lumping, rare alleles. Also, the biological interpretation of lumped alleles is less straightforward. An alternative also is to delete genotypes containing rare alleles from the analysis. We suggest that the data be analyzed, both for *overall* and *individual* genotype tests, using both the original data as well as data in which all rare alleles have been removed or combined (at least singletons, but possibly alleles with counts < 3. Also, one can perform a hierarchical set of analyses, first testing genotypes of the two, or a few, most common alleles, and adding alleles.

IX. Application to population and patient data sets

A. Introduction

Many lines of evidence indicate that HLA variation is shaped by natural selection and that some form of balancing selection is operating on the classical HLA genes, except for DPB1 (which nonetheless may still be selected). Some studies have observed deviations from HWP that appear to be due to selection on HLA genes, ruling out typing errors, non-random mating etc. (reviewed in Meyer and Thomson 2000, also see Black and Salzano 1981; Markow et al. 1993; Chen et al. 1999). One trend observed in HLA data is that isolated populations tend to show more deviation from HWP. Selection explanations are possible, but also Hedrick (1990) has pointed out that the lower genetic diversity of isolated populations makes tests of HWP more powerful.

Nonetheless, genotyping errors are always the first possibility to consider in quality control (QC) of HLA, and other, data. For HLA genotyping, which often involves a series of probes and primers being applied in sequence to progressively resolve an allele, it is possible that, for example, a probe which is critical in resolving an allele may be omitted, or not functioning as expected, leading to an incorrect allele assignment. Bugawan et al. (1999) observed significant excess over HWP for the allele A*3401 in populations from Papua New Guinea. These deviations resulted from a single base-pair mismatch between a PCR primer and the A*3401 allele, so that the A*3401 allele was not detected in these populations. After previously reported homozygous samples in these populations were re-typed using a second method to avoid this issue, the data were then in HWP.

B. Microsatellite data

Microsatellite polymorphisms are commonly used markers in genetic analyses for both disease and population genetic studies. While microsatellites provide an abundant and cost effective source of genetic markers, there are issues regarding microsatellite typing that need to be considered in downstream analyses. One such issue is the preferential amplification of shorter fragment sizes compared to larger sizes. When the difference in amplification is large it can be difficult to distinguish the longer allele from background noise resulting in an overrepresentation of homozygotes for the shorter allele(s) (Demers et al. 1995). This leads to a condition referred to as extreme preferential amplification (EPA). Another related issue in microsatellite genotyping is that of allele dropout (Rodriguez et al. 2001). In this situation a specific allele does

not amplify (unrelated to allele length) possibly due to low quality or low concentration template DNA, or to variation in the sequence where PCR primers anneal. As with EPA, allele dropout can result in an artificial increase of specific homozygous genotypes. Overall deviation from HWP, followed by an examination of patterns of deviations for individual homozygotes and heterozygotes can provide insight into the possibility of EPA and/or allele dropout. When detected, these issues can often be resolved by retyping with a modified PCR process.

Chen et al. (2005) reported on an example of EPA with the MogCA microsatellite in a sample of 47 unrelated CEPH individuals. We give an additional example here from a sample of 48 Mixe individuals (Hollenback et al. 2001). The Mixe, a native American group from Mexico, were sampled from the mountains east of the Oaxacan Valley. Seven different MogCA alleles were found in this sample. A significant overall test of HWP along with past experience with this marker suggested possible preferential amplification problems. A modified PCR protocol, designed to compensate for preferential amplification, was used to retype this sample. The table below provides additional detail regarding the pattern of deviation from HWP for individual homozygotes (before retyping) that led to identification and resolution of problems from EPA.

Table IX.1: HWP testing of individual homozygotes to identify EPA problems with microsatellite typing

Mixe data set

	Before Retyping				After Retyping			
Homozygote	obs.	exp.	Chi-sq	p-value	obs.	exp.	Chi-sq	p-value
122/122	3	0.33	25.388	<0.001	0	0.13	0.145	0.71
132/132	34	27.76	24.478	<0.001	17	16.92	0.002	0.97
136/136	0	0.02	0.022	0.89	0	0.02	0.022	0.89
146/146	0	0.02	0.022	0.89	0	0.08	0.091	0.77
148/148	0	0.05	0.050	0.83	0	0.63	0.804	0.37
150/150	1	0.26	2.529	0.12	1	1.33	0.120	0.73
152/152	0	0.01	0.005	0.95	0	0.01	0.005	0.95

C. Admixed data

With admixed data, when specific alleles (or haplotypes) can be identified as coming from a different population, removal of these individuals from tests of HWP may give a closer fit to HWP, e.g., see, Hollenbach et al. (2001). Bone marrow registry data is one source of HLA data where there is typically a good deal of admixture, especially in the full registry dataset. Large significant deviations from HWP are often seen when working with the entire registry dataset, but less significance is found when subsets of the data, based on self identified race and ethnicity (SIRE) codes, are analyzed separately. While differences in sample size, and thus power to

detect deviations from HWP, are certainly at play here, the difference in frequencies for various sets of alleles among the SIRE groups is a major contributing factor.

D. Type 1 diabetes and HLA DR-DQ

In disease studies, except for a strictly recessive model, deviation from HWP in a patient population may be observed for a causative genetic variant, as well as markers in LD with the causative variant (Thomson 1993, 1995a, b, c, Zou and Donner 2006). Models of disease are in fact equivalent to selection models: the general population represents the before-selection genetic variation, the patient population that after selection (disease in this case) (Valdes and Thomson 1997). Thus, if the control population is in HWP, a genetic marker should *not be removed* from analyses due solely to lack of fit to HWP in a patient population.

