Lecture 08:
Detecting selection with marker data.
2: Polymorphism-based tests: I

UNE course:

The search for selection

3 -- 7 Feb 2020

Bruce Walsh (University of Arizona)

jbwalsh@email.arizona.edu

Outline

- Overview of polymorphism-based tests
- Tests based on allele-frequency change
  - Waples adjusted tests
  - Fisher-Ford test
  - Schaefer linear trend test
- Tests based on spatial variation
  - Lewontin-Krakauer tests
  - Allele-environmental associations
- Tests based upon diversity pattern over a chromosomal region

Polymorphism-based tests

- Several different sampling approaches are used in attempts to detect ongoing (or very recent) selection
  - A population sampled at two (or more) time points
  - A series of populations sampled a single time
  - A single sample from a population

Temporal or spatial sampling

1) Excessive allele-frequency change. The first formal test of selection was proposed by Fisher and Ford (1947), who used the machinery developed in Chapter 2 for the divergence under drift to test for excessive change in a time-series of allele frequencies from a single population. While perhaps the most unambiguous signature of selection, this approach requires long-term monitoring of a population and having some reasonably independent estimate of N_e. The ever-increasing availability of **ancient DNA (aDNA)** samples opens up exciting new data sets for this type of analysis (Mathieson et al. 2015, Malaspinas 2016; Schraiber et al. 2016).

2) Excessive allele-frequency divergence. Lewontin and Krakauer (1973) proposed using the divergence between a series of contemporaneously sampled populations (presumably from a common ancestor) to test for selection. The machinery from Chapter 2 predicts the expected divergence under drift, as measured by Wright's F_{ST} statistic for population structure. Loci displaying excessive F_{ST} values relative to drift are selection candidates. Using an incorrect model of population structure can seriously compromise these tests.

Single population sample

The above two categories require samples from multiple populations (either temporally or spatially), which limits their widespread use. A less demanding design is a single population sample, as employed by the three remaining categories.

3) Chromosomal spatial patterns of variation. As detailed in Chapter 8, a sweep leaves a characteristic decrease in polymorphism around a selected site, and a number of formal likelihood tests are based on the expected pattern of the nucleotide diversity, π, as a function of the recombination distance, c, from the sweep (Equation 8.8a). Early versions of these tests assumed that the population was in mutation-drift equilibrium at the start of the sweep, while more recent versions have relaxed this strong assumption.

The final two categories divide tests by whether they assume an infinite-sites, or an infinite-alleles, framework, using the neutral equilibrium results for these models developed in Chapter 2. Recall that the infinite-sites framework considers a sequence as a series of separate sites (e.g., SNPs), while the infinite-alleles framework treats each different DNA sequence (haplotype) as a different allele (Figure 2.9). Both models assume that the region being considered is small enough that recombination within the sample can be ignored. Given the large (and diverse) number of tests in both of these categories, each section reviewing these different approaches concludes with a summary table of proposed tests (Table 9.1 for infinite-sites and Table 9.3 for haplotypes).

4) Changes in the site-frequency spectrum. Under the infinite-sites model, the frequency spectrum of neutral sites at mutation-drift equilibrium is given by the Watterson distribution (Equation 2.34). Starting with Tajima (1989), a number of tests have been proposed that search for shifts in this spectrum following a sweep, such as an excessive number of sites with rare alleles or with high-frequency derived alleles. The major complication with this class of tests is that changes in population demography (such as a recent expansion or contraction) or the presence of population structure (migration between partly isolated populations) can mimic signatures of selection.

5) Tests based on haplotype information. Under the infinite-alleles model, the number of alleles (haplotypes) in a sample at mutation-drift equilibrium is given by the Ewens sampling formula (Equation 2.30a) and their allele-frequency spectrum by Equation 2.33b. Starting with Ewens (1972) and Watterson (1977, 1978), a number of tests have used these expressions to detect departures from the neutral equilibrium model. As with tests based on the site-frequency spectrum, significant departures can occur for neutral alleles if the population is not in equilibrium or if population structure is present.

