Heredity (2015) 114, 249–254
© 2015 Macmillan Publishers Limited All rights reserved 0018-067X/15
www.nature.com/hdy

ORIGINAL ARTICLE

A parameter to quantify the degree of genetic mixing among individuals in hybrid populations

ST Kalinowski and JH Powell

Hybridization between genetically distinct taxa is a complex evolutionary process. One challenge to studying hybrid populations is quantifying the degree to which non-native genes have become evenly mixed among individuals in the population. In this paper, we present a variance-based parameter, m_d, that measures the degree to which non-native genes are evenly distributed among individuals in a population. The parameter has a minimum value of 0 for populations in which individuals from multiple taxa are present but have not interbred, and a maximum value of 1 for populations in which all individuals have the same amount of non-native ancestry. A recurrence equation showed that relatively few generations of random mating are required for m_d to approach 1 (indicating a well-mixed population), and that m_d is independent of initial amounts of non-native ancestry. The parameter is mathematically equivalent to F_{ST} and we show how existing formulae for F_{ST} can be used to estimate m_d when diagnostic loci are available. Computer simulations showed this estimator to have very little bias for realistic amounts of data.
Heredity (2015) **114,** 249–254; doi:10.1038/hdy.2014.93; published online 12 November 2014

INTRODUCTION

Hybridization between genetically distinct taxa is a complex ecological process that has important implications for a diverse array of population genetic questions. For example, medical geneticists study admixed human populations to identify and map genes causing diseases (for example, Mao *et al.*, 2007; Cheng *et al.*, 2010; Winkler *et al.*, 2010). Ecological geneticists study admixed populations to understand how outbreeding affects survival and reproduction (for example, Hogg *et al.*, 2006; Johnson *et al.*, 2010; Vander Wal *et al.*, 2012). Conservation geneticists study admixed populations to manage the spread of introgressive hybridization (for example, Rhymer and Simberloff, 1996; Hedrick, 2009; Muhlfeld *et al.*, 2009). Agricultural geneticists and environmental activists monitor wild populations to detect genes from genetically modified organisms (for example, Watrud *et al.*, 2004; Piñeyro-Nelson *et al.*, 2009; Zapiola and Mallory-Smith, 2012). And finally, evolutionary biologists study hybrid zones to better understand how selection, gene flow and mate choice shape the genetic structure of natural populations (for example, Barton and Hewitt, 1985).

Hybridization often begins when non-native individuals enter a population and mate with individuals from a native taxon. At this point, the population may have a 'bimodal' distribution of genotypes (Harrison and Bogdanowicz, 1997). Such a population might consist of mostly genetically 'pure' individuals of both taxa, and, potentially, a few hybrid individuals. For some taxa, this will be as far as hybridization proceeds (for example, Steeves *et al.*, 2010). However, if hybrids are fertile and pre-zygotic isolating mechanisms are weak, hybridization may continue until all individuals in the population are hybrids (Rhymer *et al.*, 1994). At this point, ecologists sometimes call the population a 'hybrid swarm' (for example, Allendorf *et al.*, 2001). If interbreeding continues, the distribution of non-native genes among individuals in the population will become 'unimodal' and will eventually approach the point in which all individuals have the same amount of non-native genes—and the mixing can be viewed as complete.

Genetic data are often used to quantify the amount of non-native genes in admixed populations and the degree to which these genes have become mixed in the population. The analysis of such data is relatively straightforward when there are fixed genetic differences between the taxa (for example, Rhymer *et al.*, 1994), and sophisticated methods are available to study hybridization when fixed genetic differences between taxa are not present (for example, Pritchard *et al.*, 2000; Anderson and Thompson, 2002; Hey, 2010). Therefore, it is usually relatively straightforward to use genetic data to estimate the ancestry of each individual in a hybrid population.

One challenge to interpreting such data is the potentially complex distribution of different levels of hybridity among individuals in the population. As discussed above, the amount of non-native genes present in individuals sometimes varies widely among the individuals in a population. Quantifying this variability is relevant to many analyses, but there is no widely accepted way to do this. A common practice is to report the total proportion, *P*, of non-native genes in a population and present a graph showing the distribution of non-native genes among all individuals (for example, Pertoldi *et al.*, 2010). This approach conveys a lot of information about the distribution of non-native genes among individuals in a population, but makes it difficult to compare the degree of genetic mixing among different populations.