Further, the form of deviations from HWP can indicate specific genotype effects on disease risk. With type 1 diabetes (T1D), in Caucasian populations, the two most frequent haplotypes associated with increased predisposition to disease are DRB1*0301 DQA1*0501 DQB1*0201 and DRB1*0401 DQA1*0301 DQB1*0302. Abbreviating these as DR3 and DR4, the observed genotype counts in patients and the HWP are given in Table IX.1 (data from Erlich et al. 2008). The excess of heterozygote DR3/DR4's, reflecting a higher disease risk than for the respective homozygote combinations, is highly significant (Chi-sq. = 24.6, p < 000001) and consistently seen.

Table IX.2: HLA DR-DQ genotype counts in type 1 diabetes for predisposing DR-DQ haplotypes*

	DR3/DR3	DR3/DR4	DR4/DR4
Observed genotype counts:	46	157	35
Expected HWP:	65.1	118.8	54.1
(O - E)	-19.1	38.2	-19.1
$(O - E)^2/E$	5.6	12.3	6.7

Total chi-square: 24.6, $df = 1$, $p < 0.000001$

* Data are from the study of Erlich et al. (2008): DR3 denotes the DRB1*0301 DQA1*0501 DQB1*0201 haplotype, and DR4 the DRB1*0401 DQA1*0301 DQB1*0302 haplotype; these are the 2 most common predisposing DR-DQ haplotypes in the study.

E. HIV-1 and CCR5 Δ32

The *CCR5* gene is a co-receptor for HIV-1, the virus predisposing to AIDS. A variant, *CCR5 Δ32*, with a 32 base pair deletion resulting in a frame shift and premature stop codon, has a truncated protein product that is not expressed on the cell surface. Individuals homozygous for *CCR5 Δ*32 have nearly complete protection against HIV-1 infection, despite repeated exposures (see Martin and Carrington 2002 for review). This gives highly significant deviations from HWP, see example (1) of Table IX.2 below. Conversely, in individuals with repeated exposures

who are not infected, there is a significant excess of individuals homozygous for *CCR5 Δ*32, see example (2) of Table IX.2. In both cases, most of the contribution to the chi-square statistic comes from the homozygous *Δ32/Δ32* classes. The deviation is in opposite directions in these 2 samples; and the test of heterogeneity of the 3 genotypes in the 2 samples is highly significant ($p < 2.7E\text{-}16$) (data not shown). Compare these data to those for a random population sample, which shows fit to HWP (example (2) of Table II.1).

Table IX.3: CCR5 *Δ32* variant and deviations from HW in HIV+ and HIV- samples*

(1) CCR5 *Δ32* variant in HIV+ sample: $p_1 = 0.909$, $p_2 = 0.091$

Genotypes:	*A/A*	*A/Δ32*	*Δ32/Δ32*	
Observed numbers	1988	440	1	Total: 2,429
Expected HWP	2007.0	401.9	20.1	Total: 2,429
$(O - E)$	-19.0	38.1	-19.1	
$(O-E)^2/E$	0.18	3.61	18.15	

Total Chi-square = 21.94, $df = 1$, $p < 0.000003$

(2) CCR5 *Δ32* variant in HIV- sample: $p_1 = 0.8835$, $p_2 = 0.1165$

Genotypes:	*A/A*	*A/Δ32*	*Δ32/Δ32*	
Observed numbers	793	174	29	Total: 996
Expected HWP	777.5	205.0	13.5	Total: 996
$(O - E)$	15.5	-31.0	15.5	
$(O-E)^2/E$	0.31	4.68	17.80	

Total Chi-square = 22.79, $df = 1$, $p < 0.000002$

*Data are from Martin and Carrington (2002)

X. Discussion

We have not discussed the literature on using different test statistics for evaluating fit or not to HWP (see Rousset and Raymond 1995), especially those that take into account the nature of the alternative hypothesis. For example, population structure and selfing can result in heterozygote deficiency, and other tests may have more power to detect such deviations than those we have discussed. Nonetheless, with alternative test statistics, the conditional distribution under the null hypothesis can still be computed using Levene's distribution (see eq. IV.1) (Levene 1949). While different test statistics may define different rankings of all possible genotype tables, the appropriate p-value in each test is still defined as the sum of probabilities of samples with more extreme ranks. Hence, our discussions of exact (complete enumeration) and approximation

(resampling) MC and MCMC tests of HWP are relevant to other test statistics, with a simple modification of the test statistic to be considered.

As mentioned in Section IV, further research will focus on increasing the range of sample size (n) and number of alleles (k) for which the *overall* genotype level *exact* test can be carried out. Strategies for treating rare genotype (and allele) counts requires further research. Notwithstanding, we caution interpretation of *individual* genotype p-values when the expected value is ≤ 2. Future research will also focus on determining the exact HWE for a given sample, i.e., finding the table(s) of possible genotype counts of integer values, whose conditional probability is greater than or equal to all the other possible tables of genotype values, and therefore whose p-value is 1. In this way we could compare observed values to the true HWE for that sample, rather than to expected values based on non-integer values. This approach should be particularly useful for data with rare expected genotype counts.

Acknowledgements: This research was supported by NIH contracts and grants: AI40076 (GT, RMS, LFB) and AI67068 (SJM, JAH).

References:

Black FL, Salzano FM. 1981. Evidence for heterosis in the HLA system. American Journal of Human Genetics 33: 894-899.

Bugawan TL, Mack SJ, Stoneking M, Saha M, Beck HP, Erlich HA. 1999. HLA class I allele distributions in six Pacific/Asian populations: evidence of selection at the HLA-A locus. Tissue Antigens 53: 311-319.