Two other strategies use haplotype information. The first searches for the distinct signatures in the pattern of pairwise linkage disequilibrium (LD) predicted around a hard or a soft sweep (Table 8.2). The second considers the frequency of a neutral allele as a function of its age (Equation 2.12). Under neutrality, a common allele is an old allele, with shorter blocks of LD, reflecting a longer history of recombination. The presence of high-frequency alleles with **long haplotypes** (large blocks of LD) offers a signature of selection (these are often called **LRH**, for **long-range haplotype**, tests). A key point is that haplotype structure provides ***signals that can be missed by site-frequency and hard-sweep tests***, and thus offers more power in some settings.

Attempts to Account for Departures From the Equilibrium Model

Most tests for selection are based on the null hypothesis of the neutral equilibrium (or standard neutral) model (Chapter 2). While rejection of this null can indeed imply a signature of selection, rejection can also occur if a neutral population is not in mutation-drift equilibrium. Cavalli-Sforza (1966) noted that demography and population structure should leave a common signal over all genes within a genome, and this observation has been used in attempts to correct for any genome-wide nonequilibrium features in the data. The simplest approach is the **outlier method**, whereby values of the test statistic are computed for a large number of genes, with outliers suggesting potential targets of selection. This is an *enrichment method*, not a formal test. The second approach is to use data from presumably neutral markers unlinked to a region of interest to infer the population history (e.g., bottlenecks, expansions, population structure). These histories can then be used to simulate the coalescent structure (Chapter 2) for neutral alleles under this nonequilibrium model, which in turn can be used to generate the distribution of the test statistic under this more appropriate null. A final approach is to use presumably neutral sites to generate an **empirical site-frequency spectrum** to use in place of the equilibrium Watterson distribution.

These approaches are based on information from a large number of loci obtained in a genomic scan, with the assumption that most sites are not under positive selection and hence provide information to better shape the null hypothesis. This critically relies on the validity of Cavalli-Sforza's assumption of a common demographic or population structure signal over all loci, upon which any additional signal from selection is placed. Unfortunately, this need not be the case, especially in a population that is expanding over space. **Allelic surfing** can occur, wherein random alleles (and new mutations) on the leading edge of a wave of population expansion can "surf " (this wave) rather quickly to high frequencies in newly founded parts of the population (Edmonds et al. 2004; Klopfstein et al. 2006; Hallatschek et al. 2007; Travis et al. 2007; Excoffier and Ray 2008; Hallatschek and Nelson 2008, 2009; Excoffier et al. 2009a; Hallatschek 2011). Because neutral alleles on the leading wave of expansion are largely random, surfing ***does not affect all genomic locations equally***, and as a result can mimic signatures of selection even after correcting for demography or structure based on others markers within the sample. This is especially troublesome as the model species most surveyed for recent selection—humans, cosmopolitan human commensal *Drosophila* (*melanogaster* and *simulans*), and *Arabidopsis*—all have undergone massive range expansions. Hofer et al. (2009) found that while a large fraction of the human single-nucleotide polymorphisms (SNPs), short tandem repeats (STRs), and indels show large (greater than 0.3) differences in frequency across world populations, this pattern is easily accounted by allelic surfing, suggesting that this phenomenon can be a considerable problem in the search for sites under recent selection in humans.

What is the correct null model?

