Journal of Heredity 2015:106(1):1–19
doi:10.1093/jhered/esu062
Advance Access publication November 25, 2014

Testing for Hardy–Weinberg Proportions: Have We Lost the Plot?

Robin S. Waples

From the Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd. East, Seattle, WA 98112.

Address correspondence to Robin S. Waples at the address above, or e-mail: robin.waples@noaa.gov.

Abstract

Testing for Hardy–Weinberg proportions (HWP) is routine in almost all genetic studies of natural populations, but many researchers do not demonstrate a full understanding of the purposes of these tests or how to interpret the results. Common problems include a lack of understanding of statistical power and the difference between statistical significance and biological significance, how to interpret results of multiple tests, and how to distinguish between various factors that can cause statistically significant departures. In this perspective, which focuses on analysis of genetic data for nonmodel species, I 1) review factors that can cause departures from HWP at individual loci and linkage disequilibrium (LD) at pairs of loci; 2) discuss commonly used tests for HWP and LD, with an emphasis on multiple-testing issues; 3) show how to distinguish among possible causes of departures from HWP; and 4) outline some simple steps to follow when significant test results are found. Finally, I 5) identify some issues that merit particular attention as we move into an era in which analysis of genomics-scale datasets for nonmodel species is commonplace.

Subject areas: *Conservation genetics and biodiversity; Bioinformatics and computational genetics*

Key words: *genotypic ratios, Hardy–Weinberg equilibrium, heterozygotes, linkage disequilibrium, multiple testing, statistical tests*

Every biologist with even the most cursory understanding of genetics knows about the Hardy–Weinberg (HW) principle (also known as the HW law), which was elucidated at the very beginning of the field of population genetics, soon after Mendel's work was rediscovered (Crow 1988). The HW principle makes 2 postulates of fundamental importance: 1) after a single episode of random mating, genotypic frequencies can be expressed as a simple function of allele frequencies, and 2) in the absence of perturbing forces (such as selection, genetic drift, mutation, migration), genotypic and allele frequencies remain constant over time. The first point is particularly important because it means that genetic characteristics of populations can be described in terms of frequencies of a relatively few alleles rather than the much larger arrays of all possible genotypes (Hedrick 2000; Allendorf et al. 2013). May (2004) noted that the HW principle is the evolutionary analogue to Newton's First Law of Motion (bodies at rest remain at rest unless perturbed), although he emphasized the constancy of allele frequencies rather than the genotypic frequencies that are the core of the HW principle.

If we let p and $1 - p$ be frequencies of alleles A and a at a gene locus, then Hardy–Weinberg *equilibrium* (HWE) is said to occur when frequencies of the genotypes AA, Aa, and aa are p^2, $2p(1 - p)$, and $(1 - p)^2$, respectively. Extension to multiple alleles is straightforward. Although the HWE terminology is widely used, it is not a true equilibrium, particularly when real populations are considered that experience random genetic drift. In the following, therefore, I will discuss how to evaluate agreement of observed genotypic frequencies with the Hardy–Weinberg proportions (HWP) expected to occur if the assumptions of the Hardy–Weinberg principle are met.

The original papers by Hardy (1908) and Weinberg (1908) considered only single genes, but Weinberg (1909) soon showed that a similar principle applies to pairs of loci: If the loci assort independently and other HW assumptions are met, the frequencies of 2-locus gametes are simple functions of allele frequencies at the 2 loci. Curiously, whereas single-locus analyses are generally cast in terms of HW equilibrium, results of 2-locus analyses typically are expressed in terms of linkage *dis*equilibrium (LD). Although gametic disequilibrium might be a more exact general term, the LD terminology is in such widespread use that I adopt it here to include all nonrandom associations of alleles at different loci, whether or not the loci are physically linked on the same chromosome. With 2 loci, a single generation of random mating does not remove all LD; instead, the approach is asymptotic at a rate that depends on the probability of recombination between the loci.

Statistical tests of HWP (see next section for details) attempt to answer the question, "After accounting for sampling error, are genotypic frequencies observed in a sample compatible with those expected under HWP?" Agreement between observed and expected genotypic frequencies can be conveniently expressed using Wright's coefficient of inbreeding, F_{IS}:

$$F_{IS} = 1 - \frac{H_o}{H_e} = \frac{H_e - H_o}{H_e}, \quad (1)$$

where H_o is the observed fraction of heterozygotes and H_e is that expected under HWP. A positive F_{IS} indicates a deficiency of heterozygotes compared with the HWP expectation, while a negative F_{IS} indicates an excess.

Prior to the 1960s, genotypic data were available for only a relatively few well-studied markers whose genetic basis was understood. Under those conditions, tests of agreement of observed and expected genotypic frequencies generally focused on whether departures from HWP could be attributed to nonrandom mating or to other factors such as selection or population mixture. With the advent of protein electrophoresis, however, for the first time it became possible to generate data for dozens of new markers for a wide range of species (Lewontin and Hubby 1966; Utter et al. 1974; Powell 1975). Because pedigree studies to validate the genetic basis of all this variation are difficult to conduct, researchers routinely found a new use for HW tests: to help screen out data that reflected scoring errors, posttranslational modification of phenotypes, or other nongenetic artifacts. The early years of protein electrophoresis witnessed some dubious misuses of HW tests for these purposes. For example, Koehn (1972) complained that many papers on North Atlantic eels (*Anguilla* spp.) blithely interpreted electrophoretic phenotypes in terms of genetic variation in spite of resounding refutation of agreement with HWP. The most egregious offense noted by Koehn involved a sample in which all 43 individuals were reported to be heterozygotes for a locus with 2 alleles; this produced a huge χ^2 value which the author reported but then went on to use the data to draw inferences about population structure.

Researchers (as well as reviewers and journal editors) gradually became more vigilant in restricting inferences to enzyme systems that produced results compatible with HWP. The tests themselves have been rigorously evaluated and incrementally improved over time (Vithayasai 1973; Emigh 1980; Guo and Thompson 1992; Wigginton et al. 2005; Kulinskaya and Lewin 2009), and vast improvements in computational power and more efficient computer algorithms now make it easy for researchers to rapidly assess the degree of conformity to HWP of large datasets (Excoffier and Heckel 2006). However, although these developments are encouraging, they do not in any sense indicate that issues related to testing HWP and LD have been "solved." It is not enough to have an accurate assessment of the degree of conformity to HWP; it is also essential to be able to properly interpret results of those assessments and take appropriate actions. An informal survey of recent published literature identifies several general problems in this regard.

First, it is axiomatic in frequentist statistics that failure to reject a hypothesis does not prove that the hypothesis is true, but this important issue is often forgotten or ignored. Agreement with HWP is not a guarantee that observed variation has a genetic basis, nor does it guarantee that other factors such as selection, drift, and nonrandom mating are not influencing genotypic frequencies. This has huge importance for testing of HWP, as the test generally has low power to detect actual departures unless sample sizes are large. For example, Fairbairn and Roff (1980) showed that, if presumed genotypes in a diallelic system are scored randomly rather than according to a validated genetic model, a sample of 50 individuals would fail to detect any departures from HWP over one-third of the time using an $\alpha = 0.05$ significance criterion. Furthermore, it is also possible to have large departures from HW assumptions that nevertheless do not create any departures from HWP (see Box 1). In these scenarios, strong evolutionary forces are acting but cannot be detected by a simple test of HWP, regardless how large a sample size is obtained. Many modern researchers do not appear to be aware of these key findings.

A second pervasive issue is that researchers typically want to evaluate HWP for many loci sampled in many populations, often with temporal replicates as well, which means that it is essential to consider issues related to multiple testing. Rigorous methods are available to account for multiple testing, but researchers often misuse the resulting information in drawing conclusions. Multiple testing issues are particularly complex for tests of LD, as the number of comparisons increases as a function of the square of the number of markers.

Another chronic problem is a failure to distinguish between statistical significance and biological significance (Waples 1998; Hedrick 1999). By itself, the level of statistical significance, or *P* value, associated with a test of HWP or LD tells one nothing about 1) the magnitude of any evolutionary forces that might be acting on the population, or 2) whether any departures are likely to have a substantial impact on subsequent analyses of the genetic data. Statistical tests, therefore, should be only the first step in evaluating potential departures from equilibrium assumptions; however, many researchers never get beyond this first step.

Finally, many researchers appear to have lost the plot regarding tests of HWP—that is, lost track of *why* they are being done in the first place. Admittedly, the reasons for performing HWP and LD tests can be varied; for example, some researchers have a specific alternative hypothesis to evaluate for a specific gene locus or set of loci. However, for population genetic studies of nonmodel species, the most common reason for performing HWP tests (and the focus of this Perspective) is that the researcher knows that doing so is a required first step before moving on to the interesting stuff: using a variety of sophisticated and sexy software to analyze their data. The desire to dispense with HW testing as quickly as possible is understandable, but moving on to the next step should be contingent not only

Box 1. Strong selection but Hardy–Weinberg agreement in American eels

An example using American eels (*Anguilla rostrata*) shows how strong selection can be compatible with HWP. Williams et al. (1973) and Koehn and Williams (1978) found allele frequency differences on the order of 10% among nearby North American rivers, in spite of the generally-accepted assumption that all North American eels breed randomly in the Sargasso Sea. The authors postulated that this reflected locally-adapted alleles in each stream. If true, this would require strong selection to substantially change allele frequencies within a single cohort of panmictic offspring. Williams and Koehn also reported that samples at each locality agreed with HWP. But does not the Hardy–Weinberg principle assume no selection? It turns out that these 2 apparently contradictory results can be reconciled using the principle described by Lewontin and Cockerham (1959), who showed that selection does not cause any HW deviations provided that $W_1W_3 = W_2^2$, where W_1 and W_3 are fitnesses of the alternative homozygotes and W_2 is fitness of the heterozygote. To illustrate, assume a cohort of 10 000 eels starts out with frequency 0.8 for the *A* allele and 0.2 for the *a* allele, and further assume the genotypes are in HWP, so the numbers of eels having each of the genotypes are $AA = 10000*0.8^2 = 6400$, $Aa = 10000*2*0.8*0.2 = 3200$ and $aa = 10000*0.2^2 = 400$ (Table 1.1). Now assume that the eels undergo selection, with probabilities of survival for the 3 genotypes given by $[W_1, W_2, W_3] = [0.36, 0.6, 1]$. Note that $W_1W_3 = W_2^2 = 0.36$. After selection, the numbers with genotypes [*AA*, *Aa*, *aa*] are reduced to [2304, 1920, 400], and frequency of the *A* allele has dropped to 0.706. But postselection genotypes are in perfect agreement with HWP, given the postselection allele frequencies, even though relative fitness of the genotypes differed by a factor of almost 3 and the cohort underwent strong directional selection that resulted in genetic deaths of over half of the population.