Bourgain C, Abney M, Schneider D, Ober C, McPeek MS. 2004. Testing for Hardy-Weinberg equilibrium in samples with related individuals. Genetics 168: 2349-2361.

Chen JJ, Thomson G. 1999. The variance for the disequilibrium coefficient in the individual Hardy-Weinberg test. Biometrics 55: 1269-1272.

Chen JJ, Duan T, Single RM, Mather KA, Thomson G. 2005. Hardy-Weinberg testing of a single homozygous genotype. Genetics 170: 1439-1442.

Chen JJ, Hollenbach JA, Trachtenberg EA, Just JJ, Carrington M, Rønningen KS, Begovich A, King MC, McWeeney SK, Mack SJ, Erlich HA, Thomson G. 1999. Hardy-Weinberg testing for HLA class II (DRB1, DQA1, DQB1 and DPB1) loci in 26 human ethnic groups. Tissue Antigens 54: 533-542.

Cochran W. 1954. Some methods for strengthening the common chi-squared test. Biometrics 10: 417-451.

Demers, D. B., E. T. Curry, M. Egholm and A. C. Sozer. 1995. Enhanced PCR amplification VNTR locus D1S80 using peptide nucleic acid (PNA). Nucleic Acids Research 23(15): 3050-5.

Ebrahimi N, Bilgili D. 2007. A new method of testing for Hardy-Weinberg equilibrium and ordering populations. Journal of Genetics 86: 1-7.

Elston RC, Forthofer R. 1977. Testing for Hardy-Weinberg equilibrium in small samples. Biometrics 33: 536-542.

Emigh T. 1980. A comparison of tests for hardy-Weinberg equilibrium. Biometrics 36: 627-642.

Erlich H, Valdes AM, Noble J, Carlson JA, Varney M, Concannon P, Mychaleckyj JC, Todd JA, Bonella P, Fear AL, Lavant E, Louey A, Moonsamy P, T1DGC. 2008. HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetic consortium families. Diabetes 57: 1084-1092.

Freeman MF, Tukey JW. 1950. Transformations related to the angular and the square root. The Annals of mathematical Statistics 21: 607-611.Guo SW, Thompson EA. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48(2): 361-72.

Hardy GH. Mendelian proportions in a mixed population. 1908. Science 28:49-50.

Hedrick PW. 2003. A heterozygote advantage. Science 302: 57.

Hedrick PW. 1990. Evolution at HLA: possible explanations for the deficiency of homozygotes in two populations . Human Heredity 40: 213-220.

Hernández JL, Weir BS. 1989. A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics 45(1): 53-70.

Hollenbach JA, Thomson G, Cao K, Fernandez-Viña M, Erlich HA, Bugawan TL, Winkler C, Winter M, Klitz W. 2001. HLA diversity, differentiation, and haplotype evolution in Mesoamerican natives. Human Immunology 62: 378-390.

Huber M, Chen Y, Dinwoodie I, Dobra A, Nicholas M. 2006. Monte Carlo Algorithms for Hardy-Weinberg Proportions. Biometrics 62, 49–53.

Kalinowski ST. 2006. HW-QUICKCHECK: an easy-to-use computer program for checking for agreement with Hardy-Weinberg expectations. Molecular Ecology Notes 6: 974-979.

Lancaster A. 2006. Interplay of selection and molecular function in HLA genes. PhD thesis, University of California at Berkeley.

Lancaster A, Nelson MP, Single RM, Meyer D, Thomson G. 2003. PyPop: a software framework for population genomics: analyzing large-scale multi-locus genotype data. Pac Symp Biocomput 2003: 514-525.

Lancaster A, Nelson MP, Single RM, Meyer D, Thomson G. 2007a. Software framework for the Biostatistics Core of the International Histocompatibility Working Group. In: Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress, Volume I, ed. Hansen JA. IHWG Press, Seattle, WA, pp. 510-517.

Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. 2007b. PyPop update - a software pipeline for large-scale multi-locus population genomics. Tissue Antigens 69 (Suppl. 1): 192-197.

Levene H. 1949. On a matching problem arising in genetics. Annals of Mathematical Statistics 20: 91-94.

Li CC. 1955. Population Genetics. University of Chicago Press, Chicago.

Louis EJ, Dempster ER. 1987. An exact test for Hardy-Weinberg and multiple alleles. Biometrics 43(4): 805-11.

Maldonado Torres H. 2009. HLA genetic diversity in the UK population -- population genetics tools and methods. Ph.D. thesis, University College London.

Maldonado-Torres H, Lancaster AK, Single RM, Madrigal AJ, Thomson G, Marsh SGE. 2009. A complete enumeration algorithm for Hardy-Weinberg equilibrium: overall and individual genotype-level exact tests, in preparation.

Mantel N, Li CC. 1974. Estimation and testing of a measure of nonrandom mating. Annals of Human Genetics 37: 445-454.

Martin M, Carrington M. 2002. The role of human genetics in HIV-1 Infection. In O'Brien T (Ed): Chemokine Receptors and AIDS. Marcel Dekker, Inc., NY, pp. 133-162.

Markow T, Hedrick PW, Zuerlein K, Danilovs J, Martin J, Vyvial T, Armstrong C. 1993. HLA polymorphism in the Havasupai: evidence for balancing selection. American Journal of Human Genetics 53: 943-952.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of State Calculations by Fast Computing Machines. 1953. Journal of Chemical Physics 21:1087-1092.