Vernesi *et al.* (2003) developed a parameter, which they called the 'true hybridization index' (THI), that solves this problem by quantifying how well the genes of multiple taxa are mixed in a population. THI has a range of 0–1, with 0 indicating that no mixing has occurred

Department of Ecology, Montana State University, Bozeman, MT, USA
Correspondence: Professor ST Kalinowski, Department of Ecology, Montana State University, 310 Lewis Hall, Bozeman, MT 59717, USA.
E-mail: skalinowski@montana.edu
Received 24 April 2013; revised 17 July 2014; accepted 21 July 2014; published online 12 November 2014

(all individuals are genetically pure), and 1 indicating that the population is thoroughly mixed (all individuals have the same amount of ancestry from each contributing taxon).

The purpose of this present investigation is to extend Vernesi *et al.* (2003) work in several ways. First, we show how their genetic mixing parameter (which we call the 'degree of genetic mixing') can be derived in a simple, biologically meaningful way. This new definition allows us to show how the amount of genetic mixing in a population is related to the amount of gametic disequilibrium present in a population, and to show how the parameter increases in a randomly mating population. We also present a nearly unbiased formula for estimating this parameter when diagnostic loci are available and discuss how best to estimate the parameter when diagnostic loci are not available.

METHODS AND RESULTS

A genetic mixing parameter, m_d

We seek to derive a parameter (We use 'parameter' to refer to a numeric characteristic of an entire population and 'statistic' to refer to a quantity calculated from a sample (Everitt and Skrondal, 2010)) that quantifies the degree of genetic mixing among taxa in a population and has a range of 0 to 1. We will derive such a parameter here and compare it to the parameter THI of Vernesi *et al.* (2003) in the Discussion.

A parameter quantifying the degree of genetic mixing among taxa can be derived as follows. As above, let P represent the overall proportion of non-native genes in a population. Furthermore, let P_i represent the proportion of non-native genes in the genome of the ith individual. Finally, let N represent the total number of individuals in the population. With this notation, $P=\frac{1}{N}\sum_i P_i$. The variance of P_i, $\mathrm{Var}(P_i)$, serves as a useful measure of how well mixed native and non-native genes are in the population. If the two taxa have interbred for a long time, all individuals in the population will have similar values of P_i, and $\mathrm{Var}(P_i)$ will be low. $\mathrm{Var}(P_i)$ will take a minimum value of zero when all individuals in the population have exactly the same amount of non-native ancestry. On the other hand, if there has not been extensive interbreeding between the taxa present in the population, $\mathrm{Var}(P_i)$ will be high. $\mathrm{Var}(P_i)$ will take a maximum value when all individuals in the population are genetically pure members of either taxa (P_i for every individual is either 0 or 1). The variance of P_i for this case is $P(1-P)$, the variance of a Bernoulli random variable. Given this, we propose a measure, which we will call m_d, of how well mixed non-native genes are in a population

$$m_d = 1-\frac{\mathrm{Var}(P_i)}{P(1-P)} \tag{1}$$

This parameter is equivalent to THI as defined by Vernesi *et al.* (2003) (see below), but we call it the 'degree of genetic mixing' in a population because we do not believe it is appropriate to call any parameter describing the amount of hybridization in a population the 'true hybrid index'. There are lots of reasonable ways of quantifying hybridization in a population, and, therefore, no single 'true' index. This parameter has a minimum value of 0, which occurs when a population consists of individuals from two species that have not yet interbred, that is, $\mathrm{Var}(P_i)=P(1-P)$. This parameter has a maximum value of 1.0, which occurs when all the individuals in the population have the same amount of non-native ancestry. Note that m_d is undefined if P equals 0 or 1. This is appropriate, as it does not make sense to quantify how genetically well mixed a population is when there are genes from only one taxon in the population.

Our parameter can also be defined for hybrid populations having more than two taxa. Let P_j represent the proportion of the genes in the population that belong to the jth taxa and let P_{ij} represent the proportion of genes in the ith individual that are from the jth taxa. With this notation, m_d is equal to

$$m_d = 1-\frac{\sum_j \mathrm{Var}(P_{ji})}{\sum_j P_j(1-P_j)} \tag{2}$$

Genetic mixing and gametic disequilibrium

It is well known that recently hybridized populations have high amounts of gametic disequilibrium (even at unlinked loci), and that this disequilibrium declines with time. This suggests that there may be a relationship between m_d and the amount of gametic disequilibrium in a population. This, indeed, is the case. Let D represent the parametric amount of gametic disequilibrium present at a pair of loci in a hybrid population, and let $\overline{D}$ represent the average value of D across all the pairs of loci in the population. Barton and Gale (1993); Equation 2b have shown that for populations in Hardy–Weinberg equilibrium

