Molecular Ecology (2003) **12**, 903–915

A test for deviation from island-model population structure

ADAM H. PORTER
Department of Entomology & Program in Organismic and Evolutionary Biology, University of Massachusetts, Amherst, MA 01003–2410, USA

Abstract

The neutral island model forms the basis for several estimation models that relate patterns of genetic structure to microevolutionary processes. Estimates of gene flow are often based on this model and may be biased when the model's assumptions are violated. An appropriate test for violations is to compare F_{ST} scores for individual loci to a null distribution based on the average F_{ST} taken over multiple loci. A parametric bootstrap method is described here based on Wright's β-distribution to generate null distributions of F_{ST} for each locus. These null distributions account for error introduced by sampling populations, individuals and loci, and also biological sources of error, including variable alleles/locus and inbreeding. Confidence limits can be obtained directly from these distributions. Significant deviations from the island model may be the result of selection, deviations from the island model's migration pattern, nonequilibrium conditions, or other deviations from island-model assumptions. Only strong biases are likely to be detected because of the inherently large sampling variation of F_{ST}. Nevertheless, a coefficient, *Nb*, describing bias in the spread of the β-distribution in units comparable to the gene flow parameter, *Nm*, can be obtained for each locus. In samples from populations of the butterfly *Coenonympha tullia*, the loci Idh-1, Mdh-1, Pgi and Pgm showed significantly lower F_{ST} than expected.

Keywords: *Coenonympha*, F_{ST}, gene flow, Lewontin–Krakauer test, neutrality, parametric bootstrap, population structure, selection

Received 23 October 2001; revision received 2 December 2002; accepted 11 December 2002

Introduction

Population geneticists have long searched for simple statistical tests to determine which marker loci are appropriate for use in the analysis of genetic population structure. Inferences about gene flow are often based on patterns of allele frequencies at polymorphic marker loci using Wright's (1931) infinite island model of population structure, and therefore depend on the assumptions of this model (Lewontin 1974; Slatkin 1987; Whitlock & McCauley 1999). Among the most important of these are that migrants may move to any other population with equal probability, that only genetic drift and gene exchange, but not selection, affect patterns of allele frequency differences, and that the frequency distribution of allele frequencies among populations has settled into a stable shape (but see Rannala 1996). Data that violate these assumptions can produce biased estimates. Loci experiencing balancing selection will have allele frequencies more similar than expected under neutrality, giving the false impression of higher gene flow, and the opposite effect occurs for loci under differentiating selection. Migration patterns different from that of the island model can also bias gene flow estimates. Depending on the migration pattern, greater or lesser variation in allele frequencies than the island model predicts, or simply differently shaped allele frequency distributions, may be produced. The spectrum of possible explanations for observed genetic patterns makes the interpretation of genetic population structure difficult and there is legitimate scepticism towards numerical estimates of gene flow from genetic patterns (Whitlock & McCauley 1999). Nevertheless, the central problem remains of teasing apart these competing causes of genetic patterns and estimating their relative contributions. Statistical tests are needed that allow biologists to identify these effects.

Numerous tests are available to detect the presence of selection within individual populations. None can be readily extrapolated to studies of spatial population structure. Early tests (e.g. Ewens 1972) compared the distribution of allele frequencies at a locus to that expected from neutral

Correspondence: A. H. Porter. Fax: (413) 545–0231; E-mail: aporter@ent.umass.edu

variation alone. Large data sets are needed, but accumulating these test statistics over populations to increase statistical power would tacitly invoke the assumption that the populations are independent replicates, and gene exchange violates this assumption. More recent tests of selection assess patterns of nucleotide replacements in DNA sequence data for a single locus (e.g. Tajima 1989; Fu & Li 1993; Simonsen *et al*. 1995; Fu 1997) or linked loci (Kelly 1997) in a single population sample. They detect the traces of rapid change in the frequency of alleles in the recent past, but it is much more difficult to assess the relative fitnesses of genotypes currently segregating in the population.

The Lewontin–Krakauer test (Lewontin & Krakauer 1973) and variants of it (Tsakas & Krimbas 1976; Bowcock *et al*. 1991; McDonald 1994; Beaumont & Nichols 1996; Vitalis *et al*. 2001) assess patterns of variation among populations and can be used to test for deviation from island-model assumptions. These tests are based on the null distribution of Wright's F_{ST} (Wright 1931, 1951, 1978) among loci, using the same underlying island model as used when estimating gene flow. F_{ST} is the variance of allele frequencies among subpopulations, relative to the total variance available in the pool of populations: $F_{ST} = \sigma_q^2 / Q(1-Q)$, where Q is the allele frequency in the pool of populations. (Weir & Cockerham (1984) provide an unbiased estimator of F_{ST}). The rationale for these tests is simple: all loci experience the effects of gene flow equally, but selection acts on loci independently (Cavalli-Sforza 1966). If selected loci are present in the sample, the variance in single-locus F_{ST} estimates should be higher than if only neutral loci were sampled. In principle, these tests require knowledge of the shape of the null sampling distribution of F_{ST}, which has not been derived analytically for any migration pattern. Available tests therefore rely either on summary statistics for this distribution (namely the variance) that can be roughly approximated analytically for the island model (Lewontin & Krakauer 1973), on coalescent models of drift following complete isolation of subpopulations (Vitalis *et al*. 2001), or on simulation to obtain the null distribution numerically (Bowcock *et al*. 1991; Beaumont & Nichols 1996; Balanovskaya & Nurbaev 1998a,b).

A direct approach is described here that generates a null distribution of F_{ST} for each locus separately. This parametric bootstrap method circumvents the purely statistical (but not the biological) criticisms of the Lewontin–Krakauer test (Jacquard 1974; Nei & Maruyama 1975; Robertson 1975a,b; Ewens & Feldman 1976; Ewens 1977; Nei & Chakravarti 1977; Nei *et al*. 1977). The new test uses the multiallele version of Wright's β-distribution of allele frequencies among populations (Wright 1931, 1978) (the same island-model distribution underlying the theory of gene flow estimation) as a basis for constructing the null distribution of F_{ST}. This multivariate β-distribution, also called a Dirichlet distribution (Rannala 1996; Rannala & Hartigan 1996; Burr 2000), is a function of the multilocus F_{ST} value and the among-population average allele frequencies of each locus. F_{ST} is recalculated from genotypes that are repeatedly resampled from this distribution, following the sampling structure of the original data, and the null distribution of F_{ST} is accumulated. The observed values of F_{ST} for each locus are compared to their null distributions to see if they show significant bias. Any bias in F_{ST} at individual loci can be measured. If it is independently justified, these loci can be removed from the analysis prior to estimation of genetic structure. To demonstrate the test, an analysis of simulated data and the reanalysis of a data set from a previously published study (Porter & Geiger 1988) are presented. A computer program and source code are available (www-unix.oit.umass.edu/~aporter/software/).

Sampling model

Under the null hypothesis that alleles are neutral, the primary factors responsible for patterns of differentiation among island-model populations are mutation, gene flow and genetic drift. These processes result in an equilibrium degree of differentiation under constant demographic conditions, and mutation usually has a negligible effect on allele frequencies relative to migration. In an ideal population subdivided according to the infinite island model of Wright (1931), ignoring the mutation terms, the steady-state distribution of allele frequencies for the single locus, two-allele case is approximately

$$\varphi(q|Nm,Q) = \frac{\Gamma(4Nm)}{\Gamma(4NmQ)\Gamma(4Nm[1-Q])} q^{4NmQ-1}(1-q)^{4Nm(1-Q)-1} \quad (1)$$

(Wright 1931), where q is the allele frequency in a single subpopulation, Q is the average allele frequency in the total population, N is the subpopulation size and m is the proportion of individuals migrating among populations each generation. For multiple alleles at a locus, this distribution extends to

$$\varphi(\mathbf{q}\,|\,Nm,\mathbf{Q}) = \Gamma(4Nm)\prod_{a=1}^{h} \frac{q_a^{4NmQ_a-1}}{\Gamma(4NmQ_a)}, \quad (2)$$

where $\mathbf{Q}$ is a vector of frequencies for each of the h alleles in the total population, and $\mathbf{q}$ is the corresponding vector of allele frequencies for a subpopulation (Wright 1978). $\varphi(\mathbf{q}\,|\,Nm,\mathbf{Q})$ represents a null distribution of allelic frequencies among island-model subpopulations, under the null hypothesis that only gene flow and drift, but not selection, influence allelic frequencies.