Meyer D, Thomson G. 2001. How selection shapes variation of the human major histocompatibility complex: a review. Annals of Human Genetics 65: 1-26.

Pompanon F, Bonin A, Bellemain E, Taberlet P. 2005. Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics 6: 847-859.

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

Smith CAB. 1986. Chi-squared tests with small numbers. Annals of Human Genetics 50: 163-167.

Teo YY, Fry AE, Clark TG, Tai ES, Seielstad M. 2007. On the usage of HWE for identifying genotyping errors. Annals of Human genetics 71: 701-703.

Thomson G. 1993. The AGFAP method: applicability under different ascertainment schemes and a parental contributions test. Genetic Epidemiology 10: 289-310.

Thomson G. 1995a. HLA disease associations: models for the study of complex human genetic disorders. Critical Reviews in Clinical Laboratory Science 32: 183-219.

Thomson G. 1995b. Analysis of complex human genetic traits: an ordered-notation method and new tests for mode of inheritance. American Journal of Human Genetics 57: 474-486.

Thomson G. 1995c. Mapping disease genes: family based association studies. American Journal of Human Genetics 57: 487-498.

Valdes AM, Thomson G. 1997. Detecting disease-predisposing variants: the haplotype method. American Journal of Human Genetics 60: 703-716.

Weinberg W. 1908. Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64:368-382.

Weinberg W (translated by SH Boyer). 1963. On the demonstration of heredity on man, 1908, In: Papers on Human Genetics, Prentice-Hall, Englewood Cliffs.

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. American Journal of Human Genetics 76(5): 887-893.

Yates F. 1934. The analysis of multiple classifications with unequal numbers in the different classes. Journal of the American Statistical Association 29: 51-66.

Yates F. 1963. Contingency tables involving small numbers and the chi-squared test. Journal of the Royal Statistical Society (Supplement) 1: 217-235.

Yuan A, Bonney GE. 2003. Exact test of Hardy-Weinberg equilibrium by Markov chain Monte Carlo. Mathematical Medicine and Biology - A Journal of the IMA 20: 327-340.

Zou GY, Donner A. 2006. The merits of testing Hardy Weinberg equilibrium in the analysis of unmatched case-control data: a cautionary note. American Journal of Human Genetics 70: 923-933.

Appendix A: Derivation of Hardy-Weinberg proportions (HWP)

A. History

Mendelian genetics was rediscovered in the early 1900's. The founding principle of population genetics—the Hardy-Weinberg law—was derived independently in 1908 by Hardy, an English mathematician, and Weinberg, a German physician (Hardy 1908, Weinberg 1908, and for a translation see Weinberg 1963). The geneticist Punnett brought to the attention of Hardy a remark of Yule (also a geneticist), which criticized the concept of Mendelian inheritance by suggesting that if brachydactyly (short fingeredness) was dominant "in the course of time one would expect, in the absence of counteracting factors, to get 3 brachydactylous individuals to 1 normal." Of course we now understand that Yule was confusing the Mendelian 3:1 ratio in an F1 cross (*Aa* x *Aa* gives 3*A*-:1*aa*) with population features. (The notation *A*- is used with dominant traits to denote the dominant phenotype which includes *AA* and *Aa* genotypes.) Thus the Hardy-Weinberg law describes the application of Mendelian principles at the *population*, as opposed to the individual, level.

B. Estimating allele frequencies: the method of gene (allele) counting

Before proceeding with derivation of the HWP, we need to show how allele frequencies are estimated in the specific case where we can test fit to HWP, i.e., when all heterozygotes can be distinguished from the homozygous states (e.g., with incomplete dominance or codominance and in most situations with molecular typing). Allele frequencies in this case are obtained by the method of gene (allele) counting. Strictly speaking this should be referred to as allele counting rather than gene counting; the latter term however will be used for consistency with most of the literature.

In a two allele system with alleles denoted A_1 and A_2, with respective allele frequencies in the sample under study of p_1 and p_2, the maximum likelihood estimates (MLEs) of these allele frequencies are simply those found by counting the numbers of alleles, as would be expected:

$$p_1 = f(A_1) = f(A_1A_1) + f(A_1A_2)/2, \qquad \text{(eq. A.1)}$$

$$p_2 = f(A_2) = f(A_1A_2)/2 + f(A_2A_2), \text{ with } p_1 + p_2 = 1,$$

where $f(A_1A_1)$ is the frequency of homozygous A_1A_1 individuals, $f(A_1A_2)$ of heterozygous A_1A_2 individuals, etc. Note that no assumption of HWP is required. This is the method of gene counting.

Table A.1: Example genotype counts and frequencies

Genotypes	A_1A_1	A_1A_2	A_2A_2	
Observed numbers	50	40	10	Total: 100
Frequencies:	0.50	0.40	0.10	Total: 1.00

Consider the example in Table III.1. The allele frequencies in this case are:

$p_1 = 0.5 + (0.4)/2 = 0.7$, $p_2 = (0.4)/2 + 0.10 = 0.3$ (check $p_1 + p_2 = 1$).

Of course these calculations can also be done using the observed *numbers* of genotypes:

A_1A_1	A_1A_2	A_2A_2	
n_{11}	n_{12}	n_{22}	Total: n

$p_1 = (2n_{11} + n_{12})/2n$, and $p_2 = (n_{12} + 2\,n_{22})/2n$. (eq. A.2)

C. The variance of the allele frequency estimates

Denoting by p_{A1} the true frequency of allele A_1 at the population level, the variance of the allele frequency estimated by the method of gene counting is $p_{A1}\,(1 - p_{A1}) / (2n)$ where n is the sample size. An estimate of this variance is obtained using this formula and the estimated allele frequency p_1:

$p_1\,(1 - p_1) / (2n)$. (eq. A.3)

The standard deviation (*sd*) is the square root of the variance. The 95% confidence interval for the range of the true allele frequency p_{A1} is 1.96 times the *sd* on either side of the estimate. The larger the sample size n, the smaller the variance and *sd*, and theoretically the more accurate the estimate of the allele frequency, as one would expect.