$$\mathrm{Var}(P_i)=\tfrac{1}{2}\overline{D} \tag{3}$$

If we divide both sides of Equation 3 by $P(1-P)$, we obtain

$$\frac{\mathrm{Var}(P_i)}{P(1-P)}=\tfrac{1}{2}\overline{D}' \tag{4}$$

where D' is a popular standardized measure of gametic disequilibrium that has a maximum value of 1.0 (Lewontin, 1964; Hedrick, 2011). Combining Equations 1 and 4 shows the relationship between m_d and D'

$$m_d = 1-\tfrac{1}{2}\overline{D}' \tag{5}$$

This is a very useful result for two reasons. First, it relates the degree of mixing in a population to the amount of gametic disequilibrium present in the population, a quantity frequently estimated in genetic samples. Second, it allows us to make quantitative statements about how quickly genes in a population will mix when there is random mating.

Genetic mixing in randomly mating populations

Interpreting empirical estimates of m_d would be easier if we knew how quickly it could increase in simple evolutionary scenarios, for example, if there was random mating in a population. The relationship between m_d and D' makes it easy to make some simple statements about the behavior of m_d in cases like this. For example, it is well known that D and D' for unlinked loci decrease by a factor of 0.5 in a randomly mating population (for example, Hedrick, 2011). This fact tells that if we use unlinked loci to estimate the ancestry of individuals, $1-m_d$ will decrease by a factor of 0.5 every generation of random mating. Therefore, if a population begins with genetically pure individuals from two taxa ($m_d=0$), and the individuals in the population mate randomly for t generations, m_d at generation t, $m_d^{(t)}$, will equal

$$m_d^{(t)} = 1-\left(\tfrac{1}{2}\right)^t \tag{6}$$

For the first three generations of random mating, m_d will equal 1/2, 3/4 and 7/8 (Figure 1). After five generations of random mating, m_d will be ~0.97, and mixing will be nearly complete. This relationship assumes that the population begins with genetically pure individuals from two taxa, but is independent of the proportion of non-native individuals that enter the population.

Estimation

We will discuss estimation of m_d in two contexts: cases in which diagnostic loci are available and cases in which diagnostic loci are not available.

Estimation using diagnostic loci

Estimating m_d using diagnostic loci is facilitated by noting the term $\mathrm{Var}(P_i)/P(1-P)$ in Equation 1 is mathematically equivalent to Wright's F_{ST} (Wright, 1951), or more specifically, F_{ST} for one locus with two alleles. The only difference between m_d and F_{ST} is that P_i in Equation 1 refers to the frequency of alleles in an individual whereas P_i in Wright's (1951) definition of F_{ST} refers to the frequency of alleles in a population. The similarity is not coincidental. F_{ST} quantifies how allele frequencies vary among populations; our parameter, m_d, quantifies how allele frequencies vary among individuals. The mathematical equivalence between m_d and F_{ST} allows us to use the well-developed literature on F_{ST} to estimate m_d (see below).

If diagnostic loci are available to unambiguously discriminate between native and non-native alleles, standard methods for estimating F_{ST} can be used to estimate m_d. For example, Weir and Cockerham's $\hat{\theta}$ (1984) can be used to produce an estimate of m_d, $\hat{m}_d$

$$\hat{m}_d = 1 - \hat{\theta} \tag{7}$$

Weir and co-workers have presented a few alternative methods for estimating $\hat{\theta}$ (Weir and Cockerham, 1984; Weir and Hill, 2002; Weir, 2010). The method most appropriate for the application here is the estimator designed to compare gene pools in randomly mating populations (Weir and Cockerham, 1984, unlabeled equation at the top of page 1363; Weir and Hill, 2002, Equation 5). Using our notation, this estimator is calculated as

$$\hat{\theta} = \frac{\frac{1}{N_s-1}\sum_i n_i(\hat{P}_i - P)^2 - \frac{1}{\sum_i (n_i-1)}\sum_i n_i\hat{P}_i(1-\hat{P}_i)}{\frac{1}{N_s-1}\sum_i n_i(\hat{P}_i - P)^2 + \frac{(\overline{n}-1)}{\sum_i (n_i-1)}\sum_i n_i\hat{P}_i(1-\hat{P}_i)} \tag{8}$$

where N_s is the number of individuals sampled from a hybrid population, n_i is the number of amplified alleles in individual i, $\hat{P}_i$ is the estimated proportion of non-native alleles in the genome of individual i, and $\overline{n}$ is the average number of alleles amplified in each genotyped individual

$$\overline{n} = \frac{1}{N_s - 1}\left(\sum_{i=1}^{n} n_i - \frac{\sum_i n_i^2}{\sum_i n_i}\right) \tag{9}$$

These equations, as noted above, are only applicable for loci having diagnostic alleles.