When the shape of $\varphi(\mathbf{q}\,|\,Nm,\mathbf{Q})$ has settled to equilibrium, the correlation of alleles within subpopulations (F_{ST}) is related by a simple function to the gene flow rate,

$$4Nm \approx F_{ST}^{-1} - 1 \quad (3)$$

(Wright 1931). Substituting into equation (2) yields

$$\varphi(\mathbf{q}|F_{ST},\mathbf{Q}) = \Gamma(F_{ST}^{-1} - 1)\prod_{a=1}^{h}\frac{q_a^{[F_{ST}^{-1}-1]Q_a-1}}{\Gamma([F_{ST}^{-1} - 1]Q_a)}. \quad (4)$$

(Rannala & Hartigan 1996). This describes the null allele frequency distribution in terms of the standardized variance of allele frequencies among subpopulations, F_{ST}, which is easily estimated from data. (For haploid loci, the quantity $4Nm$ in equations (1) to (3) is replaced by $2Nm$, and equation (4) remains the same.)

In finite samples, the realized value of F_{ST} also depends on the extent of inbreeding within subpopulations, F_{IS}. In the hierarchical relationship $(1 - F_{IT}) = (1 - F_{ST})(1 - F_{IS})$ of Wright (1951), F_{IT} is a measure of homozygote excess seen in the total population, and it must be partitioned between F_{ST} and F_{IS}. Sampling of genotype frequencies within subpopulations affects these values (Cockerham 1970; Weir & Cockerham 1984). The expected genotype frequencies can be obtained from known values of F_{IS} and $\mathbf{q}_i$, the vector of allele frequencies for subpopulation i. It is

$$\phi(\gamma_i|\mathbf{q}_i, F_{IS}) \equiv \begin{cases}\gamma_{i,aa} = q_{i,a}^2 + F_{IS}q_{i,a}(1 - q_{i,a}) \\ \gamma_{i,ab} = q_{i,a}q_{i,b}(1 - F_{IS}), a \neq b,\end{cases} \quad (5)$$

where γ_i is an $h \times h$ square matrix of (gene-ordered) genotypic combinations for subpopulation i, and a and b are indices of alleles. The diagonal elements of this matrix are homozygote frequencies and the off-diagonal elements are frequencies of (gene-ordered) heterozygotes. Equation (5) can give spurious negative genotype frequencies when there are more than two alleles and F_{IS} is negative; in that case, a more general form is given in the Appendix. This approach assumes that alleles of different loci are uncorrelated, an assumption that should be tested by estimating gametic disequilibrium when analysing data.

If $\mathbf{Q}$, F_{ST}, and F_{IS} are known, then distributions (4) and (5) can be used to obtain a random sample of genotypes under the null hypothesis that only gene flow, drift and local inbreeding determine genetic structure. From these genotypes, a new $\tilde{F}_{ST}$ may be calculated. The new $\tilde{F}_{ST}$ differs only by sampling error from the original F_{ST}, and explicitly includes sampling effects for the number of subpopulations, the number of individuals in each subpopulation, and the deviation from Hardy–Weinberg proportions in the subpopulations. By repeating this process, a null sampling distribution of $\tilde{F}_{ST}$ values may be accumulated under the neutrality hypothesis. It is important to follow the sampling patterns of the original data set, so that the null $\tilde{F}_{IS}$ distribution for hypothesis testing explicitly includes these sampling patterns.

Typically, however, the population parameters $\mathbf{Q}$, F_{ST}, and F_{IS} are unknown and are instead estimated from the sample. I use $\hat{\mathbf{Q}}$ to represent the estimate of average gene frequencies, and $\hat{F}_{ST}$ and $\hat{F}_{IS}$ for the weighted averages of F_{ST} and F_{IS} taken over sampled loci and alleles, using the Weir & Cockerham (1984) method (their θ and f, respectively). Averages are appropriate for the generation of null distributions because all loci should experience the same degree of inbreeding (Cavalli-Sforza 1966; Lewontin & Krakauer 1973). These estimates may replace their respective parameter values. However, this substitution introduces new sources of sampling error into the null distribution, namely the standard errors of $\hat{\mathbf{Q}}$, $\hat{F}_{ST}$ and $\tilde{F}_{IS}$. I account for this variation by first resampling from the data to get standard bootstrap estimates of $\hat{\mathbf{Q}}$, $\hat{F}_{ST}$ and $\hat{F}_{IS}$, then using these to obtain the parametric-bootstrap samples $\tilde{F}_{ST}$ (see Appendix).

Confidence limits

The 2.5 and 97.5 percentiles from the null distribution of the multilocus $\tilde{F}_{ST}$ can be inserted into equation (3) to obtain confidence limits on the $\hat{F}_{ST}$ estimate, and on the estimated effective gene flow rate, denoted $\langle Nm \rangle$. The null $\tilde{F}_{ST}$ distribution is skewed (see example below), so the confidence limits will be asymmetrical around the mean and somewhat broader than those advocated by Weir & Cockerham (1984). They obtain confidence limits by jackknifing over loci, assuming a normal distribution of F_{ST}, and so do not incorporate sampling variation within and among subpopulations. Because of the skew, the confidence limits from the resampled null distribution will also be better for testing hypotheses involving the lower end of the distribution near $F_{ST} = 0$, where the distribution falls sharply. Of course, these confidence limits rely on the island-model assumptions. The true confidence limits may differ to the extent that the island-model assumptions are violated.

Hypothesis testing

Observed values $\hat{F}_{ST,k}$ for each locus are compared to the expected null distributions, and judged to differ significantly if they fall below the 2.5 percentile ($< \tilde{F}_{ST(0.025),k}$) or above the 97.5 percentile ($> \tilde{F}_{ST(0.975),k}$). Two-tailed significance levels are then estimated from the percentile of the null distribution into which the observed values fall. Some biologists may wish to adjust their significance levels to account for multiple comparisons. A sequential Bonferroni correction would be appropriate (Rice 1989), based on the number of loci tested.

Estimating the strength of the bias

Fitness, and therefore selection strength, enters the parametric distribution of allele frequencies (equation 2) as an exponential

scaling coefficient (Wright 1931, 1937) and therefore is not easily obtainable from F_{ST}. Non-island-model migration patterns have effects that are even more obscure. Nevertheless, a bias b_k in the effective gene flow estimate for locus k can be characterized to a first order of approximation as

$$4N(m + b_k) \approx F_{ST,k}^{-1} - 1. \quad (6)$$

For a locus where significant bias is detected, equation (6) applies, whereas for a neutral locus in the island model, $b_k = 0$ and this simplifies to equation (3). Taking their difference yields an estimate of the effective bias, Nb_k, as

$$\langle Nb_k \rangle \approx \frac{1}{4}(\tilde{F}_{ST,k}^{-1} - \hat{F}_{ST}^{*-1}), \quad (7)$$

where $\hat{F}_{ST,k}$ is the observed estimate of F_{ST} for the kth locus, $\hat{F}_{ST}^{*}$ is the observed estimate of F_{ST} after biased loci have been removed from the analysis, and the bracket notation $\langle\rangle$ represents the estimate of a product. Since b does not assume or specify an explicit functional relationship to fitness (e.g. $b \neq s$, a selection coefficient), it should be interpreted as a heuristic device for comparing different data sets and different loci. Balanovskaya & Nurbaev (1998a,b) refer to Nb_k (their R_S) as the index of selection intensity under the assumption that $\hat{F}_{ST}^{*}$ represents neutrality and the island model holds.