Felsenstein (1965) described an analogous situation involving epistatic interactions between 2 gene loci: even with strong selection, certain fitness relationships would produce no linkage disequilibrium (LD). Conversely, interaction of the introduced virus myxomatosis with European rabbits illustrates how epistatic selection can create strong (albeit transitory) LD for genes that are not physically linked. Functional antibodies require polypeptides encoded by 2 different genes. van der Loo et al. (1987) found consistent LD among European rabbit antibody genes located on different chromosomes, suggesting strong epistatic selection. Within the lifetime of a cohort, selection favors associations of alleles that produce effective antibodies, but recombination erodes those associations every generation, creating a large genetic load. The myxomatosis/European rabbit association is very recent on evolutionary time scales (~one century), which could explain why chromosomal rearrangements have not yet brought the favorable gene combinations into tight linkage, as theory would predict (Kimura 1956).

Table 1.1 Effects on genotypic frequencies of an episode of strong selection within a hypothetical cohort starting with $N = 10000$ individuals

	AA	*Aa*	*aa*	*N*	*p*
Before	6400	3200	400	10 000	0.800
$[W_1, W_2, W_3]$	0.36	0.6	1		
After					
Observed	2304	1920	400	4624	0.706
Expected	2304	1920	400		

$[W_1, W_2, W_3]$ are relative fitnesses of the 3 genotypes [*AA*, *Aa*, *aa*], and *p* is allele frequency.

on having performed the required tests, but also on having shown that results are compatible with fundamental assumptions made in subsequent analyses. All too often, "completion of HWP tests" is treated merely as a box that has to be checked rather than a process that involves careful evaluation of the results.

In this Perspective, I will 1) review factors that can cause departures from HWP at individual loci and LD at pairs of loci; 2) discuss commonly used tests for HWP and LD, with an emphasis on multiple-testing issues; 3) show how to distinguish among possible causes of departures from HWP; and 4) outline some simple steps to follow when significant test results are found. Finally, I 5) identify some issues that merit particular attention as we move into an era in which analysis of genomics-scale datasets for nonmodel species is commonplace. Fine treatments of some of these topics can be found elsewhere, in papers by Fairbairn and Roff (1980) and Lessios (1992) or textbooks by Hedrick (2000), Allendorf et al. (2013), and others. However, this information is routinely ignored or misused in the implementation of testing for HWP and LD, so a new, comprehensive treatment seems warranted.

Factors That Can Cause Deviations from HWP

The Hardy–Weinberg principle depends on a number of assumptions, including simple Mendelian inheritance in a diploid organism with discrete generations, random mating, an infinite population, and no mutation, migration, or selection. In testing for agreement with HWP, there is also an implicit assumption of random sampling from the population as a whole.

Mutation

Mutations in gametes can be passed on to the next generation, while somatic mutations accumulate over time and can change allele frequencies within a cohort. In theory, either of these processes could disrupt the relationship between allele frequencies in parents and genotypic frequencies in their offspring. In reality, however, mutations are rare enough that these single-generation effects generally can be ignored in evaluating HWP and LD. However, cumulative effects of mutation are important for understanding patterns of LD because loci that are tightly linked to new advantageous

mutations can rapidly increase in frequency through genetic hitchhiking (Barton 2000).

Finite Population Size

The assumption of infinite population size is of course never satisfied in nature. The practical relevance is that finite populations have a tendency to produce heterozygotes at higher frequencies than predicted from HWP; this is the basis for the heterozygote-excess method for estimating effective population size, N_e (Pudovkin et al. 1996). The traditional explanation for this phenomenon has been that an excess of heterozygotes occurs when allele frequencies differ between the sexes, and this is more pronounced in small populations (Robertson 1965). Balloux (2004) showed that hermaphrodites that cannot self-fertilize also produce an excess of heterozygotes of the same magnitude:

$$E\frac{H_o}{H_e} = 1 + \frac{1}{2N_e + 1}, \tag{2}$$

where $E(H_o/H_e)$ is the expected value of the ratio (H_o/H_e). A simple rearrangement of Equation 1 produces $H_o/H_e = 1 - F_{IS}$, which means Equation 2 can be rewritten as

$$\begin{aligned} E(1 - F_{IS}) &= 1 + \frac{1}{2N_e + 1}, \quad \text{so} \\ E(F_{IS}) &= -\frac{1}{2N_e + 1}. \end{aligned} \tag{3}$$

With a finite number of parents, therefore, we expect that F_{IS} will be negative by the approximate magnitude $1/(2N_e)$. Unless N_e is very small (no more than a few dozen or so), this will be a small effect, but it potentially can be detected in species with high fecundity, in which case it might be possible to obtain large samples of offspring produced by just a few individuals (Hedgecock et al. 2007).

A comparable effect occurs at pairs of gene loci: A finite number of parents produce offspring with random LD due to drift that can be measured by the statistic r^2, which is the squared correlation of alleles at different gene loci. If the loci are unlinked, the expected magnitude of LD is approximately

$$E(r^2) \approx \frac{1}{3N_e} + \frac{1}{S}, \tag{4}$$

where S is the number of individuals in the sample (Hill 1981). Because the number of pairwise comparisons increases rapidly with the number of alleles and loci, power to detect overall drift effects across the genome can be high, and the LD method is widely used to estimate effective population size (Waples and Do 2008; Palstra and Fraser 2012).

Selection

Natural selection operates through differential survival and/or reproduction of individuals with different genotypes, so it is easy to see how selection can distort genotypic frequencies from HWP. This is most likely to occur if selection favors either heterozygotes (overdominance) or homozygotes (underdominance), in contrast to directional selection for or against a particular allele. In addition, selective pressures that differ between males and females (or other factors) can cause allele-frequency differences between the sexes, and this produces an excess of heterozygotes in their offspring (see Hedrick 2000 for details). For diploid autosomal genes, these sex-based allele frequency differences disappear in the offspring, and if those offspring randomly mate, HWP are restored in their progeny.

Two issues are important regarding selection. First, because HWP are generated following a single episode of random mating, selection in previous generations has no effect on current genotypic frequencies; selection that affects HWP must occur over the lifetime of a cohort. [An exception would be selection at a sex-linked locus; see "Sex Linkage" below.] Second, even very strong selection within a cohort can produce an array of genotypes that does not deviate from HWP (Wallace 1958; Lewontin and Cockerham 1959; see Box 1).

When fitness effects of a gene depend on the genetic background, selection for or against certain combinations of genes can create LD. Because recombination breaks down gene–gene combinations unless they are close together on the same chromosome, many have concluded that these effects are ephemeral, and as a consequence epistasis plays a relatively small overall role in selection and adaptation. However, that view has been challenged (Hansen 2013), and Box 1 shows how strong epistatic selection can create recurrent LD among markers on different chromosomes (and large genetic load) in a population facing novel selective pressures.

Population Structure

A key assumption of the HW principle is that the sample in question is drawn from a single, randomly mating population. This assumption might be violated for a variety of reasons: individuals might be sampled on feeding grounds or during migrations at a time and place where multiple populations overlap; population boundaries might be fuzzy or hard to detect; or the species might be continuously distributed with localized breeding structure. If the sample includes a mixture of individuals from more than one breeding unit, then (on average) H_o will be less than H_e, leading to a deficiency of observed heterozygotes (Wahlund 1928) and a positive F_{IS}. The magnitude of this "Wahlund effect" increases with the degree of population differentiation and evenness of the mixture proportions. In the notation below, we assume a mixture that includes populations 1 and 2 in proportions m and $(1 - m)$. Let the frequencies of a given allele be p_1 in population 1 and p_2 in population 2, so $\bar{p}_w = mp_1 + (1 - m)p_2$ = the weighted mean of p_1 and p_2. A simple derivation combining results from Wahlund (1928) and Wright (1951) (see Appendix) leads to the following result:

$$E(F_{IS}) = F_{ST}[4m(1-m)/C], \tag{5}$$

where C is a function of allele frequencies and mixture fractions. This produces the elegantly simple result that the magnitude of the Wahlund effect in a mixed sample (measured by F_{IS}) should be an increasing function of the standardized variance of allele frequency between the populations in the mixture (F_{ST}). That is, F_{IS} should be larger at loci with large F_{ST}. Under 2 special cases (equal mixture fractions or fixed allele differences), $4m(1-m)/C = 1$, in which case theory predicts a linear relationship between F_{IS} and F_{ST} with a slope of 1. More generally, we expect a positive correlation with a slope that decreases as the mixture fractions become more unequal.

A 2-locus analogue to the Wahlund effect, whereby LD is generated in a sample that includes more than one gene pool, was described by Nei and Li (1973) and Sinnock (1975). The magnitude of this effect is a function of the product of F_{ST} values for the 2 loci (Waples and England 2011):

$$E\left(r^2_{\text{mix}}\right) \approx C_2 F_{\text{ST}(1)} F_{\text{ST}(2)}, \tag{6}$$

where r^2_{mix} is the component of overall r^2 due to population mixture and C_2 is a constant that depends on the number of populations involved and their relative proportions in the mixed sample (and, probably, allele frequencies at the loci involved). That is, pairs of loci for which the product of single-locus F_{ST} values are largest should be the loci for which r^2 values are largest in a population mixture.

Whereas the single-locus Wahlund effect disappears with a single generation of random mating, LD created by population admixture (in which matings occur between individuals from different populations) decays gradually over time at a rate determined by the probability of recombination. Assuming no new interbreeding events, half the residual LD decays each generation for unlinked loci, but if the recombination rate is low the effects of population admixture can persist for many generations.

Age Structure

The Hardy–Weinberg principle implicitly assumes generations are discrete. Age structure can create Wahlund-like effects within single populations that are otherwise mating randomly. In species with overlapping generations, individuals in a single cohort are produced by adults that participate in 1 reproductive cycle, not by random mating of all adults across a generation. Parents in different reproductive cycles will differ somewhat in allele frequencies, so their offspring will as well. Therefore, a sample of mixed-age individuals is in essence composed of a number of subpopulations, with the expected result being a deficiency of heterozygotes compared with HW expectations at individual loci and a component of mixture LD at pairs of loci. These effects will generally be small but can be important in some cases, at least for mixture LD (Waples et al. 2014). When mixed-age parents randomly mate to produce a single cohort of offspring, this mini-Wahlund effect disappears at individual loci but only decays by half at pairs of unlinked loci.

Assortative Mating

In assortative mating, mate choice depends on the phenotype. By itself, this does not change allele frequencies but can affect genotypic frequencies. To the extent that the phenotype predicts the genotype, positive assortative mating (among phenotypically similar individuals) will tend to reduce H_o, leading to positive F_{IS}, while negative assortative mating will have the opposite effect. The most extreme form of assortative mating is self-fertilization, which occurs in many species. At equilibrium in a system involving partial self-fertilization with probability s, the frequency of Aa heterozygotes will be (Hedrick 2000):

$$\text{Freq}(Aa) = \frac{4p(1-p)(1-s)}{2-s}. \tag{7}$$

This differs from the HWP expectation $[2p(1-p)]$ by the factor $2(1-s)/(2-s)$, which reduces to 1 for $s = 0$. The equilibrium relationship in Equation 7 is only approached gradually over many generations, and HWP are restored after a single episode of random mating.

As we found with selection, it is possible to identify patterns of assortative mating that do not create any departures from HWP at all; Li (1988) referred to these scenarios as pseudo-random mating.