$\hat{\theta}$ can also be calculated using software designed for estimating F_{ST} (for example; GENEPOP; Rousset, 2008). Doing this requires reorganizing the data so that each individual in the sample is represented as a population and the diagnostic alleles from all loci are pooled into a single locus. Appendix shows how this can be done to create a GENEPOP file (Rousset, 2008).

Weir and Cockerham's (1984) $\hat{\theta}$ is essentially unbiased when used to estimate F_{ST}. However, $\hat{\theta}$ may not be as unbiased when used to estimate m_d. Weir and Cockerham's formula for $\hat{\theta}$ (Equation 8 in our paper) assumes alleles sampled from an individual are random and independent draws from a gene pool (which, in our application, is an individual's genome). With this assumption, the estimated proportion of non-native ancestry in the ith individual, $\hat{P}_i$, will be binomially distributed. This will not be true when there is gametic disequilibrium in a population (for example, in the early generations of hybridization). When there is gametic disequilibrium in a population, the alleles present in diploid genotypes are not always going to be independent. This is easily seen by considering an F_1 individual (that is, an individual having one parent from each hybridizing taxon). In an F_1

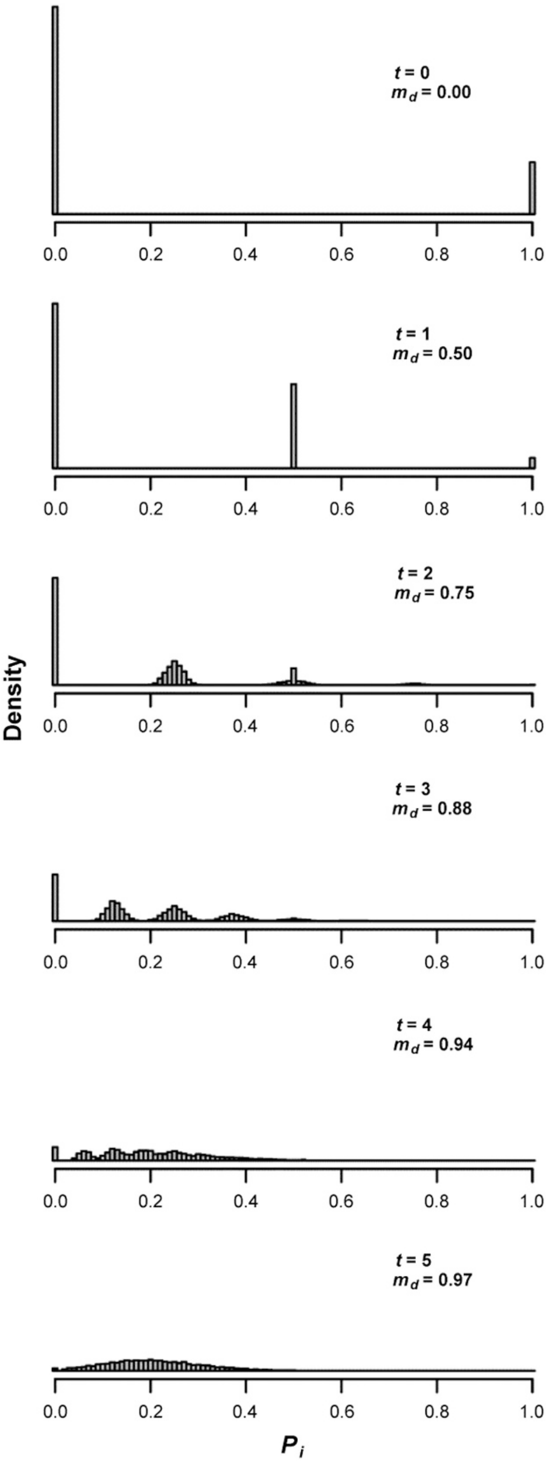

Figure 1 The distribution of hybridity coefficients (P_i) among individuals in a randomly mating populations over five generations (t=0–5). Individuals with P_i equal to 0 or 1 are genetically pure members of alternative taxa.