Interpretation of significant biases

There are many reasons that loci may deviate from the null expectation, and they are probably impossible to disentangle without independent information on spatial structure or selection. Selection, deviations from island-model migration patterns, or lack of time for the distribution to settle into a steady state could be biological causes, and bad data (e.g. from improperly resolved genotypes or undetected gene duplications) can also introduce biases. Deviations from the island model's migratory pattern in some cases can increase the allele-frequency variance among populations. This can manifest itself empirically in finite population samples as a small subset of 'outliers' — neutral loci that show excess deviation. This effect of migratory pattern would probably have to be strong to increase the probability of detection appreciably when the number of sampled loci is small. However, if detected, this pattern could readily be misinterpreted as being the result of selection on the divergent loci, and of course, selection could also be influencing the pattern. Additional information should always be sought to support conclusions about the reasons for significance. At minimum, however, the detection of bias should lead to new, testable hypotheses about the causes of observed genetic patterns.

Removing biased loci

Perhaps most biologists would wish to stop upon finding any significant deviation from island-model assumptions. For some systems, however, it may be independently justified to believe that the true population structure is reasonably approximated by an island model, and to try to remove biased loci in an effort to gain a better estimate of neutral population differentiation. To ensure that loci exhibiting significant biases in the locations and shapes of their null distributions do not contribute to the estimation of gene flow, loci with the most extreme biases should be sequentially dropped from the analysis and the entire process repeated until no more significant biases can be identified. If isolation-by-distance or other spatially restricted patterns of gene flow are suspected as the reason for the bias, then instead of eliminating loci, the sampling pattern could be limited to a smaller spatial scale and the analysis repeated. By choosing a smaller spatial scale, differentiation at the extremes of the range will be less and the distribution of allele frequencies will be more like that of the island model (Beaumont & Nichols 1996; Hudson 1998). On the other hand, if source-sink, range-expansion, or other population structures generating excess similarity are suspected, it is perhaps better to estimate the migration matrix (Beerli & Felsenstein 2001).

Simulation

Here is a simulation of 25 diploid subpopulations followed at 10 unlinked, diallelic loci, to demonstrate the analysis under controlled conditions. One locus in the simulation violates the island-model assumptions by being under selection, and the remaining loci are neutral. The selected locus experienced a simple pattern of heterozygote superiority, with homozygote fitnesses of 1 and heterozygote fitness of $1 + s$. The simulation was seeded using genotype frequencies obtained randomly from Wright's β-distribution (equation 4), with $N = 50$ individuals and $m = 0.1$. The source code relies on a C++ library that has been described elsewhere (e.g. Johnson & Porter 2000; Porter & Johnson 2002) and is freely available (http://www.oit-unix.umass.edu/~aporter/software/).

The analysis followed these steps:

(i) Estimate the allele frequencies for all loci the pooled subpopulations, $\hat{\mathbf{Q}}$. Estimate $\tilde{F}_{ST,k}$ for each locus k and the mean $\hat{F}_{ST}$ and $\hat{F}_{IS}$ among loci.

(ii) Generate a null distribution of $\tilde{F}_{ST,k}$ for each locus following the protocol in the Appendix, based on 1000 resampled replicates.

(iii) Compare the observed $\hat{F}_{ST,k}$ to the null distribution of $\tilde{F}_{ST,k}$ for each locus. Use the percentile method to determine the probability of obtaining an equal or more extreme value of $\hat{F}_{ST,k}$.

Table 1 Single-locus estimates of genetic structure among 25 simulated island-model populations from null distributions seeded using the weighted average $\hat{F}_{ST}$ over alleles and loci as described in the Appendix

Locus k	$\hat{F}_{ST}$	$\tilde{F}_{ST(0.025)}$	$\tilde{F}_{ST(0.975)}$	$\langle Nb_k \rangle$	$\langle Nb_{k(0.025)} \rangle$	$\langle Nb_{k(0.975)} \rangle$	P
1	0.024	0.030	0.098	5.82	–1.99	3.68	0.008*
2	0.046	0.029	0.101	0.88	–2.05	4.01	0.460
3	0.052	0.045	0.128	0.24	–2.58	1.03	0.162
4	0.063	0.043	0.128	–0.54	–2.58	1.26	0.424
5	0.042	0.035	0.114	1.42	–2.38	2.70	0.172
6	0.039	0.033	0.104	1.86	–2.13	2.97	0.136
7	0.079	0.041	0.139	–1.36	–2.73	1.60	0.860
8	0.066	0.030	0.100	–0.76	–2.04	3.87	0.710
9	0.080	0.036	0.107	–1.42	–2.20	2.32	0.480
10	0.067	0.031	0.104	–0.81	–2.13	3.50	0.704
F^*_{ST}	0.055	0.049	0.077	0	–1.30	0.59	0.332

Locus 1 is under stabilizing selection for global heterozygote superiority; the remaining loci are neutral and all are unlinked. $\hat{F}_{ST}$ is the value of F_{ST} estimated from data; $\tilde{F}_{ST(0.025)}$ and $\tilde{F}_{ST(0.975)}$ are the 95% confidence limits of the null distributions, and $\tilde{F}^*_{ST}$ is the weighted average over alleles and loci of the resampled data. $\langle Nb_k \rangle$ estimates the strength of bias, in the units of gene flow (Nm), shown by locus k and P is its significance level. $\langle Nb_{k(0.025)} \rangle$ and $\langle Nb_{k(0.0975)} \rangle$ estimate the 95% confidence limits around $\langle Nb_k \rangle$, representing the strength of selection or other bias that would be statistically detectable in the given sample.

It is also possible to seed each $\tilde{F}_{ST,k}$ distribution with an average $\hat{F}_{ST}$ calculated by omitting the locus being tested, a slightly more conservative approach. I did not test this but it should yield similar results.

Simulation results

The method was able to identify the locus that violated the island-model assumptions. After seeding allele frequencies at their equilibrium distribution using equation (4) with $F_{ST} = 0.05$ and allowing an additional 100 generations to settle, the calculated F-statistics using the Weir & Cockerham (1984) method were $\hat{F}_{ST} = 0.055$ and $\hat{F}_{IS} = -0.006$. Locus 1, under selection for heterozygote superiority, showed significantly lower F_{ST} (Table 1), as expected. The remaining nine neutral loci showed no significant differences from the average estimate (Table 1). All loci had comparable confidence limits around their parametric expectation, consistent with the fact that all had similar mean allele frequencies **Q**.

Example — *Coenonympha tullia*

Here is an example with multiple alleles and unbalanced sampling patterns, using allozyme data from a published study of butterfly populations. The multiple alleles and the variation among loci in their mean frequencies make the analysis less straightforward than the simulation above.

Porter & Geiger (1988) showed that several subspecies of the ringlet butterfly, *Coenonympha tullia* (Satyrinae), characterized by minor differences in wing pattern, were highly polymorphic and broadly homogeneous in their allozyme allele frequencies across California, Nevada and Oregon (USA). These are grassland butterflies that occur in large, contiguous populations in western North America, and extinction/recolonization or source-sink migration patterns are not suspected. Pairwise gametic disequilibria were not significant in these populations. Porter & Geiger assumed the allozymes were nearly neutral and concluded from their low $\hat{F}_{ST}$ estimate [obtained using the formula from Wright (1978) that does not correct for sample size] that gene flow was high within and among subspecies. However, the high similarity among allozymes could also be the result of balancing selection on some of the loci. On the other hand, isolation by distance or disruptive selection on some loci could be masking even higher gene flow. To test for heterogeneity among loci in their population structures, the analyses above were run on the pooled set of subspecies. The single sample from the disjunct subspecies *C. tullia mono*, showing allele frequency divergence at several loci, was dropped from the analyses. The data therefore consisted of 432 individual butterflies distributed among 20 subpopulations, scored for 14 polymorphic allozyme loci (Ak-1, Gapdh, α-Gpd, Got-1, Got-2, Hk, Idh-1, Idh-2, Mdh-1, Mdh-2, Me-1, Me-2, Pgi, and Pgm).