- Historically, this has been the equilibrium-neutral model
- However, in many respects, background selection (BGS) is more biologically motivated
- Hence, have to adjust for the gene density to recombination ratio

Allele-frequency change over time

- Would seem to be the most logical test
- Power issues
  - Selection time scale is ~ 1/s
  - Drift time sale is ~ $1/(2N_e)$
  - Hence, need tn >> 1/s

In the early literature, a number of workers tested for excessive divergence by simply querying whether allele frequencies in two samples were significantly different. As noted by Gibson et al. (1979) and Waples (1989b), this is inappropriate, as it does not account for the evolutionary *drift variance* in allele frequencies

$$\sigma^2(p_t) = p_0(1-p_0)\left[1-\left(1-\frac{1}{2N_e}\right)^t\right] \simeq p_0(1-p_0)\frac{t}{2N_e} \quad \text{for } t \ll N_e$$

where p_0 is the initial allele frequency (Equation 2.14a). Consider a population sampled at two time points (0 and t), with sample sizes of n_0 and n_t, respectively. The estimated divergence is

$$\widehat{\delta}_t = \widehat{p}_t - \widehat{p}_0 \tag{9.1a}$$

This divergence has an expected value of zero (as $E[\widehat{p}_t] = p_0$), with a variance of

$$\sigma^2\left(\widehat{\delta}_t\right) = \sigma^2\left(\widehat{p}_t - \widehat{p}_0\right) = \sigma^2\left(\widehat{p}_t\right) + \sigma^2\left(\widehat{p}_0\right) - 2\sigma\left(\widehat{p}_t, \widehat{p}_0\right) \tag{9.1b}$$

where $\widehat{p}_i = p_i + e_i$, the true value plus an error due to finite sampling. Because these are draws from a binomial distribution, the sampling variance of the initial frequency is

$$\sigma^2\left(\widehat{p}_0\right) = \frac{p_0(1-p_0)}{2n_0} \tag{9.2a}$$

while the final allele frequency is influenced by both the drift and sampling variances (Waples 1989a, 1989b)

$$\sigma^2(\widehat{p}_t) = p_0(1-p_0)\left[1-\left(1-\frac{1}{2n_t}\right)\left(1-\frac{1}{2N_e}\right)^t\right]$$
$$\simeq p_0(1-p_0)\left[\frac{1}{2n_t}+\frac{t}{2N_e}\left(1-\frac{1}{2n_t}\right)\right] \quad (9.2b)$$

If sampling is done (without replacement) before reproduction, then $\sigma(\widehat{p}_t, \widehat{p}_0) = 0$, and substitution of Equations 9.2a and 9.2b into Equation 9.1b yields

$$\sigma^2\left(\widehat{\delta}_t\right) \simeq p_0(1-p_0)\left[\frac{1}{2n_0}+\frac{1}{2n_t}+\frac{t}{2N_e}\left(1-\frac{1}{2n_t}\right)\right] \quad (9.2c)$$

If sampling is done either after reproduction or with replacement, this generates a covariance between the sample estimators $\widehat{p}_t$ and $\widehat{p}_0$; see Nei and Tajima (1981b) and especially Waples (1989a, 1989b) for details. Assuming this is not the case, so that $\sigma(\widehat{p}_0, \widehat{p}_t) = 0$, Equation 9.2c yields the correct variance for the null hypothesis of random genetic drift, giving the test statistic as

$$\frac{\widehat{\delta}_t^2}{\sigma^2\left(\widehat{\delta}_t\right)} \quad (9.2d)$$

The application of this test requires an accurate estimate of p_0 to compute the sample variance (Equation 9.2c). While the sample estimate, $\widehat{p}_0$, can be used, this can be improved upon by noting that the expected allele frequency change is zero, meaning that $\widehat{p}_t$ also contributes information about p_0. A simple average of the two frequencies is not appropriate, as $\widehat{p}_0$ has a smaller drift variance and the two estimates may differ in informational value due to differences in sample size, n_i. Given these concerns, Schaffer et al. (1977) and Waples (1989b) proposed a generalized (i.e., weighted) least-squares (GLS) estimator (LW Chapter 8) for p_0. Let $\mathbf{p} = (\widehat{p}_0, \widehat{p}_t)^T$ denote the two sample estimates and denote its sampling variance-covariance matrix by

$$\mathbf{V} = \begin{pmatrix} \sigma^2(\widehat{p}_0) & \sigma(\widehat{p}_0, \widehat{p}_t) \\ \sigma(\widehat{p}_0, \widehat{p}_t) & \sigma^2(\widehat{p}_t) \end{pmatrix} \tag{9.3}$$