Table 1 Statistical bias of $\hat{m}_d$ calculated from Equation 6 and 9 estimated from computer simulations for samples of varying sizes (N_{Loci}, $N_{\text{Individuals}}$) from a population of 200 individuals after *t* generations of random mating (see text for details)

N_{Loci}	$t=2$: $m_d=0.75$			$t=5$: $m_d=0.97$		
	$N_{Individuals}$			$N_{Individuals}$		
	10	*20*	*50*	*10*	*20*	*50*
8	0.0330	0.0326	0.0325	0.0065	0.0062	0.0059
16	0.0154	0.0150	0.0150	0.0036	0.0034	0.0032
48	0.0043	0.0039	0.0039	0.0015	0.0015	0.0014
96	0.0016	0.0012	0.0012	0.0010	0.0010	0.0010

Note that the bias was positive in all the cases; this indicates that the estimated values of m_d tended to be higher than the parametric value.

individual, every locus will be heterozygous and P_i will equal 0.5. Because each locus is heterozygous, $\hat{P}_i$ will also equal 0.5—no matter how few or many loci are genotyped. In other words, when an F_1 individual is genotyped, there will be no sampling error in $\hat{P}_i$. Equation 8 assumes binomial sampling error, so terms in Equation 8 intended to eliminate sampling bias will not work as intended. This could result in a negative estimate of $\mathrm{Var}(P_i)/P(1-P)$, which would produce an estimate of m_d that is >1.0. This is equivalent to an estimate of F_{ST} being <zero.

We used computer simulations to estimate how much bias there was in the estimates of m_d calculated from Equation 7 and 8 for realistic amounts of data. We simulated populations of 50, 200 and 2000 randomly mating individuals that were founded with 20% non-native individuals. We modeled these populations after cutthroat trout (*Oncorhynchus clarkii*)—a species which frequently hybridizes with non-native rainbow trout (*O. mykiss*)—and assumed the ratio of the genetic effective population size, N_e, to census size, N, was 0.23 (Finger *et al.*, 2011). We achieved this ratio of N_e to N by varying the reproductive success according to the method of Anderson (2001). For each individual, 52 chromosomal arms were simulated with 10 equally spaced, species–specific diagnostic di-allelic loci. Recombination was allowed to occur once per chromosomal arm per generation (Danzmann *et al.*, 2005). In each generation we calculated the true degree of mixing, m_d, for each population. Then we drew 1000 simulated samples of 10, 20, 50 or 100 individuals with genotypes at 8, 16, 48 and 96 loci. The bias in estimates of m_d was calculated by comparing the average estimate of m_d with the parametric value for the populations, that is, bias = average($\hat{m}_d$) − m_d.

The simulations were performed using a program written by us in R (R Development Core Team, 2008).

The computer simulations showed that there was only a modest amount of bias in the first few generations of mating, and this bias was becoming negligible by the fifth generation (Table 1). In all cases, estimates of m_d tended to be slightly higher than the parametric value (that is, the bias was positive). As expected, the amount of bias was greatest in small samples, and for populations in the early stages of hybridization. The smallest samples we examined had 10 individuals genotyped at eight diagnostic loci. Even with these small samples, the bias observed in the second generation (0.0330), was only 4.4% of the parametric value (0.75). When the number of diagnostic loci was increased to 16, the bias was reduced to just over 2%. The bias present in the estimates of m_d for populations that had five generations of random mating ($m_d=0.97$) was much lower. Even for the smallest samples, it was only 0.6% of the parametric value.

Estimation using non-diagnostic loci

When diagnostic loci are not available, m_d can be estimated by estimating the ancestry of each individual in a sample, and then inserting these estimates into Equations 1 or 2, (depending on the number of taxa involved). There are a variety of methods for estimating the ancestry (P_i) of hybrid individuals using loci that do not have diagnostic alleles (for example, Pritchard *et al.*, 2000; Anderson and Thompson, 2002). The computer programs STRUCTURE and NEWHYBRIDS have been popular for this type of analysis. The accuracy of estimates of m_d obtained from $P_i's$ calculated by these programs will depend on how well the $P_i's$ are estimated. m_d is calculated from the variance of P_i, so if there is a bias in this variance, there will be a bias in estimates of m_d. Little is known about the error structure of STRUCTURE or similar analytic approaches, but it is likely that these errors will be affected by the amount of genetic differentiation among the taxa being studied, the number of loci genotyped and the amount of genetic variation at those loci.

We performed a series of computer simulations to explore how estimation error in $P_i's$ affected estimates of m_d. A systematic investigation of the error structure present in STRUCTURE or NEWHYBRIDS is beyond the scope of this investigation, so we examined how three plausible, generic models of estimation error affected estimates of m_d. In all simulations, we assumed there were two hybridizing taxa. As above, we simulated populations of 50, 200 or 2000 individuals and kept track of the parametric values of P_i for each individual in the population. Estimates of P_i were obtained in the simulation by assuming one of the three statistical models of estimation error. The first model assumed that the ancestry of one of the taxa was always underestimated by a constant amount. The second model assumed that estimates of the ancestry of two taxa were biased towards 0.5 by a constant proportion. The third model assumed that estimation error was normally distributed. We tested different magnitudes of estimation error for each model and ran 10 000 simulations for each case to quantify the bias in estimates of m_d.