It is important to realize that tests of HWP provide no information about cumulative levels of inbreeding in a population; F_{IS} only reflects the most recent generation of mating. Doyle (2014) described a cautionary tale on this theme. In the tropical shrimp (*Penaeus* spp.) farming industry, commercial hatcheries that provide breeders to grow-out operations maintain large, genetically diverse broodstocks. However, the breeders they provide to individual farmers typically consist of 2 full-sib families, with careful instructions about how to conduct the matings (by hybridizing across, not within, families) to produce high-quality and uniform offspring with minimal inbreeding. Under this "authorized" scenario, the hybrid offspring of these sanctioned matings are closely related, but are all sold for consumption, which requires the farmers to obtain new breeders from the suppliers each generation. Problems arise when these genetically uniform offspring are used for subsequent generations of breeding, either in the same farm or at "copy" farms they are shipped to. These unauthorized distribution channels, which according to sources Doyle cites might represent 50–90% of total production in many areas of the world, produce inbred offspring with reduced productivity. Although recent disease epidemics have severely reduced production by tropical shrimp aquaculture, responses primarily have focused on containment, with little or no attention to inbreeding (e.g., Jones 2012). Doyle (2014) argued that increased disease susceptibility due to pervasive inbreeding throughout the unofficial propagation channels is a root cause of the recent epidemics, and that genetic factors have been overlooked at least in part because a general lack of significant departures from HWP in copy hatcheries was interpreted as evidence that the populations were not suffering from inbreeding.

Sex Linkage

In mammals and *Drosophila*, females have 2 X chromosomes and males 1 X and 1 Y, the latter being devoid of many genes (Rice 1996). In birds and Lepidoptera, the situation

is reversed, with females having 2 different chromosomes (Z&W). In some species, individuals of the heterogametic sex (e.g., XY or ZW) carry only one copy of sex-linked genes, while members of the homogametic sex carry 2. With random mating and no other disturbing forces, XX females and ZZ males will have genotypic frequencies in HWP, provided that allele frequencies are the same in the 2 sexes; if that is not the case, the homogametic sex will generally show an excess of heterozygotes.

In many species, both sex chromosomes have functional genes in what are known as pseudoautosomal regions (see Hedrick 2000 and Allendorf et al. 2013 for details). Genes in these regions often behave like autosomes unless they are tightly linked to the sex-determining gene(s). With tight sex linkage, however, these pseudoautosomal genes can differ sharply in allele frequency between males and females (Clark 1988), and this produces an excess of heterozygotes in the heterogametic sex and in the population as a whole, even though the genotypic frequencies in the homogametic sex will generally conform to HWP. Berlocher (1984) and Marshall et al. (2004) provide examples of this phenomenon.

Nonrandom Sampling

Even if genotypes in the population as a whole are in HWP, those in a sample might not be, either because of random sampling error or because individuals with certain genotypes have a higher or lower probability of being sampled. Statistical tests address the first possibility, but the second is more insidious. In a truly random sample, every individual in the population has an equal opportunity of being sampled, independent of every other individual. This is nearly impossible to achieve in any real-world situation, but with luck it can be approximated. For tests of HWP, the key is whether individuals that are heterozygotes are more or less likely to appear in the sample than would occur purely by chance. This might happen, for example, if susceptibility to sampling depends on the phenotype, which reflects at least in part the underlying genotype. Samples dominated by offspring from only a few families can also lead to deviations from HW proportions and other problems (Hansen et al. 1997; Jankovic et al. 2010).

Genotyping Errors

Although modern technology has made mass genotyping fast, cheap, and efficient, it is virtually impossible to remove all sources of error, particularly when one includes factors such as mis-labeling of samples and recording, transcribing, and analyzing data. Mis-labeled samples, for example, can produce results that can be mistaken for migration. Random genotyping errors generally lead to only weak departures from HWP unless they are very extensive or sample sizes are very large (Fairbairn and Roff 1980; Cox and Kraft 2006). Errors that are more likely to affect HWP include "null" alleles (Pompanon et al. 2005), a term that has been widely used to describe several different phenomena: alleles having a mutation that prevents production of a functional gene product; alleles that produce functional products that are not detected by the analytical method used; and DNA sequences that are not detected because of a mutation in the primer-binding region. Each of these phenomena will cause true heterozygotes to be missed or scored as homozygotes, leading to a heterozygote deficiency and positive F_{IS}.

Statistical Tests of HWP and LD

Individual Tests

All statistical tests involve tradeoffs between 2 types of errors of inference: falsely rejecting the null hypothesis when it is true (Type I error), and failing to reject the null hypothesis when it is false (Type II error). Researchers typically specify the probability of a Type I error they are willing to tolerate (α), recognizing that this will lead to some unavoidable fraction (β) of Type II errors. Statistical power (the probability of rejecting a null hypothesis that is false) is $1 - \beta$. Selecting a more stringent α causes β to increase and reduces power. From at least the time of Levene (1949) and Haldane (1954), the standard method for evaluating HWP has been to use the sample allele frequencies to generate expected genotypic frequencies according to the HW principle and then compare these with observed genotypic frequencies using a chi-square or related test. For a locus with 2 alleles, 3 genotypes are possible and the chi-square test has 1 degree of freedom, so $\chi^2 > 3.84$ is required to reject HWP at the traditional $\alpha = 0.05$ level. Low expected values in some genotypic classes make the chi-square test less reliable, and these problems increase dramatically as the number of alleles becomes large (as with microsatellites). Most recent implementations therefore rely on variations of "exact" tests, which calculate the fraction of all possible genotypic arrays that produce HW deviations more extreme than the sample in question. If the number of alleles at a locus is no larger than 4 or 5, the method of Louis and Dempster (1987) can be used to exhaustively sample all possible outcomes; for loci with more alleles it is necessary to use Monte Carlo methods to sample from the vast parameter space (Guo and Thompson 1992). With 2 alleles, the test of HWP is equivalent to a test of whether observed and expected frequencies of heterozygotes are statistically different. This is not necessarily the case with multiple alleles, as an overall test can be significant even if there is no overall excess or deficiency of heterozygotes.

Whereas tests of HWP consider frequencies of 2 alleles at the same locus but on different gametes, tests of LD consider frequencies of 2 alleles at different loci but on the same gamete (Weir 1996). For nonmodel species, one generally can only compute frequencies of genotypes, not gametes. Double heterozygotes (*AaBb*) create a problem because this genotype could be formed in 2 ways: gametes *AB*/*ab*, or gametes *Ab*/*aB*. Although a maximum-likelihood method (Hill 1974) that assumes random mating at the individual loci was used in a number of early studies, most recent assessments of LD for genotypic data use the composite (Burrows) method (Weir 1979), which is simple to calculate and does not assume random mating. Zaykin et al. (2008) describe a multi-allele version of a chi-square test of composite LD at pairs of dialleic loci.

Multiple Testing

If a dataset has L loci, there are L tests of HWP and $L(L-1)/2$ pairwise tests for LD for each sample. Experimental designs that include many loci scored in samples from many temporal and/or spatial strata can therefore involve a large number of different tests. If a statistical test is accurate and the nominal Type I error rate for each test is $\alpha = 0.05$, then, on average, 5% of the tests will be significant by chance, even if all assumptions for HWP and linkage equilibrium are met. The probability that at least one test will be significant by chance is termed the experiment-wide error rate (EWR), which is $1-(1-\alpha)^k$ for k independent tests. For example, with $\alpha = 0.05$, EWR = 0.40 and 0.99 for tests of HWP at 10 and 100 loci, respectively. In testing LD, the numbers of 2-locus comparisons are $k = 45$ and 4950 for $L = 10$ or 100, respectively, leading to EWRs of 0.90 and >0.999. As k gets large, therefore, it rapidly becomes a near certainty that at least some tests will be significant by chance alone unless a multiple-testing correction is implemented.

Several options are available to quantitatively account for simultaneous tests of similar hypotheses. The simplest (Bonferroni) procedure is to require that an individual test has a P value $< \alpha/k$ to be considered significant. This procedure ensures that the EWR is $\leq\alpha$, but it has reduced power to detect multiple departures from the null hypothesis (Miller 1981). For this reason, a sequential Bonferroni procedure (Holm 1979; Rice 1989), in which the P values are first ordered by magnitude, has been widely used. Although the sequential Bonferroni sacrifices less power, when k is large the test nevertheless becomes very stringent, and many real deviations from the null hypothesis can go undetected (Sunnucks and Hansen 2013). A conceptually different approach is to control the fraction of rejected hypotheses that are actually true, rather than trying to ensure that no hypotheses are falsely rejected at all. Rejected hypotheses can be considered "discoveries" because they provide evidence of an effect, so this alternative approach has been called the False Discovery Rate or FDR (Benjamini and Hochberg 1995). The FDR procedure can greatly increase power to detect departures from the null hypothesis, especially when k is large. The original FDR assumed independence of the tests; a modified version (Benjamini and Yekutieli 2001) accommodates dependencies at some cost in power, which is still much higher than for the sequential Bonferroni (Narum 2006).

Although the procedures themselves are relatively straightforward, proper application and interpretation of multiple tests of HWP and LD is challenging. First, one must decide how to group the tests into "families" or "experiments." Should each of J populations be considered a separate "experiment," in which case there are J different overall tests of HWP and J tests of LD, with each test integrating information across all loci? Or should the focus be on individual loci or pairs of loci, with each of the L or $L(L-1)/2$ tests integrating information across all populations? Note that either of these options has additional multiple-testing issues nested within. A final option is to consider all tests of all loci/locus pairs in all populations to be part of a single experiment, which would produce JL tests of HWP and $JL(L-1)/2$ tests of LD. The most appropriate design depends on the type of inferences one wants to draw (Miller 1981). In general, as discussed below, it will be more useful to use smaller aggregations to evaluate patterns across loci or populations.

Problems also arise regarding interpretation of results after adjusting for multiple testing. It is not uncommon to see researchers report some significant deviations after adjusting for multiple tests of HWP or LD but then ignore this result, especially if they are few in number. If a Bonferroni adjustment is used with $\alpha = 0.05$ to adjust for many HW tests, it means that, if the underlying assumptions of HWP are true for every locus in every population tested, then 95% of the time zero tests will be significant after adjusting the critical P value. This means that even a single significant departure after Bonferroni correction is unlikely to occur by chance and requires some other explanation. The False Discovery Rate approaches were developed to address a slightly different goal: to control the fraction of "discoveries" that are false and hence red herrings. For example, gene expression arrays can simultaneously assay many thousands of gene products (Schwanhäusser et al. 2011), and it is important to try to distinguish those that reflect meaningful up or down regulation from those that have a high or low signal just by chance. If a researcher uses a FDR of 0.05, therefore, it means they are willing to accept that 5% of the rejected null hypotheses will actually be true. On the other hand, this means that 95% of FDR-corrected discoveries would accurately reflect violations of the null hypothesis, so simply ignoring them is risky.

Discussion and Recommendations

For many decades, tests of HWP have played an important role as gatekeepers, helping to screen out genotypic arrays that cannot plausibly be explained by random mating and simple Mendelian inheritance. Accordingly, I begin the discussion by considering appropriate use of the tests for that role.