As above, where ~2 *sd* on each side of the estimated allele frequency represents the approximate 95% confidence interval for the true allele frequency, ~1 *sd* on both sides of the estimated allele frequency is the 68% confidence interval, while ~3 *sd* corresponds to the 99.7% confidence interval.

D. Derivation of Hardy-Weinberg proportions (HWP)

We consider a parental population with two alleles at the locus of interest, with

$p_1 = f(A_1)$, $p_2 = f(A_2)$, with $p_1 + p_2 = 1$.

Under the assumptions listed above in Section II.C, to determine the genotype frequencies in the offspring we consider the 2x2 table of female and male gametes (assuming equal allele frequencies in both sexes), see Table A.1 below, and random union of these gametes to form the genotypes in the offspring generation. (We should consider random union of individuals, rather than random union of gametes, but in fact the result is the same.)

Table A.2: Random union of male and female gametes resulting in HWP

female gametes:	$A_1\,(p_1)$	$A_2\,(p_2)$
male gametes: $A_1\,(p_1)$ /	$A_1A_1\,(p_1^2)$	$A_1A_2\,(p_1p_2)$
$A_2\,(p_2)$ /	$A_2A_1\,(p_2p_1)$	$A_2A_2\,(p_2^2)$

The genotype frequencies in the offspring are the HWP:

$A_1A_1 - p_1^2$, $A_1A_2 - 2p_1p_2$, $A_2A_2 - p_2^2$.

The factor 2 arises in the term for heterozygotes as either the father can supply the A_1 allele and the mother the A_2 allele, or vice versa. Note here that with the assumption of an *infinite*

population, the genotype frequencies are *exactly* determined by the products of the allele frequencies.

Also, in particular, note that in the offspring generation, by gene (allele) counting:

$f(A_1) = p_1$, $f(A_2) = p_2$.

Thus, allele frequencies *do not change* if the conditions listed in Section II.C are met, and genotype frequencies (proportions) can be *predicted* from allele frequencies. As stated above, to conduct a test of HWP, we must be able to distinguish all heterozygote and homozygote individuals.

E. Evolutionary implications of HW

The importance of the Hardy-Weinberg law from the point of view of population genetics and evolution is that if there are no counteracting forces, then the allele frequencies do not change in a population. In other words, under a Mendelian system, the genetic variation in a population tends to be maintained. This is in contrast to the concept of 'blending inheritance' (a common belief at the time of Darwin), whereby genetic variation decreases each new generation.

Since all populations are in fact finite there is an inevitable change in the allele frequencies, in both the short-term and long-term evolution of populations, but from the point of view of testing for HWP this can be ignored.

F. Two extreme examples showing fit and lack of fit to HWP

In a test of HWP, allele frequencies are estimated and observed genotype *counts* are compared to those expected under HWP. The following two examples (Table AI.3) have the same sample size ($n = 100$) and the same allele frequencies ($p_1 = 0.6$, and $p_2 = 0.4$), and hence the *same HWP*; *but* the observed genotype frequencies in the two examples are very different, and respectively illustrate fit to HWP (example 1) and lack of fit to HWP (example 2). (In this case the examples are extreme and the results obvious; the actual statistical test is given below in Table B.1.)

Table A.3: Two examples with the same allele frequencies showing close fit to HWP and large deviation from HWP

(1) Example with close fit to HWP: $p_1 = 0.6$, $p_2 = 0.4$

Genotypes:	C_1C_1	C_1C_2	C_2C_2
Observed: counts (frequencies)	35 (0.35)	50 (0.50)	15 (0.15)
Expected HWP: counts (frequencies)	36 (0.36)	48 (0.48)	16 (0.16)

(2) Example with large deviation from HWP: $p_1 = 0.6$, $p_2 = 0.4$

Genotypes:	E_1E_1	E_1E_2	E_2E_2
Observed: counts (frequencies)	56 (0.56)	8 (0.08)	36 (0.36)
Expected HWP: counts (frequencies)	36 (0.36)	48 (0.48)	16 (0.16)

G. Hardy-Weinberg equilibrium for an X-linked trait

Females have two X (sex) chromosomes, while males have one X and one Y chromosome. The Y chromosome carries very few genes, so males are effectively haploid for X-linked traits. The Hardy-Weinberg expectations for an X-linked trait are the same as for an autosomal locus for females, while for males they are simply the allele frequencies. For example, for a locus with two alleles denoted A_1 and A_2 with allele frequencies p_1 and p_2 ($p_1 + p_2 = 1$), as above, the genotypes and their frequencies are:

females: A_1A_1 - p_1^2, A_1A_2 - $2p_1p_2$, A_2A_2 - p_2^2, males: A_1 - p_1, A_2 - p_2.

If the male and female allele frequencies are not equal, HWP in the females will be reached asymptotically. Note that no test of HWP can be performed in the males.