The simulations showed that the bias in estimates of m_d was a function of how well the ancestry of each individual, P_i, was estimated (results not shown). Good estimates of P_i produced good estimates of m_d. There did not appear to be any thresholds affecting how estimation error in P_i was propagated to estimates of m_d.

DISCUSSION

We have defined and derived a parameter, m_d, that quantifies the amount of genetic mixing of native and non-native genes in a hybrid population. The parameter that we developed is related to both F_{ST} and the average amount of pairwise gametic disequilibrium in a population. The value of this parameter will rapidly approach 1.0 in randomly mating populations; after five generations of random mating, m_d will be ~0.97. Computer simulations showed that when diagnostic loci are available, one of Weir and Cockerham's (1984) estimators of F_{ST} did a very nice job of estimating m_d. If diagnostic loci are not available, m_d can be estimated by inserting the estimated ancestry of each individual into Equations 1 or 2,.

The genetic mixing parameter described here may be thought of as an alternative version of the 'true hybridization index' (THI) parameter of Vernesi *et al.* (2003). Both parameters have a range of 0 to 1.0 and both parameters quantify as to how well genes are mixed in a population. In fact, m_d and THI are mathematically equivalent. The main difference between m_d and THI is how the parameters are defined. m_d is defined as a variance of the amount of non-native genes *across* individuals in a population. As noted above, this variance will be zero when the population is genetically well mixed and all individuals have the same amount of non-native ancestry. THI is defined as the average variance of genetic ancestry within individuals. This is explained as follows. The first step in calculating THI is calculating the variance of ancestry, V_i, within each individual in the sample. If d taxa are hybridizing, and we use our notation, the variance of non-native genes in the ith individual is $V_i = \frac{\sum_j \left(P_{ij} - \frac{1}{d}\right)^2}{d}$. Notice that this variance measures how close the native and non-native ancestries of an individual are to 1/d. This quantity, somewhat surprisingly, is related to the amount of mixing in a population. If all the individuals in a population are genetically pure, the proportion of native and non-natives genes in the ith individual, P_{ij}, will be 0 or 1,and the values of V_i for every individual will be relatively high. On the other hand, if all individuals in a population have the same amount of non-native genes, the values of P_{ij} will be closer to 1/d and the values of V_i will be smaller. This is the principle THI uses to quantify the amount of mixing in a population. More specifically, THI is calculated by taking the average of V_i across individuals and then standardized to account for the minimum and maximum values of this average (given the amount of non-native genes in the population). This calculation produces a quantity with the same value as m_d. m_d has the advantage of a simple, easy-to-interpret definition. This is important because it allowed us to relate m_d to the amount of gametic disequilibrium in a population, identify how m_d changed in randomly mating populations, and estimate m_d using formulae for F_{ST}.

In randomly mating populations, m_d will rapidly approach 1.0 at a rate specified by Equation 6. There are, however, many plausible reasons as to why natural populations will behave differently. For example, if mating between the two taxa is not random, genetic mixing may proceed more slowly. Alternatively, if hybrid individuals have a lower evolutionary fitness, mixing will be slower. Mixing will also appear to be slower if non-native individuals continuously enter a population. And, finally, the rate of mixing will be affected by the genotypes of the first individuals to enter a population. If these individuals have mixed ancestry, m_d will be higher than if the first immigrants were genetically pure non-natives.

A few comments regarding the definition of a 'hybrid swarm' may be useful, as several definitions are present in the literature. Rhymer and Simberloff (1996) defined a hybrid swarm as a population containing individuals with various degrees of non-native ancestry. Allendorf *et al.* (2001) used a slightly different definition; they specified that all individuals in a hybrid swarm must be hybrids. Finally, Allendorf and Leary (1988) provided the most strict definition; they defined a hybrid swarm as a population in which all individuals in the population have the same amount of non-native ancestry (that is, $m_d = 1$). All of these definitions may be useful, but in some circumstances it may be more useful to quantify how well mixed non-native genes are in a population rather than to classify a population as being a hybrid swarm or not. This is especially likely to be true as the number of loci used to study hybridization increases. Genomic data are expected to have very high power to detect slight amounts of variation in ancestry. Therefore, when genomic data are available, it may not be useful to define a hybrid swarm as a population in which all individuals have the same amount of non-native ancestry.