Drawing Inferences about Mendelian Variation

Because protein electrophoresis documents variation in gene products rather than the genes themselves, tests of HWP helped to identify patterns that reflected posttranslational modification of gene products or other nongenetic artifacts. An important additional screening criterion for allozymes was that the phenotypic banding patterns of heterozygotes had to be consistent with the known subunit structure of the enzyme: 2, 3, and 5 bands for monomers, dimers, and tetramers, respectively. Furthermore, different taxa have different and largely predictable patterns of tissue expression that could be used for additional quality control (Utter et al. 1987).

For markers such as microsatellites or single-nucleotide polymorphisms (SNPs) that (in theory at least) directly reflect DNA sequences, it is perhaps not as important to demonstrate that the observed variation "has a genetic basis." However, it is no less important than it was for allozymes to ensure that the recorded genotypes accurately reflect the true genotypes, and all current methods

for DNA sequencing are prone to errors of this type. Furthermore, no additional criteria are available for microsatellites or SNPs that are comparable to the subunit and tissue-specificity tests routinely used as part of quality control for allozymes. For many applications involving DNA data, tests of HWP are the primary evidence used to support the hypothesis that observed patterns of variation can be interpreted in terms of simple genetic models. This is unfortunate, as even large samples generally provide low power to detect many nongenetic artifacts (Fairbairn and Roff 1980; Cox and Kraft 2006). HW tests should be considered a necessary but not sufficient step in establishing Mendelian inheritance. Researchers should not lose sight of the fact that, by themselves, tests of HWP and LD provide at most weak support for hypotheses regarding the genetic basis of observed variation. If at all possible, researchers should use breeding studies, parentage analysis, or other approaches to more rigorously evaluate this hypothesis.

Multiple Testing

Although researchers are increasingly aware of the importance of multiple-testing issues, the most widely-used approaches are not ideally suited for common applications of HW tests for nonmodel species. The FDR is most useful when one expects to find many "discoveries" (significant *P* values) and wants to ensure that not too much time and effort is spent following leads that turn out to be false. In contrast, the typical researcher evaluating genetic data for a nonmodel species does not want to find *any* departures from HWP; instead, one hopes the tests will allow the conclusion that, in the aggregate, the presumptive genotypes are consistent with what is expected for genetically-based variation within a random-mating population that meets the other criteria for HWP.

I recommend a 2-step approach to multiple testing that begins with a "big-picture" evaluation of the extent of agreement with HWP using raw (unadjusted) test results, followed by a more targeted evaluation of outlier loci and populations, which could involve use of formal multiple-testing procedures or other approaches (see Boxes 3 and 4). The simplest way to begin is to perform separate, unadjusted tests for every locus in every population and record the fraction that is statistically significant. If this fraction is no larger than the nominal Type I error rate (α), one can conclude that the tests as a whole provide no evidence to reject HWP or demonstrate LD (Table 3.1, Figure 3.1). If the HW test is accurate and all null hypotheses are true, the observed fraction of significant tests might be exactly α, but is more likely to be slightly higher or lower. Box 3 shows how to evaluate whether the observed fraction of significant tests is statistically different from α, using the cumulative binomial distribution. If the fraction of rejected hypotheses is significantly less than α, it suggests that the test might be conservative; if the fraction is significantly higher than α, it suggests that the overall hypothesis of generalized agreement with HWP is not true.

Another approach is use Fisher's combined probability test (or the weighted Z-method; Whitlock 2005), which jointly considers multiple tests of the same hypothesis and assesses the probability (*P*) that all hypotheses are true. This can be informative for checking patterns across populations or loci to identify samples or genetic markers that consistently deviate from HWP. One caveat for this approach is that the overall *P* will be reported as 0 if any single *P* value is 0. Most software that uses permutations to compute exact HW tests will report $P = 0$ if none of the n permutations considered produced a result as extreme as the data in question. But $P = 0$ implies the observed data are impossible if the null hypothesis is true; in general, the most that can be said in that case is that $P < 1/n$.

A third approach is to compare the distribution of *P* values with the expected null distribution, which should be flat across the interval 0–1 (e.g., 10% of *P* values should fall between 0.1 and 0.2, between 0.2 and 0.3, etc.; see Box 3). One can visually examine the pattern of test results to identify substantial departures from the null expectation, or evaluate them quantitatively with a goodness-of-fit test. Finally, one can partition the data to create 2 or more groups of individuals or loci to see whether the pattern of significant deviations is consistent across groups.

Whether or not initial evaluations indicate that the fraction of significant tests exceeds the nominal α level, it is important to conduct additional analyses for evidence of individual test results that are strongly divergent and therefore might substantially affect downstream analyses. One way to do this is to repeat the analyses illustrated in Box 3 using more stringent rejection criteria (e.g., $\alpha = 0.01$ or 0.001). Formal multiple-testing corrections or some of the approaches in Box 4 can be used to accomplish much the same thing.

Interpreting Results of Statistical Tests

If test results are consistent with the global assumptions of the HW principle and independent assortment, then the researcher can proceed to subsequent analyses of the data, without losing sight of the fact that this is no guarantee that all underlying assumptions have been met. Rigorous conclusions about these assumptions based on negative results are not possible in the absence of a power analysis to determine how large a departure from HWP or linkage equilibrium assumptions could occur and still go undetected, given the experimental design.

What should one do in the common situation where the assumption of global HWP is rejected—that is, when one has performed comprehensive HWP and LD tests for a large dataset and found some departures that remain significant after accounting for multiple testing? The following responses are NOT appropriate, even though it is easy to find recently-published papers where they are followed:

- Report this result and then proceed to subsequent analyses with no further discussion.

Box 2. Identifying 1- and 2-locus Wahlund effects

A significant deficit of heterozygotes is a common outcome of HW testing. A simple test can help distinguish population structure (Wahlund effect) from other causes of positive F_{IS}. If the sample includes more than one gene pool, we expect a positive correlation between F_{IS} and F_{ST} at individual gene loci (Equation 5), and we expect a positive correlation between r^2 and the product of $F_{ST(1)}$ and $F_{ST(2)}$ for pairs of loci (Equation 6). In the simulated mixtures depicted in Figure 2.1, the expected patterns for both 1- and 2-locus Wahlund effects are evident. In Panel A, the slope of the regression of F_{IS} and F_{ST} (1.015) was close to the value of 1 expected for equal mixture fractions, and the correlation was strongly linear ($r = 0.92$). Theory does not predict a 1:1 slope for r^2 versus $F_{ST(1)}F_{ST(2)}$, but the pattern again was strongly linear ($r = 0.94$). When $F_{ST} = 0$, we do not expect any Wahlund effect on F_{IS}, so in the absence of any other factors causing deviations from HWP the intercept of the regression should be close to 0, and that was the case in both simulations in Figure 2.1.

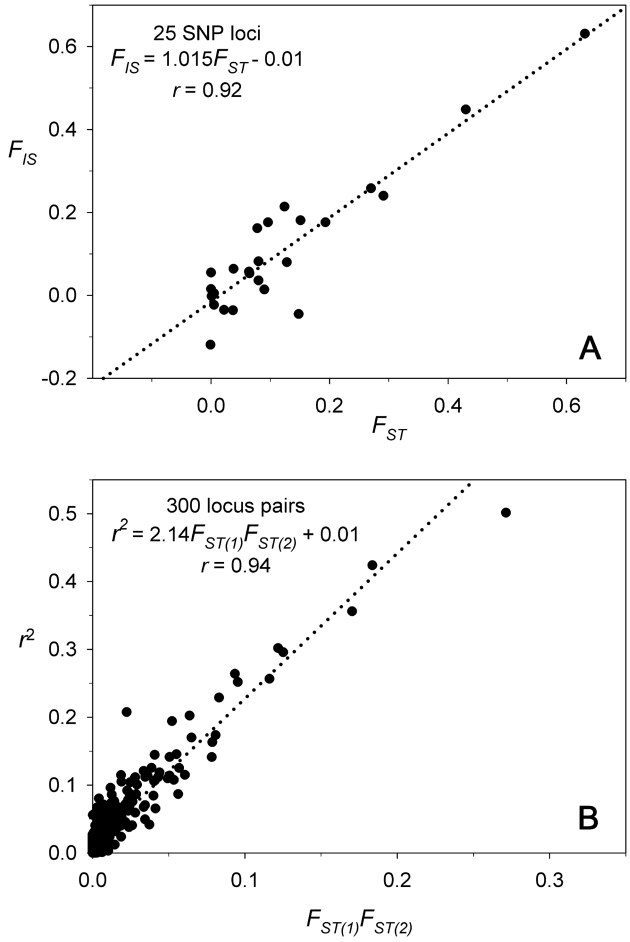

Figure 2.1. Relationship between F_{IS} and F_{ST} at 25 diallelic gene loci (Panel **A**), and between r^2 and the product of $F_{ST(1)}$ and $F_{ST(2)}$ for 300 pairs of the same loci (Panel **B**) in simulated mixtures. Two populations of size $N_e = 100$ were simulated using Easypop (Balloux 2001) and allowed to diverge until mean F_{ST} reached 0.129. FSTAT (Goudet 1995) was used to calculate F_{ST} for each locus. The populations were then combined into a single 200-individual sample; F_{IS} was calculated at each locus using FSTAT and r^2 was calculated at each pair of loci using LDNE (Waples and Do 2008). The dotted lines are the linear regressions, whose formulas are shown.

- Acknowledge the departures but dismiss them because they are few in number after multiple-testing corrections.

Over 3 decades ago, Fairbairn and Roff (1980) noted that "the effective power of the [HWP] test is further reduced by the reluctance of researchers to reject their genetic models even when a significant χ^2 value is obtained," and there is little evidence to suggest the situation is different today. Lessios (1992) went so far as to suggest that perhaps tests of HWP should be dispensed with entirely because results are so consistently ignored.

I would not go that far, but it is clear that many researchers lack a systematic approach to testing HWP and dealing with significant departures. Again I suggest a 2-step protocol: 1) Identify the most likely causes of the departures; and 2) Evaluate whether departures of the nature

Box 2. *Continued*

Figure 2.2 shows a comparable single-locus analysis for empirical data for the Leadbeater's possum. Previous analyses indicated that 2 genetically distinct demes occur in the area sampled. The correlation between F_{IS} and F_{ST} ($r = 0.66$; $P = 0.001$) is strong and positive, as expected for a Wahlund effect. The slope (0.62) is <<1, suggesting that mixture fractions might be uneven (confirmed to be about 3:1; P. Sunnucks, pers. Com). Note that F_{IS} is smaller than F_{ST} for every locus, and the intercept on the F_{IS} axis is negative (−0.1). This suggests that the heterozygote deficiency caused by the Wahlund effect is being partially offset, perhaps by the tendency toward heterozygote excess caused by small N_e—an explanation that seems plausible for this highly endangered Australian marsupial (Hansen et al. 2009).

Other factors that can cause heterozygote deficiencies and positive F_{IS} values (such as self fertilization, null alleles, or allelic dropout) are not expected to produce positive correlations like this. These patterns, however, are subject to stochastic variation. Research is needed to rigorously evaluate the effects of sampling, mixture fraction, and various sources of uncertainty on the strength of these correlations.