H. Estimation of allele frequency and carrier frequency for a recessive trait

Many traits of interest are neither codominant nor incompletely dominant, but we may be interested in estimating the allele frequencies in this case, as well as the proportion of individuals heterozygous for a recessive trait (called carriers). To do this, we often *assume* HWP in the population. Here we will use the standard notation for the two-allele case of denoting the alleles by *A* and *a*, and the genotypes by *AA*, *Aa*, and *aa*; with capital and lower case letters used to denote dominant versus recessive alleles respectively. We denote the allele frequencies by *p* and *q* ($p + q = 1$), and the HWP are AA - p^2, Aa - $2pq$, aa - q^2.

In this case we cannot distinguish the heterozygotes (*Aa*) from the homozygotes (*AA*) and hence cannot estimate the allele frequencies by gene counting. Instead, using the fact that the expected frequency of homozygotes for the recessive trait under HWP is $f(aa) = q^2$, we *estimate* the allele frequency of the recessive trait by

$$q = f(a) = [f(aa)]^{1/2}. \qquad \text{(eq. A.4)}$$

Note that we cannot test whether the population is in HWP since we have already assumed that it is so in order to obtain our estimate of the allele frequencies.

We can also use the assumption of HWP to estimate how many *carriers* for the trait (heterozygotes) there are in the population, i.e., we estimate $2pq$, using our estimate of q (and hence $p = 1 - q$), above. For *rare* recessive autosomal traits, the majority of the disease predisposing *a* alleles are found in heterozygous unaffected carriers (frequency ~$2q$, compared to a frequency of affected individuals of q^2. See examples (2) and (3) in Table II.1 (note that the heterozygote frequency was known from molecular typing in these examples). Further, most affected individuals result from heterozygote by heterozygote matings. Examples of carrier frequencies in three rare autosomal recessive traits are given for illustration**.**

For X-linked recessive traits there are always more affected males (q) than females (q^2); if the trait is rare, e.g., hemophilia, the vast majority of affected individuals are males. Likewise, for X-linked dominant traits there are always more affected females ($1-q^2$) than males (p); for rare traits, there are approximately twice as many affected females ($2p$) as males (p).

Cystic fibrosis: Characterized by malfunction of the pancreas and other glands, cystic fibrosis is one of the most frequent recessive diseases among Europeans and people of European descent. The incidence of this condition is about 1/2,500 individuals, i.e., $q^2 = 0.0004$, giving an estimate of $q = 0.02$, with a corresponding estimate for carriers of ~4%.

Sickle cell anemia: Sickle cell anemia is the most common recessive disease among African Americans, with an incidence of approximately 1/400 individuals. In this case, $q^2 = 0.0025$, the estimate of $q = 0.05$, and the estimate of carriers is ~9.5% (heterozygous carriers of the disease show no ill effects except in conditions of oxygen stress). In parts of West Africa about 1/100 individuals have sickle cell anemia, and ~18% are carriers. The high frequency of the allele is due to an advantage to heterozygous individuals in malarial environments.

Tay Sachs: Tay Sachs disease is a tragic recessive illness that results from the absence of the hexosaminidase A enzyme and leads to death by the age of 3 or 4 years. Among descendants of Ashkenazi Jews who settled in Eastern and Central Europe this disease occurs with an incidence of ~ 1/4,000; therefore $q^2 = 0.00025$, the estimate of $q = 0.016$, and the estimate of heterozygous carriers is ~3.2%. Carrier status screening is possible. If parents are both carriers then prenatal diagnosis can be performed. Among the non-Jewish population of North America, the incidence of Tay Sachs is about 1 in 500,000 births, with $q = 0.0014$.

Appendix B: The Chi-square test of HWP

A. Calculating the Chi-square test statistic

Two examples are given in Table B.1 of the calculation of $(O\text{-}E)^2/E$, where O is the observed genotype count and E is the expected count under HWP. For the moment we assume that $E \geq 5$ for all genotypes. Example (1) has genotype frequencies close to HWP, while example (2) has genotype frequencies with a large deviation from HWP. In both cases the $df = 1$, and the respective p-values are *ns* (not significant at 5%) and $p < 0.000001$ (discussed below).

Table B.1: Two examples of calculating the Chi-square test statistic for overall fit to HWP*

(1) Example with close fit to HWP: $p_1 = 0.6$, $p_2 = 0.4$

Genotypes:	C_1C_1	C_1C_2	C_2C_2	
Observed counts (O)	35	50	15	Total: 100
Expected HWP counts (E)	36	48	16	Total: 100
(O - E)	-1	2	-1	
$(O\text{-}E)^2/E$	0.028	0.083	0.063	Total Chi-square: 0.174

(2) Example with large deviation from HWP: $p_1 = 0.6$, $p_2 = 0.4$

Genotypes:	C_1C_1	C_1C_2	C_2C_2	
Observed counts (O)	56	8	36	Total: 100
Expected HWP counts (E)	36	48	16	Total: 100
(O - E)	20	-40	20	
$(O\text{-}E)^2/E$	11.1	33.3	25.0	Total Chi-square: 69.4

* These data are the same as in Table A.3.

B. Accepting or rejecting the null hypothesis

The overall Chi-square test statistic value is compared to theoretical Chi-square values based on *df* and the type 1 probability *p*-values under the null hypothesis (see Table B.2, and for a more detailed Table see http://www.math.unb.ca/~knight/utility/chitable.html). If the test statistic value is less than that given under the heading Probability (*p*-value) of 0.05 for the specified *df*, e.g., 3.84 for $df = 1$, then the null hypothesis is accepted (example (1) of Table B.1, where the overall Chi-square test statistic is 0.174). This does not mean that the null hypothesis is necessarily correct; rather, given the observed data there is no evidence to reject the null hypothesis.