Finally, let $\mathbf{1} = (1, 1)^T$ be a vector of ones. The underlying statistical model is $p_i = p_0 + e_i$, which can be written in general linear model form as $\mathbf{p} = p_0\mathbf{X} + \mathbf{e}$, where $\mathbf{V}$ is the covariance matrix for the vector, $\mathbf{e}$, of residuals and $\mathbf{X} = \mathbf{1}$. Recalling LW Equation 8.34 for GLS regression, the resulting estimate of p_0 is given by $\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{p}$, which reduces to

$$\text{GLS}(p_0) = \frac{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{p}}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}} \tag{9.4}$$

because both quadratic products are scalars.

Example 9.1. One of the classic papers in evolutionary biology is Fisher and Ford's (1947) study of the *medionigra* gene in the scarlet tiger moth *Panaxia dominula*, a colorful day-flying species with one generation per year. A single diallelic locus has a major effect on the forewing pattern. Individuals that are homozygous for the *dominula* allele have multiple forewing spots, while individuals that are homozygous for the *medionigra* allele have a darkly suffused forewing with, typically, two small spots (the *bimacula* phenotype). Heterozygotes show an intermediate pattern, which is called the *medionigra* phenotype. In 1938, Ford began studying a small colony of this species in Cothill Fen, just southwest of Oxford, England. Starting in 1941, capture-recapture data were used to estimate the census population size, with the smallest estimated size between 1941 and 1947 being 1000. In 1939 ($t = 0$) the frequency of the *medionigra* allele was estimated (from a sample size of $n_0 = 223$) as $\widehat{p}_0 = 0.092$, while by 1947 ($t = 8$), its sample frequency had decreased to $\widehat{p}_8 = 0.037$ ($n_8 = 1341$). Taking $N_e = 1000$ (this being the smallest estimated census value over any of the generations, and hence most favorable to supporting drift), do these data show evidence of a departure from drift?

dominula *medionigra* *bimacula*

For simplicity, assume sampling without replacement, so that $\sigma\left(\widehat{p}_0, \widehat{p}_t\right) = 0$, with the variances are given by Equations 9.2a and 9.2b. The resulting covariance matrix, $\mathbf{V}$, becomes

$$\frac{\mathbf{V}}{p_0(1-p_0)} = \begin{pmatrix} \frac{1}{2 \cdot 223} & 0 \\ 0 & \frac{1}{2 \cdot 1341} + \frac{8}{2000}\left[1 - \frac{1}{2 \cdot 1341}\right] \end{pmatrix} = \begin{pmatrix} 0.0022 & 0 \\ 0 & 0.0044 \end{pmatrix}$$

Because $\mathbf{V}^{-1}$ appears in both the numerator and the denominator of Equation 9.4, the unknown constant, $p_0(1-p_0)$, cancels out, allowing us to simply use the above right-hand matrix for $\mathbf{V}$, yielding

$$\text{GLS}\left(p_0\right) = \frac{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{p}}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}} = \frac{49.496}{674.762} = 0.0734$$

Equation 9.2c yields the sampling variance for the difference in allele frequencies as

$$\begin{aligned} \sigma^2\left(\widehat{\delta}_t\right) &\simeq p_0(1-p_0)\left[\frac{1}{2n_0} + \frac{1}{2n_t} + \frac{t}{2N_e}\left(1 - \frac{1}{2n_t}\right)\right] \\ &= 0.0734 \cdot 0.9266\left[\frac{1}{446} + \frac{1}{2682} + \frac{8}{2000}\left(1 - \frac{1}{2682}\right)\right] = 0.0004495 \end{aligned}$$