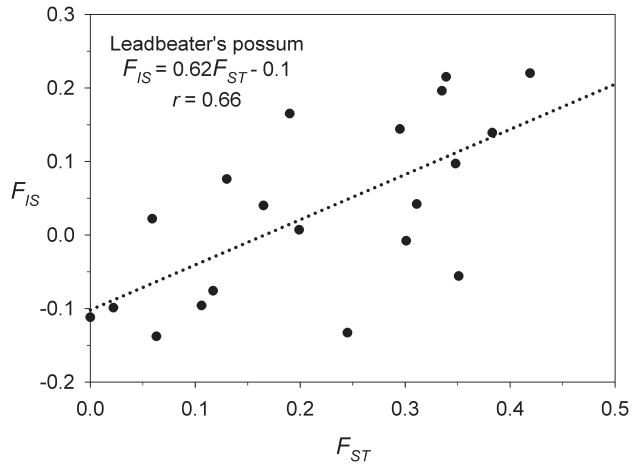

Figure 2.2. Relationship between F_{IS} and F_{ST} at 20 microsatellite loci in a sample of Leadbeater's possum taken at Yellingbo Nature Conservation Reserve in Australia that includes individuals from 2 different demes (Sunnucks and Hansen 2013). The dotted line is the linear regression of F_{IS} on F_{ST}.

and magnitude found are likely to affect conclusions of downstream analyses. These steps are discussed in more detail below.

Identifying Causes of Departures from HWP

Locus- or Sample-Specific Effects

A key to this step is finding answers to 2 questions: Can most or all deviations from HWP be traced to one or a few problem loci, or to one or a few samples? And, do the departures represent heterozygote deficiencies or excesses? Information in Table 1 can help work through this process. To answer the first question, it is useful to construct vectors of test results for each locus (across all populations) and each population (across all loci). For each vector, one can compute the fraction of significant tests and the overall combined P value using Fisher's method. The sequential Bonferroni method can also be useful here for identifying outliers. Factors that are most likely to produce locus-specific HWP deviations include assortative mating, null alleles or genotyping errors/artifacts, and sex linkage (Table 1). Nonrandom sampling could also produce locus-specific effects if heterozygotes have phenotypic traits that affect the likelihood of being sampled. If one or a few problem loci can be identified having departures from HWP that can be attributed to difficulty in recording the true genotype, those loci could be removed from the dataset for subsequent analyses. When doubts remain, robustness of the results can be evaluated by comparing results with and without the loci or samples in question. In doing this screening process, care should be taken to avoid eliminating parts of the dataset that reflect a biological signal of meaningful departures from HW assumptions.

Factors that can cause deviations from HWP only in certain samples include the Wahlund effect, small N_e, genotyping errors (which might depend on sample quality), different allele frequencies in males and females, and self-fertilization. The appropriate response to this type of result will depend on the cause of the departures (see next section). Poor-quality samples that cannot be reliably genotyped should not be used in any analyses. A generalized heterozygote excess attributable to small N_e or a generalized heterozygote deficiency attributable to self-fertilization are natural biological phenomena. Such samples can provide novel biological insights and therefore should not be simply dismissed, although care is needed in evaluating effects of these phenomena on downstream analyses.

Box 3. Simple approaches to multiple testing

A good way to begin evaluation of multiple tests is to perform unadjusted tests, count the fraction that are significant at the nominal α level, and see whether that result is consistent with a global assumption of HWP. The probability that exactly X tests will mistakenly be rejected is given by the binomial distribution, and this can be calculated for a range of values of X using a spreadsheet or similar application. For example, if 200 2-tailed tests are performed with $\alpha = 0.05$, the expected number of Type I errors is 10. To find the 95% confidence interval around this expected value, one looks for values of X for which the cumulative probability of a higher value falls in the range $0.025 < \text{Prob} < 0.975$. In this example (Table 3.1), finding more than 15 significant tests is not consistent with an assumption of global HWP; finding 3 or fewer significant tests would suggest the test might be conservative.

A second approach is to examine the distribution of P values across all tests. If the null hypothesis is true in every case and the test is accurate, the fraction of P values falling in each equally-sized bin on the scale 0–1 should be the same, within the limits of random sampling error. Figure 3.1 illustrates this with simulated genetic data. The combined distribution of P values for the 2 populations analyzed separately (2000 tests total; top panel) is generally consistent with the null expectation (5% = 100 P values should fall in each of 20 bins; solid line), although a slight excess of high P values is apparent. The number of significant or near-significant P values was slightly lower than the null expectation, so results are consistent with a global assumption of HWP. The bottom panel shows results when the 2 populations were combined and treated as a single sample; almost 25% of P values were less than 0.05, and the proportion in the range $0.05 < P < 0.15$ was also elevated from the null expectation. Note, however, that this means that the test failed to detect departures from HWP at more than 75% of the loci, in spite of the equal mixture fractions, large F_{ST}, and large sample sizes. This illustrates the generally weak power of the test for HWP.

Even when the fraction of significant, unadjusted tests is no higher than can be attributed to chance (as in the top panel in Figure 3.1), it is prudent to conduct additional analyses, using standard multiple testing procedures or some of the alternatives discussed in the text and in Box 4, to check for presence of outliers that might have important consequences for downstream analyses.

Table 3.1 The probability that 200 tests of a null hypothesis that is true will produce X or fewer false rejections (Type I errors)

X	Probability	X	Probability
1	**0.0004**	11	0.6998
2	**0.0023**	12	0.7965
3	**0.0090**	13	0.8701
4	0.0264	14	0.9219
5	0.0623	15	0.9556
6	0.1237	16	**0.9762**
7	0.2133	17	**0.9879**
8	0.3270	18	**0.9942**
9	0.4547	19	**0.9973**
10	0.5831	20	**0.9988**

The target Type I error rate was set at $\alpha = 0.05$. Bolded probabilities indicate values of X that are too high to be plausibly explained by chance ($X \geq 16$) or too low to be consistent with the nominal α level ($X \leq 3$).

Heterozygote Deficiency or Excess?

The next step is to determine whether the departures reflect a consistent excess or deficiency of heterozygotes. Factors that commonly produce heterozygote deficiencies include null alleles, the Wahlund effect, self-fertilization, and positive assortative mating (Table 1). If mixed samples are collected from populations with fuzzy boundaries, or are collected on migratory routes or feeding grounds, the Wahlund effect that produces HWP departures and positive F_{IS} will create difficulties in many downstream analyses, so these samples need careful consideration. Most evaluations of the effect of population mixture on tests of HWP and LD fail to take advantage of predictions from population genetics theory about the expected relationships between F_{IS} and F_{ST} (Equation 5) and r^2 and $F_{ST(1)}{*}F_{ST(2)}$ (Equation 6). Researchers finding some loci with heterozygote deficiencies often consider a Wahlund effect but dismiss it because the pattern is not uniform across loci. The widely-used program Microchecker warns of a possible Wahlund effect only if "all loci show an excess of homozygotes" (Oosterhout et al. 2004, p. 537). This is logically flawed. There is no expectation that the Wahlund effect will be equal across loci; instead, there is no effect at all for loci that do not differ in frequency among populations, and more generally theory predicts a positive relationship between F_{IS} and F_{ST}, with the slope being a function of the mixture fraction and allele frequency (see Equation 5 and Box 2). Clustering approaches (e.g., Pritchard et al. 2000; Corander et al. 2008; Jombart et al. 2010) can be used to estimate F_{ST} values if reference samples are not available for populations potentially contributing to the mixture. The approach in Box 2 appears to have considerable promise for identifying a Wahlund effect; however, the patterns shown in Figure 2.1 represent best-case scenarios, as they involve mixtures of equal proportions and assume that pure samples are available to estimate F_{ST}. Much work is needed to rigorously evaluate effects of random sampling error, unequal mixture fractions, and estimation of F_{ST} before the robustness of these approaches can be determined.

Self-fertilization or positive assortative mating lead to inbreeding and also can produce heterozygote deficiencies

Box 3. ***Continued***

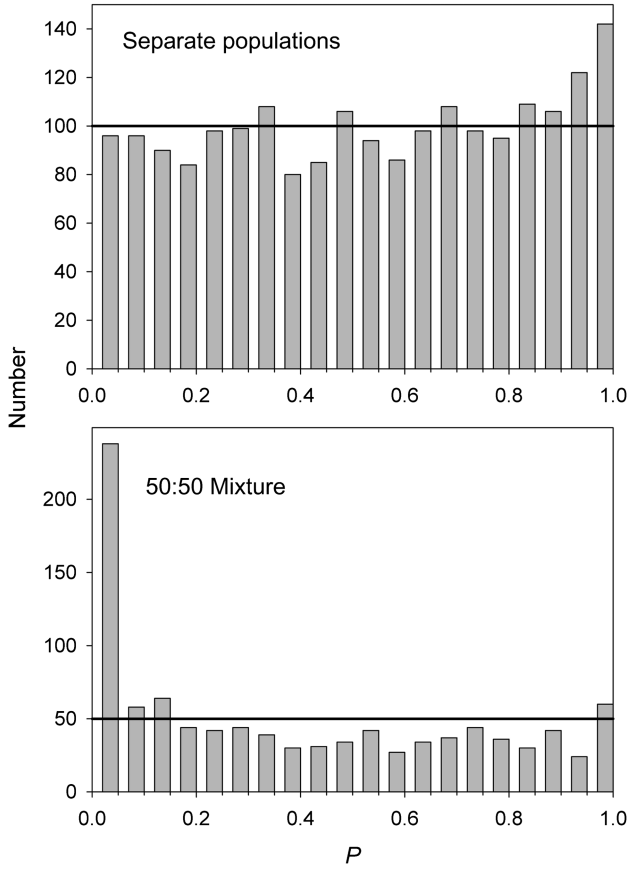

Figure 3.1. Distribution of *P* values for tests of HWP in simulated datasets. Genetic data for 1000 diallelic "SNP" loci were simulated in EasyPop (Balloux 2001). Two populations of $N_e = 1000$ were allowed to diverge until mean F_{ST} reached 0.235; the entire populations then were separately analyzed for departures from HWP using Genepop (Rousset 2008). The combined distribution of *P* values for the 2 individual populations (2000 tests total) is shown in the top panel; the bottom panel shows results for 1000 loci in an equal mixture of the 2 populations using all 2000 individuals.

that might be mistaken for a Wahlund effect. However, the effects of inbreeding on F_{IS} should affect all loci equally, so there is no expectation of a positive correlation between F_{IS} and F_{ST}. Null alleles and some types of scoring errors also produce heterozygote deficiencies, but again under this scenario there is no expectation of a positive correlation between F_{IS} and F_{ST}. If the deviations are locus-specific, that result should emerge from analyses that consider each locus separately and look for patterns across samples, using a sequential Bonferroni or one of the approaches described in Box 2.

If possible, loci showing significant heterozygote excesses should be tested for allele and genotypic frequency differences between the sexes, which can be caused by selection, population history, or sex linkage. Finite population size also is expected to produce an excess of heterozygotes and a negative F_{IS}, but in practice this excess will generally not be detectable unless N_e is tiny, because the inter-locus variance is high. Furthermore, sampling error and other factors can cause opposing effects that reduce or erase the signal from drift. It should be easier to detect finite population size effects on tests of LD because the index r^2 is always positive and effects of drift and other factors such as population mixture are largely additive (Waples and England 2011). Genetic drift should add approximately $1/(3N_e)$ to mean r^2 (Equation 4) and, all else being equal, that should increase the fraction of significant tests above the nominal α level. Surprisingly, this factor is seldom considered in evaluating routine tests of LD, even in papers where the data are later used to estimate effective size using the LD method (which explicitly assumes that LD in excess of the amount expected from sampling error is due to drift). More research is needed to determine exactly how large a problem this is likely to be for tests of LD; however, based on previous work (Waples 1989; Waples and Teel 1990), the probability of a significant test result due to drift increases with the ratio of sample size to effective size.