We conclude this paper with a discussion of what it means for genes in a hybrid population to be 'mixed.' We have assumed throughout this paper that native and non-native genes are well mixed when all the individuals in the population have the same amount of non-native ancestry. This criterion should be useful for studying populations that are in the early stages of hybridization or have stable distributions of hybrid genotypes. Other mixing criteria and parameters may be more informative for examining populations that have been hybridizing for a long time. In particular, if populations have been hybridizing for a long time, it may be more useful to quantify how thoroughly recombination has broken up and shuffled the genomes of native and non-native taxa into small chromosomal segments (for example, vonHoldt *et al.,* 2011 and references within). When genomic data are available, such an approach can be used to date the timing of admixture (Tang *et al.,* 2006). If such data are not available, and a population is in the early stages of admixture, the mixing degree parameter presented in this paper should be informative.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank reviewers for thoughtful comments that improved this manuscript. Funding for this project was provided by the National Science Foundation (DEB 0717456 to STK and DGE 1049562 to JHP) and the United States Geological Service.

Allendorf FW, Leary RB, Spruell P, Wenburg JK (2001). The problems with hybrids: setting Conservation guidelines. *Trends Ecol Evol* **16**: 613–622.

Allendorf FW, Leary RF (1988). Conservation and distribution of genetic variation in a poltypic species, the cutthroat trout. *Conserv Biol* **2**: 170–184.

Anderson EC (2001). *Monte Carlo Methods for Inference in Population Genetic Models*. University of Washington: Seattle, WA, USA, 195 p.

Anderson EC, Thompson EA (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**: 1217–1229.

Barton NH, Gale KS (1993). Genetic analysis of hybrid zones. In: Harrison RG (ed). *Hybrid Zones and the Evolutionary Process*. Oxford University Press, New York, NY, USA, pp 13–45.

Barton NH, Hewitt M (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematic* **16**: 113–148.

Cheng C-Y, Reich D, Coresh J, Boerwinkle E, Patterson N, Li M *et al.* (2010). Admixture mapping of obesity-related traits in African Americans: the Atherosclerosis Risk in Communities (ARIC) Study. *Obesity* **18**: 563–572.

Danzmann RG, Cairnet M, Davidson WS, Ferguson MM, Gharbi K, Guyomard R *et al.* (2005). A comparative analysis of the rainbow trout genome with 2 other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily: Salmoninae). *Genome* **48**: 1037–1051.

Everitt BS, Skrondal A (2010). *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge, UK.

Finger AJ, Anderson EC, Stephens MR, May BP (2011). Application of a method or estimating effective population size and admixture using diagnostic single nucleotide polymorphisms (SNPs): implications for conservation of threatened Paiute cutthroat trout (Oncorhynchus clarkii seleniris) in Silver King Creek. *Can J Fish Aquat Sci* **68**: 1369–1386.

Harrison RG, Bogdanowicz SM (1997). Patterns of variation and linkage disequilibrium in a field cricket hybrid zone. *Evolution* 493–505.

Hedrick PW (2011). *Genetics of Populations*. Jones and Bartlett Publishers: Sudbury, MA, USA.

Hedrick PW (2009). Conservation genetics and North American bison (Bison bison). *J Hered* **100**: 411–420.

Hey J (2010). Isolation with migration models for more than two populations. *Mol Biol Evol* **27**: 905–920.

Hogg JT, Forbes SH, Steele BM, Luikart G (2006). Genetic rescue of an insular population of large mammals. *Proc Biol Sci* **273**: 1491–1499.

Johnson WE, Onorato DP, Roelke ME, Land ED, Cunningham M, Belden RC *et al.* (2010). The genetic restoration of the Florida panther. *Science* **329**: 1641–1645.

Lewontin RC (1964). The interaction of selection and linkage. I. General considerations, heterotic models. *Genetics* **49**: 49–67.

Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD *et al.* (2007). A genomewide admixture mapping panel for hispanic/latino populations. *Am J Hum Genet* **80**: 1171–1178.