Analysis

The analysis follows steps (i) to (iii) above. To handle the possibility of biases introduced by significantly divergent loci found in step (iii), two additional steps were included:

(iv) Determine which loci are significantly divergent and assess potential biological causes for divergence. If selection is suspected based on independent evidence, then remove the most extreme locus and repeat steps

(i) to (iii) until no loci are found to be significantly divergent. This ensures that the estimate obtained from the average, $\hat{F}^*_{ST}$, is not biased by the loci thought to be under significant selection.

(v) Using the final estimates $\hat{F}^*_{ST}$, $\hat{F}^*_{IS}$ and $\hat{\mathbf{Q}}$, estimate $\langle Nm \rangle$, the unbiased gene flow rate, and $\langle Nb_k \rangle$, the bias on the kth locus.

If loci are removed in step (iv), the estimates of $\hat{F}_{ST}$ and $\hat{F}_{IS}$ change, possibly altering the significance levels in the next iteration. Removal of more than one locus at a time can inadvertently result in the removal of a locus that would not deviate significantly from the null expectation.

Results

The pooled data from all subspecies yielded weighted average *F*-statistics of $\hat{F}_{ST} = 0.0484$ and $\hat{F}_{IS} = 0.088$, which generates an estimate of $\langle Nm \rangle = 4.92$ (2.56–4.96; 95% confidence interval from the null distribution). The first pass of 1000 replicates yielded three loci that had significantly low $\hat{F}_{ST,k}$: Mdh-1 ($P < 0.001$), Pgi ($P < 0.001$) and Idh-1 ($P = 0.024$), and one marginally significant locus, Pgm ($P = 0.070$). The ecological genetics of allozyme loci is unstudied in *Coenonympha*, but Pgi is known to be under selection for heterozygote superiority in *Colias* butterflies (Watt *et al*. 1983, 1985, 1996; Watt 1985) and there is some evidence that Mdh-1 allozyme variants have functional differences that may be under selection in honeybees (Harrison *et al*. 1996). On the hypothesis that selection might be causing similar patterns here, I removed both highly divergent loci, Mdh-1 and Pgi, from the analysis. New *F*-statistics, $\hat{F}^*_{ST} = 0.0555$ and $\hat{F}^*_{IS} = 0.115$, were estimated and a second pass of 1000 replicates was made. Significantly low $\hat{F}_{ST,k}$ recurred in Idh-1 ($P = 0.014$) and appeared in Pgm ($P = 0.038$), now significant because of the new, higher average $\hat{F}^*_{ST}$. I then removed Idh-1, even though selection on this locus has been unstudied in butterflies and evidence of it has been found only occasionally in other organisms (e.g. Bergmann & Gregorius 1993; DaCunha & DeOliveira 1996). I obtained new *F*-statistics, $\hat{F}^*_{ST} = 0.0589$ and $\hat{F}^*_{IS} = 0.114$, and made a third pass through the analysis. This time a significantly low $\hat{F}^*_{ST,k}$ was found again in Pgm ($P = 0.042$). Pgm is under selection in *Drosophila* (Verrelli & Eanes 2001) and is likely to be in butterflies as well(Watt *et al*. 1985, 1996). I removed Pgm and, using new *F*-statistics, $\hat{F}^*_{ST} = 0.0662$ and $\hat{F}^*_{IS} = 0.144$, made a final pass through the analysis. This time no additional deviant loci were identified. Figure 1 shows the final null distributions for all loci, and the observed $\hat{F}_{ST,k}$ values for each locus are shown as arrows on the axes.

Table 2 shows the estimated bias on each locus, $\langle Nb_k \rangle$, and the corresponding significance levels, and the confidence limits around the $\langle Nb_k \rangle$ estimates. These confidence limits are the strengths of biases, in units comparable to gene flow units, *Nm*, that would be needed to achieve statistical significance. It can be seen that in most cases, only biases comparable in magnitude to high gene flow rates would be detected. In *Coenonympha*, Mdh-1 showed the least differentiation among populations relative to the average locus of $Nb = 7.75$. Pgi was next at $Nb = 3.49$, followed by Pgm at $Nb = 1.15$ and Idh-1 at $Nb = 0.68$.

Table 2 Single-locus estimates of genetic structure among populations of the butterfly *Coenonympha tullia*, using data from Porter & Geiger (1988), from null distributions seeded using the weighted average $\hat{F}_{ST}$ over alleles and loci as described in the Appendix

Locus k	$\hat{F}_{ST}$	$\tilde{F}_{ST(0.025)}$	$\tilde{F}_{ST(0.975)}$	$\hat{k}_{LK}$	$\langle Nb_k \rangle$	$\langle Nb_{k(0.025)} \rangle$	$\langle Nb_{k(0.975)} \rangle$	P
Ak-1	0.077	0.046	0.199	6.5	–0.55	–2.52	1.67	0.484
Gapdh	0.013	0.000	0.243	18.2	16.0	–2.75	infinite	0.280
Got-1	0.061	0.051	0.180	4.9	0.35	–2.39	1.10	0.138
Got-2	0.057	0.000	0.288	23.7	0.62	–2.91	infinite	0.882
a-Gpd	0.029	0.017	0.253	16.9	4.79	–2.79	11.0	0.190
Hk	0.034	0.000	0.262	18.2	3.52	–2.83	infinite	0.660
Idh-1	0.043	0.054	0.177	4.26	2.07	–2.37	0.88	0.006*
Idh-2	0.045	0.031	0.232	11.7	1.84	–2.70	4.25	0.196
Mdh-1	0.019	0.042	0.198	7.94	9.14	–2.52	2.19	< 0.001*
Mdh-2	0.052	0.024	0.233	13.0	0.99	–2.70	6.49	0.416
Me-1	0.093	0.039	0.207	7.8	–1.08	–2.57	2.66	0.960
Me-2	0.036	0.0	0.237	21.0	3.14	–2.72	infinite	0.794
Pgi	0.029	0.053	0.160	3.60	4.88	–2.22	0.97	< 0.001*
Pgm	0.040	0.045	0.177	5.12	2.55	–2.37	1.73	0.028*
F^*_{ST}	0.066	0.055	0.134	1.71	0.0	–1.91	0.77	0.248

The variable names follow Table 1, and $\hat{k}_{LK}$ estimates the Lewontin–Krakauer proportionality factor that relates F_{ST} to its sampling variance. The distributions are shown in Fig. 1.