Selection can produce almost any pattern of agreement with or departure from HWP, so effects are difficult to

Box 4. An example involving hypothetical data

The following hypothetical example illustrates some of the recommendations in this Perspective. Researcher X has genotyped 1000 SNP loci in samples of 100 individuals from each of 10 populations. Of the 10 000 population-by-locus tests of HWP, 630 (6.3%) show significant departures. This is slightly higher than the 5% expected by chance for the nominal Type I error rate ($\alpha = 0.05$)—is that a cause for concern? A good place to start is by examining population-specific effects. That analysis (Table 4.1) shows that nothing is unusual for populations 1–3 and 5–10: the fraction of significant tests ranges from 4.3% to 5.3%, and about half (44–61%) of the significant tests represent heterozygote deficiencies (positive F_{IS}), the remainder being excesses. But in population 4, 153 of the 1000 loci (over 15%) showed significant departures from HWP, and all but 2 of those represent heterozygote deficiencies. A histogram of F_{IS} values for this population (Figure 4.1) shows that the distribution is shifted to the right (toward positive values) compared with the null expectation. From Table 1, the most likely causes of sample-specific heterozygote deficiencies are self-fertilization and population mixture. Information about the biology of the species should provide insights into the plausibility of the former; for the latter, clustering (e.g., Pritchard et al. 2000) or PCA-based methods (e.g., Jombart et al. 2010) could be used to try to partition the sample into component gene pools, after which the analyses describe in Box 2 could be used to evaluate evidence for 1- and 2-locus Wahlund effects.

Even if the population-specific anomalies can be explained, it is still useful to conduct a locus-specific analysis to screen for unusual behavior. This can be done by calculating the number of populations for which each locus had a significant departure and plotting the distribution (Figure 4.2). In this case, the vast majority of loci conform to the null expectation, being found significant in 0–2 populations, but a cluster of 10 loci that deviate significantly from HWP in 5–7 populations are clear outliers. The number of loci showing this behavior is small enough that it did not raise any red flags in considering the overall tests or the population-specific tests. However, the causes of these consistent departures are important to isolate because they could affect downstream analyses. If these locus-specific departures represented heterozygote deficiencies, the most likely explanation might be null alleles or other scoring problems, but in this example the departures from HWP reflect an excess of heterozygotes. A tiny N_e could produce a generalized excess of heterozygotes but would not be expected to produce such distinctive results for just a few loci. A result like this, however, could occur if the loci in question are tightly linked to genes involved in sex determination, in which case allele frequencies could differ between males and females, leading to an excess of heterozygotes in their offspring.

Table 4.1 Hypothetical results for 2-tailed statistical tests of agreement with HWP conducted for 1000 loci in each of 10 population samples

		Heterozygote		
Population	Significant tests	Deficiency	Excess	% Deficiency
1	53	31	22	58.5
2	46	28	18	60.9
3	43	22	21	51.2
4	153	151	2	98.7
5	51	24	27	47.1
6	48	25	23	52.1
7	51	30	21	58.8
8	43	19	24	44.2
9	53	27	26	50.9
10	49	29	20	59.2

evaluate with any generality. In addition, even strong selection can produce genotypic frequencies that do not deviate from HWP (Box 1). Lachance (2009) evaluated power to detect selection-induced departures from HWP; selection for or against heterozygotes is most likely to produce HW departures (Table 1). These patterns can be locus-specific, but heterosis (which favors genome-wide heterozygosity) can produce a general pattern of heterozygote excess. Searching for "outlier" loci that have high F_{ST} values indicative of strong directional selection has become a popular pastime. It might seem incongruous to do this for loci that do not show any evidence of HW departures, but the 2 tests evaluate different processes. HW genotypic frequencies are affected only by selection occurring within a generation within 1 population, while outlier loci are produced by different selective regimes in different populations across multiple generations. Therefore, HW tests generally provide little information of relevance to tests for candidate genes for adaptations.

Substantial LD or deviations from HWP that cannot be explained by chance or by any of the above factors should alert the researcher to potential quality-control problems in data collection or recording that merit careful scrutiny before proceeding further. The researcher should revisit all aspects of the process of generating the genetic data to see if problems can be identified. Morin et al. (2009) showed that genotyping errors that create even a single apparent homozygote for a rare allele can cause significant HWP departures, so identifying influential individuals can be a useful exercise. Box 4 uses hypothetical data to illustrate how some of the above suggestions can be used to evaluate results of tests of HWP.

Evaluating the Biological Consequences of HW Departures

Armed with information about the likely causes of HW departures, the researcher can tackle the second major step:

Box 4. ***Continued***

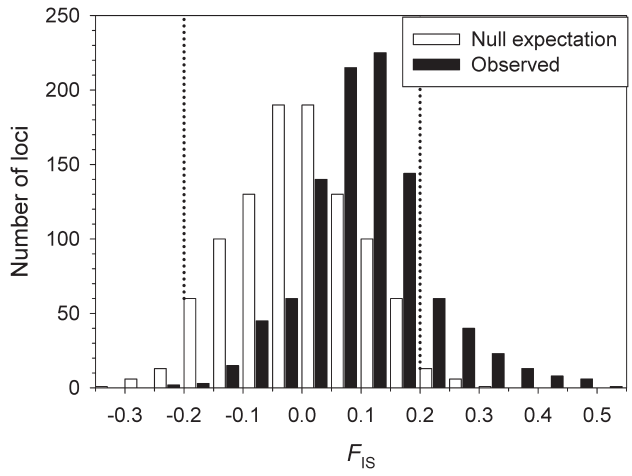

Figure 4.1. Hypothetical observed distribution of single-locus F_{IS} values at 1000 loci scored in a single sample of 100 individuals, compared with the approximate null distribution assuming all Hardy–Weinberg assumptions are met (hence mean $F_{IS} = 0$). Vertical dotted lines are approximate bounds for the 95% confidence interval for F_{IS} under the null hypothesis.

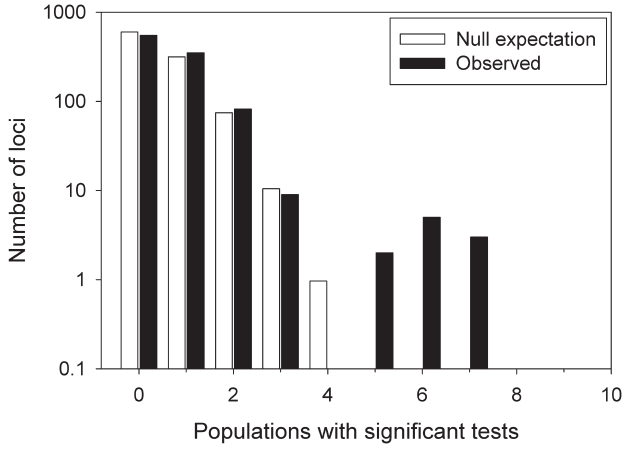

Figure 4.2. A simple way to evaluate locus-specific effects. In this hypothetical example, each of 1000 loci were tested for departures from HWP in each of 10 populations. The binomial distribution, assuming 10 independent trials for each locus, each having an $\alpha = 0.05$ probability of rejection, was used to generate the null expectation for the numbers of loci that would have 0 … 10 significant departures just by chance. In the observed data, a total of 10 loci had significant departures in 5 or more populations—a very unlikely result for any locus that actually conformed to Hardy–Weinberg principles. Note the log scale on the *Y* axis.

evaluating the consequences of using the offending loci/samples in downstream analyses. Although statistical tests play a key role in identifying the departures, when sample sizes are relatively large, significant departures from HWP could occur that have little biological importance (Waples 1998; Hedrick 1999). It is therefore important to consider not only the *P* value for each test, but the absolute magnitude of the departure (the effect size). The key question then becomes, Are departures from HWP of this magnitude likely to influence biological conclusions that emerge from downstream analyses? For example, after consideration of this issue, a researcher might reasonably argue that the departures can be ignored because the downstream results are robust to the nature and magnitude of the deviations.

How does one know when this is a "reasonable" argument? Unfortunately, it is not possible to provide much concrete guidance on this crucial issue. Pompanon et al. (2005) noted the nearly complete lack of evaluations of the consequences of scoring errors for population genetics studies. Some recent studies have begun to chip away at this huge block of

Table 1 The most common causes of LD and departures from HWP

Observation/possible cause	Locus[a] specific?	Sample specific?	Comments
Positive F_{IS}			
Positive assortative mating	Yes	Perhaps	Effect is expected only if phenotype is correlated with genotype
Self fertilization	No	Yes	
Wahlund effect	Yes[b]	Yes	F_{IS} should be positively correlated with F_{ST}[b]
True null alleles	Yes	No	
Apparent null alleles	Yes	Perhaps[c]	Could depend on sample quality[c]
Nonrandom sampling	Perhaps	Perhaps	Expected if heterozygotes are less likely to be sampled
Underdominance	Yes	Perhaps	Selection against heterozygotes
Negative F_{IS}			
Negative assortative mating	Yes	Perhaps	Effect is expected only if phenotype is correlated with genotype
Nonrandom sampling	Perhaps	Perhaps	Expected if heterozygotes are more likely to be sampled
Overdominance	Perhaps[d]	Perhaps	Selection favors heterozygotes
Selection differs in M and F	Yes	Yes	Allele frequency differences between sexes cause het excess
Sex linkage	Yes	Perhaps[e]	Allele frequency differences between sexes cause het excess
Small N_e	No	Yes	
LD; r^2 significantly > 0			
Small N_e	No	Yes	
Wahlund effect	Yes[f]	Yes	r^2 should be positively correlated with $F_{ST(1)}*F_{ST(2)}$[f]
Epistasis	Yes	Perhaps	A wide range of patterns is possible
Hitchhiking	Yes	Perhaps	

[a]Recognizing that random variation will occur among loci, even if "No" is indicated in this column.

[b]See "Comments" column.

[c]See "Comments" column.

[d]If heterozygote advantage is due to general heterosis, locus-specific effects are not expected.

[e]See Marshall et al. (2004) for an example of sample-specific departures from HWP due to sex linkage.