Table B.2: Theoretical Chi-square values with different probabilities

Degrees	Probability (*p*-value)

Of Freedom	0.05*	0.01**	0.001***	0.0001****
1	3.84	6.64	10.8	15.0
2	5.99	9.21	13.8	18.5
3	7.82	11.35	16.3	21.0
4	9.49	13.28	18.5	23.5
5	11.07	15.09	20.5	25.8

If the observed test statistic value is greater than that given under $p = 0.05$, but less than that given under $p = 0.01$, for the specified *df,* then we reject the null hypothesis at the $p < 0.05$ value. This means that the probability that the observed data were generated under the null hypothesis is low, namely less than 5%. Note that the type 1 error level listed ($p < 0.05$ in this case) gives the probability that we have *falsely* rejected the null hypothesis, i.e., there is a 5% chance under the null hypothesis of getting the observed, or more extreme, values (i.e. >3.84 and <6.4). However, we must balance the type 1 error against our ability to detect deviations from the null hypothesis.

If our test statistic value falls between the $p = 0.01$ and the $p= 0.001$ values for the specified *df* (i.e., between 6.64 and 10.83 for $df = 1$), then we would reject the null hypothesis at the $p < 0.01$ value and would have more confidence in this rejection; in this case, there is only a 1% chance that a test statistic value greater than 6.64 would be generated by chance under the null hypothesis. Similarly, if the observed test statistic is greater than 10.83 with $df = 1$, we reject the null hypothesis at the $p < 0.001$ value. For example (2) of Table B.1, where the overall Chi-square test statistic is 69.4, we reject the null hypothesis at the $p < 0.0001$ value, using Table B.2.

More detailed *p*-values are also available from the internet (e.g., http://faculty.vassar.edu/lowry/tabs.html#cs, http://www.danielsoper.com/statcalc/calc11.aspx). For example (2) of Table B.1 we can actually reject the null hypothesis with $p < 0.000001$, i.e., we can be almost absolutely certain that the observed values deviate from HWP. In this case we would need to investigate the possible reasons for this large deviation from HWP (deficiency of heterozygotes, excess homozygotes) such as typing errors, non-random mating (e.g., a self-fertilizing plant), etc.

In Table B.3 below, we consider a series of examples extending those in Table B.1, all with artificially constructed genotype counts illustrating varying degrees of fit and lack of fit to HWP. Next to the observed numbers in each case the (O - E) values are given in parentheses. The $df = 1$ in all examples (see below). As discussed above, for a Chi-square distribution with 1 *df*, the cut-off points, i.e., rejection of the null hypothesis of fit to HWP are 3.84 (5%), 6.64 (1%), 10.83 (0.1%), and 15.0 (0.01%), as given in Table VI.2. We also often put *'s after the *p*-value,

* for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$, and **** for $p < 0.0001$,

since these are easier to spot visually than the actual *p*-values.

Table B.3: Hypothetical examples showing varying degrees of departure from HWP[#]

A_1A_1	A_1A_2	A_2A_2	Chi-square test statistic
32 (-4)	56 (8)	12 (-4)	2.78
31 (-5)	58 (10)	11 (-5)	4.34, $p < 0.05$*

29 (-7)	62 (14)	9 (-7)	8.51, $p < 0.01$**
27 (-9)	66 (18)	7 (-9)	14.06, $p < 0.001$***
26 (-10)	68 (20)	6 (-10)	17.36, $p < 0.0001$****

[#] The observed genotype counts (O) are listed, with the (O - E) values in parentheses, in all examples f(A_1) = p_1 = 0.6, f(A_2) = p_2 = 0.4, and the HWP expected counts in all cases are 36, 48, and 16. The Chi-square test statistic is the sum of the $(O\text{-}E)^2/E$ values, and the df = 1 in all cases (see below).

C. Determining the degrees of freedom (df)

When the expected (E) values under the null hypothesis are independent of any features of the observed data, then the df for a particular data set is given by the number of genotype classes considered minus 1 (since the total number of genotypes is a fixed number) (as above, for the moment we are only considering examples with all $E \geq 5$). Since the observed data are used in calculations to determine the expected values when testing for HWP, the df is as above (number of genotype classes - 1) and then minus the number of independent parameters estimated.

For a two-allele system, only 1 independent parameter is estimated. If we estimate the allele frequency p_1 then we know the estimate of p_2, because $p_1 + p_2 = 1$. The number of genotype classes is 3, hence df = 3 -1 -1 = 1 as in the examples in Tables B.1 and B.3.

For a k-allele system, k - 1 independent allele frequencies are estimated (if we have k - 1 allele frequency estimates, then we know the estimate for the kth allele): the number of different genotype classes is k (homozygotes) + $k(k\text{-}1)/2$ (heterozygotes) = $k(k+1)/2$, and thus, df = $k(k+1)/2 - 1 - (k\text{-}1) = k(k\text{-}1)/2$. In this case:

$$X^2_{HW} = \sum_{i=1}^{k(k+1)/2} (O_i - E_i)^2/E_i \qquad \text{(eq. B.1)}$$

where O_i is the observed number of individuals with the ith genotype, E_i is the expected genotype count under HWP. Thus, when $E_i \geq 5$ for each genotype class, X^2_{HW} has a Chi-square distribution with $df=k(k\text{-}1)/2$.

D. Lumping genotype classes when expected values are < 5

When $E_i < 5$ for a genotype class, the observed and expected counts of all such genotype classes must be combined, and the df is adjusted appropriately. If the combined E_i is still <5, these combined classes are then combined with the genotype class with the smallest value of $E_i \geq 5$ (unless there is a biological reason for a different lumping scheme); i.e., genotype classes must be combined until all classes have an expected value $E_i \geq 5$, and the Chi-square test of goodness of fit is performed on the resulting data set. Note that there may be insufficient df to permit statistical testing of this combined dataset. For example, if one genotype in a two-allele system has an E_i <5, and therefore, is combined with another genotype class, there will be no df

for a statistical test. However, the *exact* (complete enumeration) and *approximation* (resampling) tests (described below) can be used in these cases without combining classes.