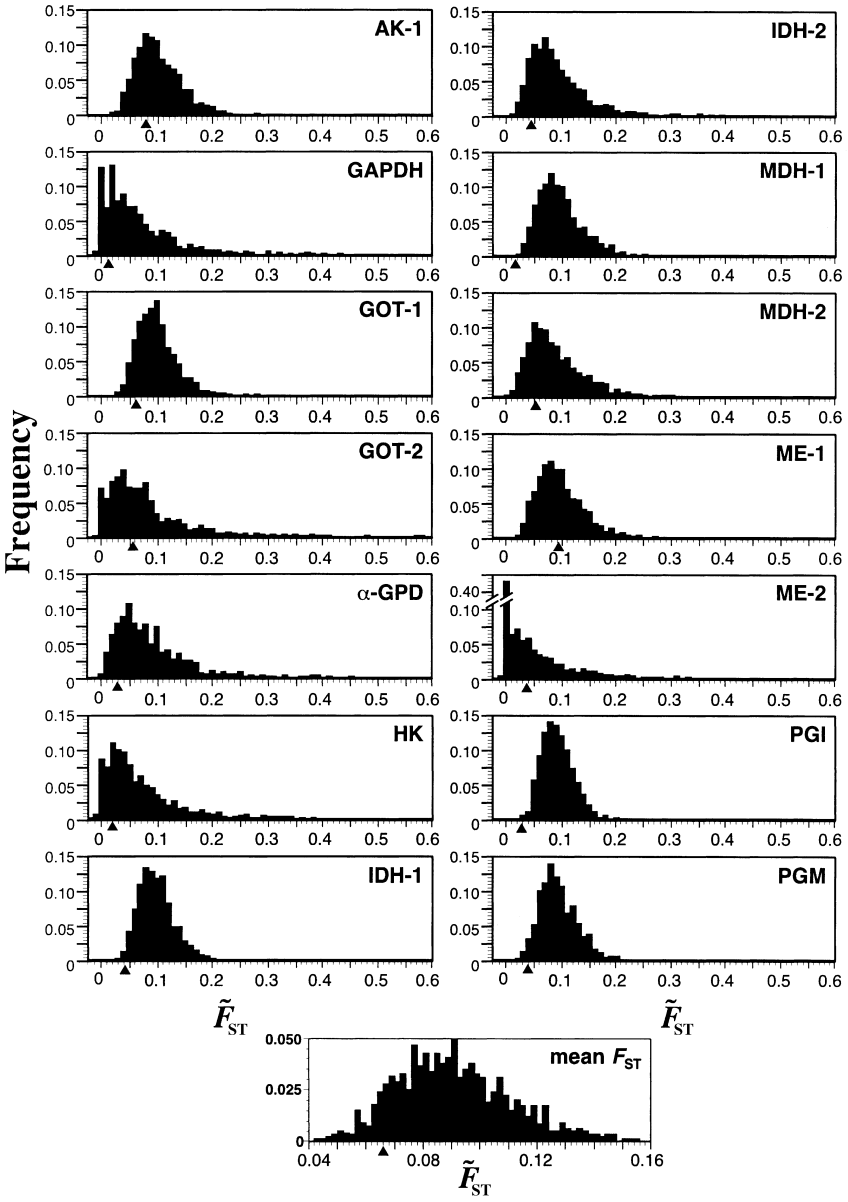

Fig. 1 Null resampling distributions of $\tilde{F}_{ST}$ for each locus in the butterfly *Coenonympha tullia* (data from Porter & Geiger 1988), seeded using the weighted average $\hat{F}_{ST}$ over alleles and loci as described in the Appendix. Observed $\hat{F}_{ST}$ values are marked by an arrow on each axis and summary statistics are given in Table 2. The distribution of the resampled mean $\hat{F}^{*}_{ST}$ has different axis scales. Idh-1, Mdh-1, Pgi and Pgm have $\tilde{F}_{ST}$ significantly lower than expected.

Interpretation

Selection vs. migration vs. nonequilibrium dynamics. This test detects deviations from the infinite-island model's assumptions, but if significant deviations are detected, it is not a simple matter to explain the basis of the deviations. Even though Mdh-1, Pgi, Idh-1 and Pgm all deviate from the null expectation in the direction characteristic of balancing selection, other violations of island-model assumptions, especially migratory patterns, could also cause this pattern. Better ecological data on population structure are needed before anything definitive can be stated. However, balancing selection would seem to be a more plausible explanation for this pattern than source-sink or extinction/recolonization dynamics, given the little we do know of the biology of this widely distributed, common species.

Statistical power

These tests are not particularly powerful in their ability to detect deviations from the island model, as can be seen by the rather wide confidence limits around the $\langle Nb_k \rangle$ estimates (Table 2). On the other hand, this low power for detection implies that estimates of gene flow rates, which have wide confidence limits anyway, are relatively insensitive to moderate variations among loci in the degree that they are affected by violations of island-model assumptions.

Shape variation among null distributions

To implement the Lewontin–Krakauer test for selection, it has been necessary to make assumptions about the shapes of the null $\tilde{F}_{ST}$ distributions for different loci (Baer 1999), treating them as constant (Lewontin & Krakauer 1973) or grouping them into a few discrete classes (Ewens 1977; Baer 1999). In fact, these shapes vary considerably and rather continuously. This can be seen by inspecting Fig. 1 and by examining in Table 2 the values of $\hat{k}_{LK,k}$, which are estimates of the Lewontin–Krakauer proportionality coefficients k_{LK} for each locus k. This constant describes the breadth of the null $\tilde{F}_{ST}$ distributions. The $k_{LK,k}$ were estimated using $\hat{k}_{LK,k} = (n-1)\langle Var(\tilde{F}_{ST,k})\rangle / \hat{F}^2_{ST,k}$ where n is the number of sampled populations and $\langle Var(\tilde{F}_{ST,k})\rangle$ was estimated directly from the 1000 replicates of each distribution. This variation among $\hat{k}_{LK,k}$ scores is mainly an effect of the varying numbers of alleles among loci (McDonald 1994) and different average allele frequencies ($\hat{\mathbf{Q}}_k$), as the sampling regime of the original data was similar among loci.

Discussion

Like the Lewontin–Krakauer test, this test detects significant deviations from the assumptions of the infinite island model of population structure. It is a more appropriate test because it accounts for the fact that the sampling distribution of F_{ST} depends on allele frequencies. It also incorporates sampling error for each locus separately and it identifies the loci responsible for any deviations. However, whether the deviations are the result of selection, migration patterns that differ from the island model's, or of violations of other island-model assumptions, requires additional information. In an analogous way, the detection of deviations from Hardy–Weinberg genotype frequencies using χ^2 tests does not of itself demonstrate whether selection, inbreeding, Wahlund effects, or some other mechanism is responsible. Nevertheless, using this island-model test, biases can be estimated. If there is independent justification from information on selection or population ecology, the deviant loci or populations can be removed from the analysis so that a better estimate of neutral gene flow rates can be obtained. Below I compare this test to its predecessors and discuss its limitations.

Variance-based approaches

Lewontin & Krakauer (1973) developed their statistical test based on an approximation of the sampling variance of F_{ST},

$$Var(F_{ST}) = \frac{k_{LK}}{n-1} F^2_{ST}, \quad (8)$$

where k_{LK} is a constant of proportionality ($k_{LK} = 2$ in their paper) and n is the number of sampled subpopulations. To reach this formula, they assumed that the multilocus F_{ST} distribution would be approximately normal, the number of loci sampled would be large and the number of individuals would be sufficiently large that there is negligible error in the estimation of allele frequencies.

When the Lewontin–Krakauer test is used with limited data sets and spatial population structure, the scaling factor k_{LK} is asked to do too much. Perhaps this is best seen in how k_{LK} is affected by relaxing the simplifying assumptions. Jacquard (1974) and Ewens & Feldman (1976) showed analytically that the value of k_{LK} depends on the kurtosis of the distribution of allele frequencies at each locus (from which F_{ST} is calculated), which in turn depends on the average allele frequencies, $\mathbf{Q}$, among populations. These shape differences can be seen in Fig. 1. The null hypothesis might be inadvertently rejected by the Lewontin–Krakauer test because of the sampling properties of low-frequency alleles. Ewens (1977) provides values of k_{LK} for various global allele frequencies $\mathbf{Q}$ in the two-allele case, and these have been used as rough corrections for the Lewontin–Krakauer test. Baer (1999) used the expected k_{LK} values for $\mathbf{Q} = \{0.5, 0.5\}$ and $\mathbf{Q} = \{0.9, 0.1\}$ to bound his tests for locus-specific effects in fish.