Muhlfeld CC, Kalinowski ST, McMahon TE, Painter S, Leary RF, Taper ML *et al.* (2009). Hybridization reduces fitness of cutthroat trout in the wild. *Biol Lett* **5**: 328–331.
Pertoldi C, Tokarska M, Wójcik J.M, Kawako A, Randi E, Kristensen TN *et al.* (2010). Phylogenetic relationships among the European and American bison and seven cattle breeds reconstructed using the BovineSNP50 Illumina Genotyping BeadChip. *Acta Theriol* **55**: 97–108.
Piñeyro-Nelson A, Van Herwaarden J, Perales HR, Serratos-Hernandez JA, Rangel A, Hufford MB *et al.* (2009). Transgenes in Mexican maize: molecular evidence and methodological considerations for GMO detection in landrace populations. *Mol Ecol* **18**: 750–761.
Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
Rhymer JM, Williams MJ, Braun MJ (1994). Mitochondrial analysis of gene flow between New Zealand mallards (Anas platyrhynchos) and grey ducks (A. supreciliosa). *Auk* **111**: 970–978.
Rhymer JM, Simberloff D (1996). Extinction by hybridization and introgression. *Annu Rev Ecol Syst* **27**: 83–109.
Rousset F (2008). Genepop'007: a complete reimplementation of the genepop software for Windows and Linux. *Mol Ecol Resour* **8**: 103–106.
R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria. ISBN. 3-900051-07-0, URL http://www.R-project.org.
Steeves TE, Maloney RF, Hale ML, Tylianakis JM, Gemmel NJ (2010). Genetic analyses reveal hybridization but no hybrid swarm in one of the world's rarest birds. *Mol Ecol* **19**: 5090–5100.
Tang H, Coram M, Wang P, Zhu X, Risch N (2006). Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet* **79**: 1–12.
Vander Wal E, Garant D, Festa-Bianchet M, Pelletier F (2012). Evolutionary rescue in vertebrates: evidence, applications and uncertainty. *Phil Trans R Soc B* **368**: 20120090.
Vernesi C, Crestanello B, Pecchioli E, Tartari D., Caramelli D, Hauffe H *et al.* (2003). The genetic impact of demographic decline and reintroduction in the wild boar (Sus scrofa): a microsatellite analysis. *Mol Ecol* **12**: 585–595.
vonHoldt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, Boyko AR *et al.* (2011). A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Resour* **21**: 1294–1305.
Watrud LS, Lee EH, Fairbrother A, Burdick C, Reichman JR, Bollman M *et al.* (2004). Evidence for landscape-level, pollen-mediated gene flow from genetically modified creeping bentgrass with CP4 EPSPS as a marker. *Proc Natl Acad Sci USA* **40**: 14533–14538.
Weir BS (2010). *Genetic Data Analysis III*. Sinauer Associates, Inc: Sunderland, MA, USA.
Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.
Weir BS, Hill WG (2002). Estimating F-statistics. *Annu Rev Genet* **36**: 721–750.
Winkler CA, Nelson GW, Smith MW (2010). Admixture mapping comes of age. *Annu Rev Genomics Hum Genet* **11**: 65–89.
Wright S (1951). The genetical structure of populations. *Ann Eugen* **15**: 323–353.
Zapiola ML, Mallory-Smith CA (2012). Crossing the divide: gene flow produces intergeneric hybrid in feral transgenic creeping bentgrass population. *Mol Ecol* **21**: 4672–4680.

APPENDIX

GENEPOP files with genotypes of standard format and reformatted to estimate $Var(P_i)/P(1-P)$. As an example, let us consider a sample of four fish, in which: two fish are native fish with no non-native genes, one fish is an F1 hybrid, and the last fish is a pure non-native fish.

Here is a sample data file for these four individuals with diploid genotypes at three loci.

```
Sample GENEPOP data in normal format
Diagnostic Locus1, Diagnostic Locus2, Diagnostic
Locus3
POP
  Ind1,              11          11          11
  Ind2,              11          11          11
  Ind3,              12          12          12
  Ind4,              22          22          22
```

The data file above can be reformatted to estimate $Var(P_i)/P(1-P)$ using GENEPOP (see below). Note that each individual in the original sample is considered a 'population' in the reformatted GENEPOP file. Note also the data is reformatted to be haploid.

```
Sample GENEPOP data reformatted to calculate Var
(Pi)/[P(1P)]
Diagnostic Loci
  POP
  Ind1Locus1a,                      01
  Ind1Locus1b,                      01
  Ind1Locus2a,                      01
  Ind1Locus2b,                      01
  Ind1Locus3a,                      01
  Ind1Locus3b,                      01
  POP
  Ind2Locus1a,                      01
  Ind2Locus1b,                      01
  Ind2Locus2a,                      01
  Ind2Locus2b,                      01
  Ind2Locus3a,                      01
  Ind2Locus3b,                      01
  POP
  Ind3Locus1a,                      01
  Ind3Locus1b,                      02
  Ind3Locus2a,                      01
  Ind3Locus2b,                      02
  Ind3Locus3a,                      01
  Ind3Locus3b,                      02
  POP
  Ind4Locus1a,                      02
  Ind4Locus1b,                      02
  Ind4Locus2a,                      02
  Ind4Locus2b,                      02
  Ind4Locus3a,                      02
  Ind4Locus3b,                      02
```

In this example, the parametric value of m_d is equal to 0.2667. An estimate calculated from GENEPOP's estimate of $Var(P_i)/P(1-P)$ is equal to 0.2571.