[f]See "Comments" column.

uncertainty associated with the broader topic of the biological consequences of departures from HWP and linkage equilibrium: effects of population mixture (Deng et al. 2001) and genotyping errors (Terwilliger et al. 1990) on association mapping; effects of genotyping errors on measures of LD (Akey et al. 2001); effects of null alleles (Dakin and Avise 2004) and other errors (Wang 2010) on parentage analysis; effects of microsatellite null alleles on estimates of inbreeding (Barker 2005) and genetic differentiation (Chapuis and Estoup 2007); impact of HW departures on gene-disease associations (Trikalinos et al. 2006); biases associated with different fixed quality cutoffs for genotype calls (Johnson and Slatkin 2008); and effects of RAD scoring errors on estimates of genetic diversity (Gautier et al. 2013; Arnold et al. 2013). Much work remains to expand this type of quantitative evaluation and synthesize the results in a way that provides practical guidelines for researchers interested in using genetic methods to study nonmodel species. In the meantime, it seems reasonable to expect that researchers who find substantial LD or departures from HWP that cannot be attributed to chance, but nevertheless want to use the data in downstream analyses, should provide an explanation of why they believe the deviations are not likely to affect subsequent conclusions or interpretations.

Emerging Issues for the Genomics Era

Some issues related to genomics data merit mention here. First, no simple, universal way exists to extract accurate genotypes from raw next-generation sequencing (NGS) data. A researcher who wants to obtain such data for a new nonmodel species must either make a variety of decisions about how to filter and package the raw data, or must delegate that job to someone else (Nielsen et al. 2011; Andrews and Luikart 2014). It is vital that researchers take an active interest in this process or they will not be able to vouch for or even understand key aspects of their data. One sobering fact that perhaps few researchers know or have paid attention to is that some popular programs for analyzing NGS data use a HW prior to call genotypes (see Andrews and Luikart 2014). This might be reasonable in well-studied species where HW assumptions have been independently verified, but could create serious problems for many nonmodel species. If HWP are assumed in calling the genotypes in the first place, any subsequent tests of conformity to HWP are of questionable value, and downstream analyses could be affected if the population is not at HWP for a biological reason (as opposed to a scoring artifact). Researchers studying nonmodel species would do better to find methods that can reliably call genotypes using independent criteria.

A second issue arises from the fact that NGS datasets can easily include thousands of genetic markers. As long as only a handful of allozymes or microsatellite loci were involved, it was convenient to assume that all markers are unlinked. This assumption does not pass the red-face test with genomics datasets: In real organisms, thousands of gene loci have

to be packaged into at most a few dozen chromosomes. As a consequence, we can expect that NGS datasets will contain pairs of loci that span a wide range of linkage relationships. Indices such as r^2 that are sensitive to physical linkage will be affected, and analyses that depend on such indices will be biased unless they account for the linkage.

Finally, standard multiple-testing procedures become difficult with genomics datasets that can have 10^4 or more markers even for nonmodel species. Default settings for software that tests for HWP by permutation and applies a Bonferroni correction might not be adequate to distinguish datasets that do and do not have P values $<\alpha/10^4$. Bonferroni-like corrections are hopeless for pairwise tests of LD, of which there are $\sim 10^8/2$ for a dataset with 10^4 loci. Furthermore, the physical linkage mentioned above produces redundancies in information content, which reduces precision. This latter issue is important to consider in the context of tests of HWP and LE, because the tests are not independent when linkage is present. Nyholt (2004) and Xu (2012) suggested procedures to deal with some of these issues for multiple tests of LD in human genomics data, but the empirical examples they used involved only small numbers of markers within particular chromosomal regions. More work is needed to develop rigorous procedures for dealing with the huge numbers of simultaneous tests of HWP and LD that will be commonplace in genome-wide studies of nonmodel species. In the meantime, some of the simple procedures discussed above and in Box 3 that look for broad patterns in the data across loci or pairs of loci can provide useful insights.

Acknowledgments

I am grateful to Fred Allendorf for inviting this Perspective. I thank Fred, Tiago Antao, Phil Hedrick, Bill Hill, and Paul Sunnucks for valuable discussions and Paul, Phil, Mike Ford, Morten Limborg, Shawn Narum, Krista Nichols, Ryan Waples, Eric Ward, and 2 anonymous reviewers for comments that considerably improved the manuscript.

Appendix. The Wahlund effect, F_{IS}, and F_{ST}

Wahlund (1928) first described the consequences for genotypic proportions of having a sample that includes individuals from more than one random mating population: homozygotes occur more frequently, and heterozygotes less frequently, than would be expected under conditions of HWP. The deficiency of heterozygotes caused by this scenario is widely known as the Wahlund effect. To illustrate the effect, assume a 2-population mixture that includes populations 1 and 2 in proportions m and $(1 - m)$. Let the frequencies of allele A at a diallelic locus be p_1 in population 1 and p_2 in population 2, so $\bar{p}_w = mp_1 + (1 - m)p_2$ = the weighted mean of p_1 and p_2. Wahlund (1928) used the following notation: mixture fractions = g and h, with $g + h = 1$; allele frequencies in the 2 populations are r_g and r_h. Translating these into the current notation produces $g = m$, $h = 1 - m$, $r_g = p_1$, and $r_h = p_2$. With these conversions, the result from Wahlund's Table 3 for the expected frequency of heterozygotes in a mixed sample can be expressed as follows:

$$\text{Freq}(Aa) = 2\bar{p}_w(1-\bar{p}_w) - 2m(1-m)(p_1-p_2)^2. \quad \text{(A.1)}$$

If we note that $(p_1 - p_2)^2 = 4\text{Var}(p)$, where $\text{Var}(p)$ is the variance of p among populations, then Equation A.1 can be written as

$$\text{Freq}(Aa) = 2\bar{p}_w(1-\bar{p}_w) - 4m(1-m) * 2\text{Var}(p). \quad \text{(A.2)}$$

The term on the left in Equation A.2 $[2\bar{p}_w(1-\bar{p}_w)]$ is the expected HW frequency of heterozygotes in the mixed sample, and the term on the right is the amount by which the frequency of heterozygotes is reduced by nonrandom mating (i.e., the Wahlund effect). Formulations similar to equation A.2 (but which typically assume equal mixture fractions, in which case $4m(1 - m) = 1$) can be found in a number of contemporary references (e.g., Frankham et al. 2002, p. 322; Allendorf et al. 2013, p. 159).

Dividing each side of Equation A.2 by $2\bar{p}_w(1-\bar{p}_w)$ produces an interesting result:

$$\begin{aligned}\frac{\text{Freq}(Aa)}{2\bar{p}_w(1-\bar{p}_w)} &= \frac{2\bar{p}_w(1-\bar{p}_w) - 4m(1-m) * 2\text{Var}(p)}{2\bar{p}_w(1-\bar{p}_w)} \\ &= 1 - 4m(1-m)\frac{\text{Var}(p)}{\bar{p}_w(1-\bar{p}_w)}.\end{aligned} \quad \text{(A.3)}$$

The term on the left of Equation A.3 is the ratio of observed to expected (HWP) heterozygosity, $H_o/H_e = 1 - F_{IS}$, while the term on the far right is similar to F_{ST} but has weighted terms in the denominator. If we let $\bar{p}_w(1-\bar{p}_w) = C\bar{p}(1-\bar{p})$, where $\bar{p} = (p_1 + p_2)/2$ = the unweighted mean of p_1 and p_2, the above equation can be written as

$$1 - F_{IS} = 1 - \frac{4m(1-m)}{C}\frac{\text{Var}(p)}{\bar{p}(1-\bar{p})} = 1 - \frac{4m(1-m)}{C}F_{ST}. \quad \text{(A.4)}$$

This implies that

$$E(F_{IS}) = F_{ST}[4m(1-m)/C], \quad \text{(A.5)}$$

where

$$\begin{aligned}C &= \bar{p}_w(1-\bar{p}_w)/\bar{p}(1-\bar{p}) \\ &= \frac{[mp_1 + (1-m)p_2][1-(mp_1 + (1-m)p_2)]}{\frac{p_1+p_2}{2}\left[1 - \frac{p_1+p_2}{2}\right]}\end{aligned} \quad \text{(A.6)}$$

is the ratio of a functions of the weighted and unweighted mean allele frequencies. We can consider some special cases:

- If $m = 0.5$ (equal mixture fractions), then $4m(1 - m) = 1$ and C also is 1, so $E(F_{IS}) = F_{ST}$; and
- If $p_1 = 0$ and $p_2 = 1$ or vice versa (populations are fixed for different alleles), then $C = 4m(1 - m)$, so $E(F_{IS}) = F_{ST}$.

In these 2 cases, therefore, we expect a linear relationship between F_{IS} and F_{ST} with a slope of 1. In other cases ($m \neq 0.5$ and alleles A and a both segregating in at least one population), the ratio $4m(1-m)/C$ will in general not equal 1. For any given mixture sample, $4m(1-m)$ will be the same at all loci, but C will vary across loci with allele frequency. This means that the expected relationship between F_{IS} and F_{ST} will not be perfectly linear, although we still expect a positive correlation between the 2 parameters.

The derivations above treat the means, variances, and F-statistics as population parameters rather than estimates based on finite samples. Although a rigorous evaluation of the effects of sampling is needed to determine the strength of these relationships in actual populations, some empirical examples are shown in Box 2 (main text).

In an evaluation of the inter-locus variance in the inbreeding coefficient, Robertson and Hill (1984) derived the following relationship:

$$E(f) = F_{ST}[(n-1)/n], \qquad \text{(A.7)}$$

where n is the number of subpopulations used to compute F_{ST} and f is comparable to F_{IS} computed over a combined sample. Although this is not explicitly stated in the text, equation A.7 assumes equal mixture fractions (Hill W, personal communication). If we back out the $n/(n-1)$ adjustment to account for computing $Var(p)$ across a finite number of subpopulations, equation A.7 produces the same result as equation A.5 for an equal mixture of 2 populations.

Funding

R.S.W. was supported by funds from the National Marine Fisheries Service, National Oceanic and Atmospheric Association.

References

Akey JM, Zhang K, Xiong M, Doris P, Jin L. 2001. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. Am J Hum Genet. 68:1447–1456.

Allendorf FW, Luikart G, Aitken SN. 2013. Conservation and the genetics of populations. 2nd ed. Oxford (UK): Wiley-Blackwell.

Andrews KR, Luikart G. 2014. Recent novel approaches for population genomics data analysis. Mol Ecol. 23:1661–1667.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol. 22:3179–3190.

Balloux F. 2001. EASYPOP (version 1.7): a computer program for population genetics simulations. J Hered. 92:301–302.

Balloux F. 2004. Heterozygote excess in small populations and the heterozygote-excess effective population size. Evolution. 58:1891–1900.

Barker JS. 2005. Population structure and host-plant specialization in two *Scaptodrosophila* flower-breeding species. Heredity (Edinb). 94:129–138.

Barton NH. 2000. Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci. 355:1553–1562.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 57:289–300.

Benjamini Y, Yekutieli D. 2001. The control of false discovery rate under dependency. Ann Stat. 29:1165–1188.

Berlocher SH. 1984. Genetic changes coinciding with the colonization of California by the walnut husk fly, *Rhagoletis completa*. Evolution. 38:906–918.

Chapuis MP, Estoup A. 2007. Microsatellite null alleles and estimation of population differentiation. Mol Biol Evol. 24:621–631.

Corander J, Marttinen P, Sirén J, Tang J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics. 9:539.

Cox DG, Kraft P. 2006. Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. Hum Hered. 61:10–14.

Crow JF. 1988. Eighty years ago: the beginnings of population genetics. Genetics. 119:473–476.

Dakin EE, Avise JC. 2004. Microsatellite null alleles in parentage analysis. Heredity (Edinb). 93:504–509.