In addition, the sample variance of $\hat{F}_{ST}$ decreases in proportion to the number of loci and individuals sampled (Weir & Cockerham 1984; Beaumont & Nichols 1996; Fernando 1997). This effect is not accounted for explicitly in equation (9), so it too is tacitly absorbed into k_{LK}, leading to underestimation of the expected $Var(F_{ST})$. The number of alleles (McDonald 1994) also has an important effect on k_{LK}. This can be seen in the variety of shapes of the single-locus distributions in Fig. 1, and in the diversity of $\hat{k}_{LK}$ estimates from these distributions in Table 2; $\hat{k}_{LK}$ are all close to 2.4 in the simulated data of Table 1 (not shown). In small data sets, rejection of the null hypothesis using the Lewontin–Krakauer test could be the result of sampling rather than underlying biological causes.

Robertson (1975b) showed that the Lewontin–Krakauer test assesses deviations from the neutral island model's assumptions in addition to selection. When there are correlations among subpopulations in their (neutral) allele frequencies, then still assuming normality of the F_{ST} distribution, the expected variance increases to $Var(F_{ST}) = k_{LK} F^2_{ST} [1/(n-1) + Var(r)]$. Here, $Var(r)$ is the variance of the correlation of allele frequencies among pairs of populations. These correlations arise when the migration patterns of the island model do not hold, such as a phylogenetic branching pattern as modelled by Robertson (1975a,b; Bowcock *et al*. 1991), isolation-by-distance (Wright 1943; Slatkin 1993), stepping-stone (Kimura & Weiss 1964), source-sink, extinction/colonization (Wade & McCauley 1988; Whitlock & McCauley 1990; Whitlock 1992), and some refugee (Porter 1999) models of population structure. In practice, when deviations from the island model's

migration patterns are ignored, Robertson's (1975b) $Var(r)$ is absorbed into k_{LK}, giving

$$k^*_{LK} = k_{LK}(1 + (n-1)\,Var(r)) = \frac{(n-1)\,Var\,(F_{ST})}{F^2_{ST}}.$$

$k^*_{LK} > k_{LK}$ so the use of expected k_{LK} leads to underestimation of the expected $Var(F_{ST})$ in the Lewontin–Krakauer test. Rejection of the null hypothesis could therefore be because of violation of the assumptions about migration patterns rather than natural selection (Robertson 1975a). It is feasible to correct for this effect by estimating $Var(r)$. Reynolds *et al*. (1983) provide an estimator for F_{ST} between population pairs that describes this correlation r, and the variance of this correlation can be obtained from the matrix of these estimates. Correcting for $Var(r)$ does not provide escape from the purely statistical biases.

Resampling approaches

In my approach, these statistical difficulties are circumvented by resampling from the theoretical null distribution of allele frequencies (equation 4) under the island model, using the sampling patterns and local inbreeding estimates ($\hat{F}_{IS}$) from the original data. The null distribution of F_{ST} is obtained directly without relying on weakly justified assumptions about its variance. Consequently, rejection of the null hypothesis can usually be attributed to violations of the biological assumptions, especially the presence of selection and deviations from the island model's migration patterns.

Others have adopted similar approaches. Bowcock *et al*. (1991) found significant locus-specific effects in human random fragment length polymorphism frequencies, which they attributed to selection. They controlled for migration patterns among their five diverse human population samples by first determining the average degree of phylogenetic divergence among these populations using genetic distances. This allowed the estimation of allele frequencies for these populations, and from these the null distributions of allele frequencies were obtained. They then calculated the probability density of F_{ST} for all possible allele frequencies using equation (2), and used this to interpret the observed $\hat{F}_{ST}$ scores for each locus. To the extent that their method captures the true migration patterns and the steady-state assumption holds, it is a test of selection *per se*, rather than a more general test of deviations from neutral island-model patterns.

Beaumont & Nichols (1996) used a coalescent-based simulation, seeded with average allele frequencies for the global population ($\hat{\mathbf{Q}}$), to obtain neutral allele frequencies (analogous to $\hat{\mathbf{q}}$) for a large array of populations. They then sampled repeatedly from this array, ignoring local inbreeding, following the sampling design of the original data. To circumvent difficulties with comparing loci having different numbers of alleles, they analysed the null distribution of the ratio of F_{ST} to heterozygosity for each locus, rather than dealing with the null distributions of F_{ST} directly. Vitalis *et al*. (2001) derive the analytical model for the case of population pairs and no gene flow following separation of the populations. The coalescent approach is particularly valuable because it allows the null distributions of F_{ST} to be generated under a wide variety of hypotheses about the historical relationships and demographics of the populations.

Balanovskaya & Nurbaev (1998a,b) generated distributions of F_{ST}, seeded with $\hat{\mathbf{Q}}$ estimates from human populations based on resampling allele frequencies from the original data. Their distributions are based on resampling 10^6 individuals per replicate and therefore do not account for the effects of sampling. These are probably the best approximations we have of the limiting shape of the parametric distribution of F_{ST}. They use these distributions to infer selection on marker loci, ignoring the consequences of sampling and of violating the assumed island-model migration patterns. McDonald (1994) and Cockerham & Weir (1993) provide information on the shapes of null F_{ST} sampling distributions based on simulations under a limited range of conditions.

Limitations

How strong must the bias Nb_k be to detect it in a given sample? A crude sense of the statistical power may be gained by inserting into equation (7) the F-values for the 95% confidence limits taken from the resampled distribution. These are

$$\begin{aligned}\left\langle Nb_{k(0.025)}\right\rangle &= \frac{1}{4}(\tilde{F}^{-1}_{ST(0.975),k} - \hat{F}^{*-1}_{ST}),\\ \left\langle Nb_{k(0.975)}\right\rangle &= \frac{1}{4}(\tilde{F}^{-1}_{ST(0.025),k} - \hat{F}^{*-1}_{ST}),\end{aligned} \qquad (9)$$

where $\tilde{F}_{ST(0.025),k}$ and $\tilde{F}_{ST(0.975),k}$ are the lower and upper confidence limits of F_{ST} obtained from the resampled distribution for locus k, and $\langle Nb_{k(0.025)}\rangle$ and $\langle Nb_{k(0.975)}\rangle$ are the estimated threshold strengths of biases that would be detected in the observed data. A proper power analysis would require comparison to the expected distribution of F_{ST} under the correct fitness function and migration pattern, which are usually unknown. Using this approach, it is clear that even for large data sets, only rather strong biases will be detected. This is a consequence of the high sampling variance around F_{ST}, and even with large data sets, there seems to be little that one can do statistically to gain better acuity.

Inherent in all these approaches, parametric or numerical, and extending also to the estimation of gene flow, is the assumption that the average F_{ST} describes patterns of neutral allelic variation among populations (Balanovskaya & Nurbaev 1998a). The individual-locus F_{ST} biases (Nb_k) can

only be estimated relative to the global average. It cannot be determined from frequency data alone whether the global average F_{ST} is neutral, or whether the average locus experiences some degree of balancing or differentiating selection. Independent data are therefore needed to calibrate the location of the neutral point along the F_{ST} distribution. However, explicit fitness models for several loci are likely to be needed before a proper calibration of the F_{ST} distribution can be attempted.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation (DEB 9981608, DEB 0075451), the U.S. Department of Agriculture, National Research Initiative Award no. 990–2471, and Cooperative State Research Extension and Education Service awarded through the Massachusetts Agricultural Experiment Station under Project No. MAS00789, Paper no. 3270. I am grateful to P. Beerli, J. Brookfield, L. Excoffier, N. Johnson, S. Karl, B. Wang, D. Weinreich and anonymous referees for their thoughtful discussions and comments. I am especially indebted to L. Excoffier and the anonymous referees, who were insightful, generous and patient. They helped me to improve the manuscript greatly, and taught me a lot about statistical biology.