To illustrate the calculation for *df* in these combined datasets, we consider a hypothetical case where there are four alleles, denoted A_1, A_2, A_3, and A_4. Suppose 2 of these alleles, A_3 and A_4, are rare, so that after combining in order that $E_i \geq 5$, there are 4 genotypic classes, namely: A_1A_1, A_1A_2, A_2A_2, and "combined" (the remaining genotypes). Note that the alleles A_3 and A_4 are only found in the "combined" category; when we calculate the *df* for the reduced data set it is valid to deduct only 2 for the number of independent allele frequencies estimated, giving *df* = 4-1-2 = 1. Note that if we had deducted the 3 independent allele frequencies, there would be insufficient *df* remaining to test fit to HWP.

The rules described here, including for the *df* calculation, apply to any combined set of genotypes. See below for two examples with three alleles that require combining.

E. Two examples of Chi-square testing of HWP that require lumping

We consider two examples that require lumping, due to $E_i < 5$ values, in a three-allele codominant system (Table B.1). The allele frequencies are the same in both examples, and example (1) shows a close fit to HWP, while example (2) has a large deviation from HWP.

Table B.4: Two examples of lumping classes in testing of HWP

(1) Example with close fit to HWP: $p_1 = 0.5$, $p_2 = 0.3$, $p_3 = 0.2$

	A_1A_1	A_1A_2	A_1A_3	A_2A_2	A_2A_3	A_3A_3	
O	26	29	19	10	11	5	Total: 100
E	25	30	20	9	12	4	Total: 100
O	26	29	19	15	11	-	Total: 100
E	25	30	20	13	12	-	Total: 100
(*O* - *E*)	1	-1	-1	2	-1	-	
$(O-E)^2/E$	0.0400	0.0333	0.0500	0.3077	0.0833	-	Total: 0.5143, ns

(2) Example with large deviation from HWP: $p_1 = 0.5$, $p_2 = 0.3$, $p_3 = 0.2$

	A_1A_1	A_1A_2	A_1A_3	A_2A_2	A_2A_3	A_3A_3	
O	33	20	14	16	8	9	Total: 100
E	25	30	20	9	12	4	Total: 100
O	33	20	14	25	8	-	Total: 100
E	25	30	20	13	12	-	Total: 100
(*O* - *E*)	8	-10	-6	12	-4	-	
$(O-E)^2/E$	2.56	3.33	1.80	11.08	1.33	-	Total: 20.10****

Since the expected value for the A_3A_3 genotype is < 5 in both examples, this category must be combined with another category, in this case A_2A_2 (the category with the next smallest expected value). Now the expected value for the new combined class of A_2A_2 and A_3A_3 is $\geq$ 5, and the standard Chi-square test can be performed. The df = 2 in both cases, there are 5 genotype classes after lumping with 2 independent allele frequency estimates, so df = 5 - 1 - 2 = 3. The data in example (2) deviate significantly from HWP with $p < 0.0001$ (Table VI.2). Note in this case that all the homozygotes have an observed excess frequency over HWP expectations, while all the heterozygote classes have an observed reduced frequency compared to HWP expectations, and this may give a clue to typing errors or admixture or inbreeding effects.

Appendix C: Malaria and sickle cell anemia and deviations from HWP

Malaria provides a strong selective pressure, and has led to polymorphism of a number of genes in humans. A well-known example, related to *Plasmodium falciparum*, involves variants at the hemoglobin β gene, notably Hb-S (homozygotes have sickle cell anemia), with heterozygote advantage in malarial environments. Individuals with sickle cell trait (AS heterozygotes) are more resistant to malaria than individuals who are homozygous AA; SS individuals are also more resistant to malaria than AA individuals, but they have sickle cell anemia which drastically reduces their fitness. An illustrative example of genotype frequencies for newborns and adults in a malarial environment is given in Table C.1

Table C.1: Sickle cell anemia genotype frequencies in a malarial environment

(1) Among newborns: f(*A*) = p_1 = 0.8820, f(*S*) = p_2 = 0.1220:

Genotypes:	*AA*	*AS*	*SS*	
Newborns (O)	780	204	16	Total: 1,000
Newborns (E)	777.9	208.2	13.9	Total: 1,000
(O - E)	2.1	- 4.2	2.1	
$(O - E)^2/E$	0.006	0.085	0.317	

Total chi-square: 0.408, *df* =1, *ns*

(2) Among adults: f(*A*) = p_1 = 0.8835, f(*S*) = p_2 = 0.1165.

Genotypes:	*AA*	*AS*	*SS*	
Adults (O)	769	229	2	Total: 1,000
Adults (E)	780.6	205.9	13.5	Total: 1,000
(O - E)	-11.6	23.1	-11.5	
$(O - E)^2/E$	0.17	2.59	9.80	

Total chi-square: 12.56, *df* =1, $p < 0.0004$***

The adult sample shows very significant deviation from HWP, while the newborn sample does not. This reflects the fact that at birth (pre-selection) the genotype frequencies are expected to be in HWP, whereas after selection has acted (malaria and sickle-cell anemia), the adult population shows significant deviation from HWP. The fact that the allele frequencies are very similar in newborns and adults may indicate that the population is close to equilibrium.