Deng HW, Chen WM, Recker RR. 2001. Population admixture: detection by Hardy-Weinberg test and its quantitative effects on linkage-disequilibrium methods for localizing genes underlying complex traits. Genetics. 157:885–897.

Doyle RW. 2014. Inbreeding and disease in tropical shrimp aquaculture: a reappraisal and caution. Aqua. Res. Advance Access published May 9, 2014, doi:10.1111/are.12472.

Emigh TH. 1980. A comparison of tests for Hardy-Weinberg equilibrium. Biometrics. 36:627–642.

Excoffier L, Heckel G. 2006. Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet. 7:745–758.

Fairbairn, DJ, Roff DA. 1980. Testing genetic models of isozyme variability without breeding data: Can we depend on the χ^2? Can J Fish Aquat Sci. 37:1149–1159.

Felsenstein J. 1965. The effect of linkage on directional selection. Genetics. 52:349–363.

Frankham R, Briscoe DA, Ballou JD. 2002. Introduction to conservation genetics. Cambridge (UK): Cambridge University Press.

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet JM, Estoup A. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol Ecol. 22:3165–3178.

Goudet J. 1995. FSTAT version 1.2: a computer program to calculate *F*-statistics. J Heredity. 86:485–486.

Guo SW, Thompson EA. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics. 48(2):361–372.

Haldane JBS. 1954. An exact test for randomness of mating. J Genet. 52:631–635.

Hansen TF. 2013. Why epistasis is important for selection and adaptation. Evolution. 67:3501–3511.

Hansen BD, Harley DK, Lindenmayer DB, Taylor AC. 2009. Population genetic analysis reveals a long-term decline of a threatened endemic Australian marsupial. Mol Ecol. 18:3346–3362.

Hansen MM, Nielsen EE, Mensberg K-LD. 1997. The problem of sampling families rather than populations: relatedness among individuals in samples of juvenile brown trout *Salmo trutta* L. Mol Ecol. 6:469–474.

Hardy HG. 1908. Mendelian proportions in a mixed population. Science. 28:49–50.

Hedgecock D, Launey S, Pudovkin AI, Naciri Y, Lapègue S, Bonhomme F. 2007. Small effective number of parents (Nb) inferred for a naturally spawned cohort of juvenile European flat oysters *Ostrea edulis*. Mar Biol. 150:1173–1183.

Hedrick PW. 1999. Perspective: Highly variable loci and their interpretation in evolution and conservation. Evolution. 53:313–318.

Hedrick PW. 2000. Genetics of populations. 2nd ed. Sudbury (MA): Jones and Bartlett.

Hill WG. 1974. Estimation of linkage disequilibrium in randomly mating populations. Heredity (Edinb). 33:229–239.

Hill WG. 1981 Estimation of effective population size from data on linkage disequilibrium. Genet Res (Camb). 38:209–216.

Holm S. 1979. A simple sequential rejective multiple test procedure. Scand J Stat. 6:65–70.

Jankovic I, vonHoldt BM, Rosenberg NA. 2010. Heterozygosity of the Yellowstone wolves. Mol Ecol. 19:3246–3249.

Johnson PL, Slatkin M. 2008. Accounting for bias from sequencing error in population genetic estimates. Mol Biol Evol. 25:199–206.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics. 11:1471–2156.

Jones B. 2012. Transboundary movement of shrimp viruses in crustaceans and their products: a special risk? J Invertebr Pathol. 110:196–200.

Kimura M. 1956. A model of a genetic system which leads to closer linkage by natural selection. Evolution. 10:278–287.

Koehn RK. 1972. Genetic variation in the eels: a critique. Mar Biol. 14:179–181.

Koehn RK, Williams GC. 1978. Genetic differentiation without isolation in the American eel, *Angiuilla rostrata*. II. Temporal stability of geographic patterns. Evolution. 27:192–204.

Kulinskaya E, Lewin A. 2009. Testing for linkage and Hardy-Weinberg disequilibrium. Ann Hum Genet. 73:253–262.

Lachance J. 2009. Detecting selection-induced departures from Hardy-Weinberg proportions. Genet Sel Evol. 41:15

Lessios HA. 1992. Testing electrophoretic data for agreement with Hardy-Weinberg expectations. Mar Biol. 112:517–523.

Levene H. 1949. On a matching problem arising in genetics. Ann Math Stat. 21:91–94.

Lewontin RC, Cockerham CC. 1959. The goodness-of-fit test for detecting natural selection in random mating populations. Evolution. 13:561–564.

Lewontin RC, Hubby JL. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics. 54:595–609.

Li CC. 1988. Pseudo-random mating populations. In celebration of the 80th anniversary of the Hardy-Weinberg law. Genetics. 119:731–737.

Louis EJ, Dempster ER. 1987. An exact test for Hardy-Weinberg and multiple alleles. Biometrics. 43:805–811.

Marshall AR, Knudsen KL, Allendorf FW. 2004. Linkage disequilibrium between the pseudoautosomal PEPB-1 locus and the sex-determining region of chinook salmon. Heredity (Edinb). 93:85–97.

May RM. 2004. Uses and abuses of mathematics in biology. Science. 303:790–793.

Miller RG Jr. 1981. Simultaneous statistical inference. New York: McGraw Hill.

Morin PA, Leduc RG, Archer FI, Martien KK, Huebinger R, Bickham JW, Taylor BL. 2009. Significant deviations from Hardy-Weinberg equilibrium caused by low levels of microsatellite genotyping errors. Mol Ecol Resour. 9:498–504.

Narum SR. 2006. Beyond Bonferroni: Less conservative analyses for conservation genetics. Conserv Genet. 7:783–787.

Nei M, Li WH. 1973. Linkage disequilibrium in subdivided populations. Genetics. 75:213–219.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 12:443–451.

Nyholt DR. 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 74:765–769.

Oosterhout CV, Hutchinson WF, Wills DPM, Shipley P. 2004. Micro-Checker: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes. 4:535–538.

Palstra FP, Fraser DJ. 2012. Effective/census population size ratio estimation: a compendium and appraisal. Ecol Evol. 2:2357–2365.

Pompanon F, Bonin A, Bellemain E, Taberlet P. 2005. Genotyping errors: causes, consequences and solutions. Nat Rev Genet. 6:847–859.

Powell JR. 1975. Protein variation in natural populations of animals. Evol Biol. 8:79–119.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Pudovkin AI, Zaykin DV, Hedgecock D. 1996. On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. Genetics. 144:383–387.

Rice WR. 1989. Analyzing tables of statistical tests. Evolution. 43:223–225.

Rice WR. 1996. Evolution of the Y sex chromosome in animals. BioScience. 46:331–343.

Robertson A. 1965. The interpretation of genotypic ratios in domestic animal populations. Anim Prod. 7:319–324.

Robertson A, Hill WG. 1984. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics. 107:703–718.

Rousset F. 2008. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour. 8:103–106.

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. Nature. 473:337–342.

Sinnock P. 1975. The Wahlund effect for the two-locus model. Am Nat. 109:565–570.

Sunnucks P, Hansen BD. 2013. Guest Box 5. In: Allendorf FW, Luikart G, Aitken SN, editors. Conservation and the genetics of populations. 2nd ed. Oxford (UK): Wiley-Blackwell. p. 93–95.

Terwilliger JD, Weeks DE, Ott J. 1990. Laboratory errors in the reading of marker alleles cause massive reductions in LOD score and lead to gross overestimation of the recombination fraction. Am J Hum Genet. 47:A201.

Trikalinos TA, Salanti G, Khoury MJ, Ioannidis JP. 2006. Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations. Am J Epidemiol. 163:300–309.

Utter FM, Hodgins HO, Allendorf FW. 1974. Biochemical genetic studies of fishes: potentialities and limitations. In: Sargeant JK, editor. Biochemical and biophysical perspectives in marine biology. Vol. 1. New York: Academic Press. p. 213–238.

Utter F, Aebersold P, Winans G. 1987. Interpreting genetic variation detected by electrophoresis. In: Ryman N, Utter F, editors. Population genetics and fishery management. Seattle (WA): University of Washington Press. p. 21–45.

van der Loo W, Arthur CP, Richardson BJ, Wallage-Drees M, Hamers R. 1987. Nonrandom allele associations between unlinked protein loci: are the polymorphisms of the immunoglobulin constant regions adaptive? Proc Natl Acad Sci U S A. 84:3075–3079.

Vithayasai C. 1973. Exact critical values of the Hardy-Weinberg test statistic for two alleles. Communic Stat. 1:229–242.

Wahlund S. 1928. Zusammensetzung von population und korrelationserscheinung vom stand-punkt der vererbungslehre aus betrachtet. Hereditas. 11:65–106 [English translation. In: Weiss KM, Ballonoff PA, editors. 1975. Demographic Genetics. Dowden, Hutchinson and Ross, Stroudsburg. p. 224–263].

Wallace B. 1958. The comparison of observed and calculated zygotic distributions. Evolution. 12:113–115.

Wang J. 2010. Effects of genotyping errors on parentage exclusion analysis. Mol Ecol. 19:5061–5078.

Waples RS. 1989. Temporal stability of allele frequencies: testing the right hypothesis. Evolution. 43:1236–1251.

Waples RS. 1998. Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species. J Heredity. 89:438–450.

Waples RS, Antao T, Luikart G. 2014. Effects of overlapping generations on linkage disequilibrium estimates of effective population size. Genetics. 197:769–780.

Waples RS, Do C. 2008. LDNe: a program for estimating effective population size from data on linkage disequilibrium. Mol Ecol Resour. 8:753–756.

Waples RS, England PR. 2011. Estimating contemporary effective population size based on linkage disequilibrium in the face of migration. Genetics. 189:633–644.

Waples RS, Teel DJ. 1990. Conservation genetics of Pacific salmon. I. Temporal changes in allele frequency. Conserv Biol. 4:144–156.

Weinberg W. 1908. On the demonstration of heredity in man. In: Boyer SH, trans (1963). Papers on human genetics. Englewood Cliffs (NJ): Prentice Hall.

Weinberg W. 1909. Über Vererbungsgesetze beim Menschen. Z. Indukt. Abstammungs Vererbungsl. 1:277–330. [citation from Crow 1988].

Weir BS. 1979. Inferences about linkage disequilibrium. Biometrics. 35:235–254.

Weir BS. 1996. Genetic data analysis II. Sunderland (MA): Sinauer Associates.

Whitlock MC. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol. 18:1368–1373.

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 76:887–893.

Williams GC, Koehn RK, Mitton JB. 1973. Genetic differentiation without isolation in the American eel, *Angiuilla rostrata*. Evolution. 27:192–204.

Wright S. 1951. The genetical structure of populations. Ann Eugen. 15:323–354.

Xu S. 2012. Testing Hardy-Weinberg disequilibrium using the generalized linear model. Genet Res (Camb). 94:319–330.

Zaykin DV, Pudovkin A, Weir BS. 2008. Correlation-based inference for linkage disequilibrium with multiple alleles. Genetics. 180:533–545.

Received April 7, 2014; First decision June 19, 2014;
Accepted August 26, 2014

Corresponding Editor: Fred Allendorf