References

Abramowitz M, Stegun IA (1967) *Handbook of Mathematical Functions*, 6th printing. National Bureau of Standards, Washington D.C.

Baer CF (1999) Among-locus variation on F_{ST}: fish, allozymes and the Lewontin–Krakauer test revisited. *Genetics*, **152**, 653–659.

Balanovskaya E, Nurbaev SD (1998a) Selective structure of the gene pool. III. Estimation from F_{ST}-statistics with the use of numerical resampling. *Russian Journal of Genetics*, **34**, 1434–1446.

Balanovskaya E, Nurbaev SD (1998b) Selective structure of the gene pool. IV. Estimation from selection intensity index R_S. *Russian Journal of Genetics*, **34**, 1559–1573.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London, Series B*, **263**, 1619–1626.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences USA*, **98**, 4563–4568.

Bergmann F, Gregorius HR (1993) Ecogeographical distribution and thermostability of isocitrate dehydrogenase (Idh) allozymes in European silver fir (*Abies alba*). *Biochemical Systematics and Ecology*, **21**, 597–605.

Bowcock AM, Kidd JR, Mountain JL *et al.* (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proceedings of the National Academy of Sciences USA*, **88**, 839–843.

Burr TL (2000) Quasi-equilibrium theory for the distribution of rare alleles in a subdivided population: justification and implications. *Theoretical Population Biology*, **57**, 297–306.

Cavalli-Sforza LL (1966) Population structure and human evolution. *Proceedings of the Royal Society of London, Series B*, **164**, 362–379.

Cheng RCH, Feast GM (1979) Some simple gamma variate generators. *Applied Statistics*, **28**, 290–295.

Cockerham CC (1969) Variance of gene frequencies. *Evolution* **23**, 72–84.

Cockerham CC, Weir BS (1993) Estimation of gene flow from *F*-statistics. *Evolution*, **47**, 855–863.

DaCunha GL, DeOliveira AK (1996) Citric acid cycle: a mainstream metabolic pathway influencing life span in *Drosophila melanogaster*? *Experimental Gerontology*, **31**, 705–715.

Efron B, Tibshirani RJ (1993) *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87–112.

Ewens WJ (1977) Population genetics theory in relation to the neutralist-selectionist controversy. *Advances in Human Genetics*, **8**, 67–134.

Ewens WJ, Feldman MW (1976) The theoretical assessment of selective neutrality. In: *Population Genetics and Ecology* (eds Karlin S, Nevo E), pp. 303–337. Academic Press, New York.

Fernando N (1997) Sampling variation of *F*-statistics. MS Thesis, Bowling Green State University, Bowling Green, Ohio.

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**, 915–925.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Harrison JF, Nielsen DI, Page RE (1996) Malate dehydrogenase phenotype, temperature and colony effects on flight metabolic rate in the honey-bee, *Apis mellifera*. *Functional Ecology*, **10**, 81–88.

Hudson RR (1998) Island models and the coalescent process. *Molecular Ecology*, **7**, 413–418.

Jacquard A (1974) *The Genetic Structure of Populations* [translated by Charlesworth D, Charlesworth B]. Springer-Verlag, New York.

Johnson ME (1987) *Multivariate Statistical Simulation*. Wiley, New York.

Johnson NA, Porter AH (2000) Rapid speciation via parallel, directional selection on regulatory genetic pathways. *Journal of Theoretical Biology*, **205**, 527–542.

Jordan WC, Verspoor E, Youngson AF (1997) The effect of natural selection on estimates of genetic divergence among populations of the Atlantic salmon. *Journal of Fish Biology*, **51**, 546–560.

Kelly JK (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.

Kimura M, Weiss GH (1964) The stepping-stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, **49**, 561–576.

Lewontin RC ME (1974) *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

McDonald JH (1994) Detecting natural selection by comparing geographic variation in protein and DNA polymorphisms. In: *Non-Neutral Evolution* (ed. Golding B), pp. 88–100. Chapman & Hall, New York.

Nei M, Chakravarti A (1977) Drift variances of F_{ST} and G_{ST} statistics obtained from a finite number of isolated populations. *Theoretical Population Biology*, **11**, 307–325.

Nei M, Maruyama T (1975) Lewontin–Krakauer test for neutral genes. *Genetics*, **80**, 395.

Nei M, Chakravarti A, Tateno Y (1977) Mean and variance of F_{ST} in a finite number of incompletely isolated populations. *Theoretical Population Biology*, **11**, 291–306.

Porter AH (1999) Refugees from lost habitat and reorganization of genetic population structure. *Conservation Biology*, **13**, 850–859.

Porter AH, Geiger HJ (1988) Genetic and phenotypic population structure of the *Coenonympha tullia* group (Lepidoptera: Nymphalidae; Satyrinae) in California: no evidence for species boundaries. *Canadian Journal of Zoology*, **66**, 2751–2765.

Porter AH, Johnson NA (2002) Speciation despite gene flow when developmental pathways evolve. *Evolution*, **56**, 2103–2111.

Rannala B (1996) The sampling theory of neutral alleles in an island population of fluctuating size. *Theoretical Population Biology*, **50**, 91–104.

Rannala B, Hartigan JA (1996) Estimating gene flow in island populations. *Genetical Research*, **67**, 147–158.

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, **105**, 767–779.

Rice W (1989) Analyzing tables of statistical tests. *Evolution*, **43**, 223–225.

Robertson A (1975a) Remarks on the Lewontin–Krakauer test. *Genetics*, **80**, 396.

Robertson A (1975b) Gene frequency distributions as a test of selective neutrality. *Genetics*, **81**, 775–785.

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of selective neutrality for DNA polymorphism data. *Genetics*, **141**, 413–429.

Slatkin M (1987) Gene flow and the genetic structure of natural populations. *Science*, **236**, 787–792.

Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47**, 264–279.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tsakas S, Krimbas CB (1975) How many genes are selected in populations of *Dacus oleae*. *Genetics*, **79**, 675–679.

Tsakas S, Krimbas CB (1976) Testing the heterogeneity of *F* values: a suggestion and a correction. *Genetics*, **84**, 399–401.

Verrelli BC, Eanes WF (2001) The functional impact of Pgm amino acid polymorphism on glycogen content in *Drosophila melanogaster*. *Genetics*, **159**, 201–210.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158**, 1811–1823.

Wade MJ, McCauley DE (1988) Extinction and recolonization: their effects on the genetic differentiation of local populations. *Evolution*, **42**, 995–1005.

Watt WB (1985) Adaptation at specific loci. Differential mating success among glycolytic allozyme genotypes of *Colias* butterflies. *Genetics*, **109**, 157–175.

Watt WB, Cassin RC, Swan MS (1983) Adaptation at specific loci. III. Field behavior and survivorship differences among *Colias Pgi* genotypes are predictable from in vitro biochemistry. *Genetics*, **103**, 725–739.

Watt WB, Carter PA, Blower SM (1985) Adaptation at specific loci. IV. Differential mating success among glycolytic allozyme genotypes of *Colias* butterflies. *Genetics*, **109**, 157–175.

Watt WB, Donohue K, Carter PA (1996) Adaptation at specific loci. VI. Divergence vs. parallelism of polymorphic allozymes in molecular function and fitness-component effects among *Colias* species (Lepidoptera, Pieridae). *Molecular Biology and Evolution*, **13**, 699–709.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Whitlock MC (1992) Non-equilibrium population structure in forked fungus beetles: extinction, colonization, and the genetic variance among populations. *American Naturalist*, **139**, 952–970.

Whitlock MC, McCauley DE (1990) Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. *Evolution*, **44**, 1717–1724.

Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. *Heredity*, **82**, 117–125.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Wright S (1937) The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences USA*, **23**, 307–320.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–128.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Wright S (1969) *Evolution and the Genetics of Populations*, Vol. 2. *The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

Wright S (1978) *Evolution and the Genetics of Populations. Vol. 4. Variability within and among Natural Populations*. University of Chicago Press, Chicago.

Appendix

Generating the null distribution of F_{ST}

Consider a matrix **g** of sample genotypes with elements g_{ijk} from r loci, indexed by $k = \{1, \ldots, r\}$, obtained from n subpopulations, indexed by $i = \{1, \ldots, n\}$, each with n_i individuals indexed by $j = \{1, \ldots, n_i\}$. The sampling design may be unbalanced at any level, and it is common in empirical studies that not all individuals are sampled at all loci. The goal is to generate null distributions of F_{ST} for each locus by replicating this sampling design while sampling from the distributions of Equations 4 and 5. It is possible to use the observed sample values $\hat{\mathbf{Q}}$, $\hat{F}_{ST}$ and $\hat{F}_{IS}$ to seed these equations, and use a large number of replicates to generate the null distributions. However, this ignores important sources of sampling variation around $\hat{\mathbf{Q}}$, $\hat{F}_{ST}$ and $\hat{F}_{IS}$. To capture this variation, I first take a standard bootstrap sample (Efron & Tibshirani 1993), resampling over populations and individuals within populations. From these I obtain new estimates of $\hat{\mathbf{Q}}$, $\hat{F}_{ST}$ and $\hat{F}_{IS}$, labelled with primes ($\hat{\mathbf{Q}}'$, $\hat{F}'_{ST}$ and $\hat{F}'_{IS}$). These are inserted into equation (4) in place of the parameters **Q** and F_{ST} to create the distribution

$$\varphi(\mathbf{q} | \hat{F}'_{ST}, \hat{\mathbf{Q}}'), \tag{A1}$$

from which a parametric-bootstrap sample is then taken. A vector of allele frequencies, $\tilde{\mathbf{q}}$, is chosen randomly from (A1) by taking ratios of random deviates from a gamma distribution. A random gamma deviate X_a with parameters $\alpha = \hat{\mathbf{Q}}_a(\hat{F}_{ST}^{-1} - 1)$, $\beta = 1$ is first chosen for each allele $a = \{1, \ldots, h\}$. A vector of randomly sampled, β-distributed allele frequencies $\tilde{\mathbf{q}}$ is then obtained using $\tilde{\mathbf{q}} = X_a/(X_1 + \ldots + X_h)$ (see Abramowitz & Stegun 1967, p. 944). Algorithms for obtaining random gamma deviates are available in the public domain (Cheng & Feast 1979; Johnson 1987). This way, a set of resampled allele frequency vectors $\{\tilde{\mathbf{q}}_1, \ldots, \tilde{\mathbf{q}}_n\}$ may be obtained to represent each of the n subpopulations in the original sample.

These $\tilde{\mathbf{q}}_i$ represent parametric allele frequencies of sample populations, and the extent of sampling variance around them depends on the number of genotypes taken from each. For each of the n subpopulations, we then randomly draw n_i genotypes from the distribution of equation (5) [or if $\hat{F}'_{IS} < 0$, from equation (A3) below],

$$\phi(g_i | \tilde{\mathbf{q}}_i, \hat{F}'_{IS}), \tag{A2}$$

replacing $\mathbf{q}_i$ with the resampled vector $\tilde{\mathbf{q}}_i$ and replacing the parametric F_{IS} with the its estimated value, $\hat{F}'_{IS}$, from the bootstrap data sample. To obtain a random genotype $\tilde{g}_{ijk}$, equation (A2) is converted to a cumulative distribution, then a random number ρ is chosen from a uniform distribution, and the genotype is found that yields the cumulative sum ρ. This process is repeated for each of the n_i individuals in subpopulation i, and the process is repeated for each subpopulation and locus in the original sample. These genotypes are then used to calculate $\tilde{F}_{ST,k}$ for each locus. In the analyses presented in the text, I repeated this entire standard-bootstrap/parametric-bootstrap resampling scheme 1000 times to obtain the null sampling distribution of each $\hat{F}'_{ST,k}$.

Expected genotype frequencies when $F_{IS} < 0$

With inbreeding coefficient F_{IS}, the genotype frequency of homozygotes carrying allele a is $\gamma_{aa} = q_a^2 + F_{IS}\, q_a(1 - q_a)$. Since the genotype frequency γ_{aa} cannot be less than zero, substituting $\gamma_{aa} = 0$ and rearranging yields a minimum F_{IS} at $F_{\min,a} = -q_a/(1 - q_a)$ if $q_a = 0.5$, or $F_{\min,a} = -(1 - q_a)/q_a$ if $q_a \geq 0.5$. Ignoring this constraint generates negative genotype frequencies using equation (5). With multiple alleles of unequal frequency, each allele is constrained to a unique minimum F_{IS}. This is further complicated when generating expected genotype frequencies from allele frequencies $\mathbf{q}_i$ and F_{IS}, because the input value of F_{IS} might fall above, below, or among these minima. Here is a method that allows F_{IS} for each homozygote frequency γ_{ii} to go to its minimum, adjusting the heterozygote frequencies accordingly.

Label the alleles with the lowest frequency first, $0 \leq q_1 \leq q_2 \leq \ldots \leq q_h \leq 1$. This allows us to work sequentially with alleles, first distributing available allele copies proportionally into the respective heterozygotes (with availability determined by that allele's minimum F_{IS}), then rescale the remaining allele frequencies into the available genotype frequency space for iteration of the next allele. The scaled allele frequency for each iteration a is $q_b^{(a)} = q_b \left(1 - \sum_{i=1}^{a-1} q_i\right)^{-1}$, where a and b are indices of alleles, and $a = b$. For the first iteration, $q_b^{(1)} = q_b$. A scaling factor $\lambda^{(a)} = \lambda^{(a-1)}(1 - 2q_{a-1}^{(a-1)})$ describes the available frequency space in iteration a, with $\lambda^{(1)} = 1$.

With this notation, we can define a minimum F_{IS} for each allele in its rescaled frequency space as $F_{\min}^{(a)} = -q_a^{(a)}/(1 - q_a^{(a)})$. Finally, we define a threshold allele t (remembering that allele frequencies are arranged in ascending order), such that $F_{\min}^{(t-1)} < F_{IS} < F_{\min}^{(t)}$. From these rescaled minima, define $F_{IS}^{(a)} = \max(F_{IS}, F_{\min}^{(a)})$ to account for the possibility that the input F_{IS} may not be as low as the minimum for allele a. These lead to a function for genotype frequencies analogous to equation (5), dropping the subscript for populations,

$$\phi(\gamma|\mathbf{q},F_{IS}) \equiv \begin{cases} \gamma_{aa} = \lambda^{(a)}[(q_a^{(a)})^2 + F_{IS}^{(a)} q_a^{(a)}(1 - q_a^{(a)})], \; a < t \\ \gamma_{aa} = \lambda^{(t)}[(q_a^{(t)})^2 + F_{IS} q_a^{(t)}(1 - q_a^{(t)})], \; a \geq t \\ \gamma_{ab} = 2\lambda^{(a)}[(1 - F_{IS}^{(a)}) q_a^{(a)} q_b^{(a)}], \; a < b, \; a < t \\ \gamma_{ab} = 2\lambda^{(t)}[(1 - F_{IS}) q_a^{(t)} q_b^{(t)}], \; a < b, \; a \geq t \end{cases} \quad \text{(A3)}$$

This constrains homozygote genotype frequency to $\gamma_{aa} = 0$, except homozygotes of the allele h, with the highest frequency q_h, have $\gamma_{hh} > 0$. When $F_{IS} \geq 0$, equation (A3) reduces to equation (5).