The resulting Waples test statistic for fit to pure drift becomes

$$\frac{(0.037 - 0.092)^2}{0.0004495} = 6.729$$

The probability that a χ_1^2 random variable is this big or larger is 0.0095, implying strong rejection of neutrality. By using different values of N_e in the above calculation, we can find the largest effective population size that would still allow drift to account for these data. For $N_e = 500$, the test statistic becomes 4.19 (a p value of 0.040), while for $N_e = 250$, the statistic is 2.39 (a p value of 0.12). Hence, any effective population size slightly smaller than 500 would be compatible with a hypothesis of the observed allele-frequency change being driven by drift.

Time series of frequencies: Fisher-Ford test

Let y_t denote the transformed frequency of the allele in generation t. For a t that is small relative to N_e, we find (approximately) that

$$y_t = 2\sin^{-1}\left(\sqrt{p_t}\right) \sim N\left(y_0, t/[2N_e]\right) \tag{9.5a}$$

where $y_0 = 2\sin^{-1}\left(\sqrt{p_0}\right)$ is the transformed value of the initial frequency. Estimates of allele frequencies are made at k time points, with no requirements about the temporal spacing between samples. Let $\mathbf{y}$ denote the vector of the transformed estimates of the k sampled allele-frequencies, and let $\mathbf{1}$ denote a vector of ones of the same length

$$\mathbf{y} = 2\begin{pmatrix} \sin^{-1}\left[\sqrt{p_1}\right] \\ \vdots \\ \sin^{-1}\left[\sqrt{p_k}\right] \end{pmatrix}, \qquad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \tag{9.5b}$$

Finally, we need the covariance matrix, $\mathbf{V}$, whose elements are independent of the allele frequency (because of the variance-stabilizing transformation; Equation 9.5a). The sample indices denote the sequence of samples, not the actual sampled generation itself (see Example 9.2), with t_i the generation number associated with the ith sample. The diagonal terms of $\mathbf{V}$ are given from Equation 9.2c

$$V_{ii} = \frac{1}{2n_{t_i}} + \frac{t_i}{2N_e}\left(1 - \frac{1}{2n_{t_i}}\right) \simeq \frac{1}{2n_{t_i}} + \frac{t_i}{2N_e} \tag{9.5c}$$

Now consider the covariance between samples i and j, which correspond to generations t_i and t_j, respectively (where $i > j$ and $t_i > t_j$). The estimates for these two sample points have a shared history (from the base value, p_0) of drift up through generation t_j, yielding

$$V_{i,j} = V_{j,i} = \frac{t_j}{2N_e} \quad \text{where } \; t_j < t_i \tag{9.5d}$$

Note that the covariance with the base generation ($t = 0$) is always zero (which is why the off-diagonal covariances for $\mathbf{V}$ in Example 9.1 were set to zero). The $k \times k$ matrix, $\mathbf{V}$, contains only those rows and columns corresponding to the k specific generations sampled.

This is now a goodness-of-fit problem for a linear model. Using Equation 9.4, we obtain a generalized least-squares (GLS) estimate of the (transformed) initial frequency

$$\widehat{y}_0 = \frac{\mathbf{1}^T \mathbf{V}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{V}^{-1} \mathbf{1}} \tag{9.6a}$$

Using this value, the vector of deviations is

$$\boldsymbol{\delta}_\mathbf{y} = \mathbf{y} - \widehat{y}_0 \cdot \mathbf{1} \tag{9.6b}$$

and the test statistic, the weighted sum of the squared (transformed) allele-frequency differences,

$$\boldsymbol{\delta}_\mathbf{y}^T \, \mathbf{V}^{-1} \boldsymbol{\delta}_\mathbf{y} \tag{9.6c}$$

is expected to be approximately χ^2_{k-1} distributed due to the normality assumption on y_i.

Year	t	$\widehat{p}$	$y = 2\sin^{-1}\left(\sqrt{p}\right)$	n
1939	0	0.092	0.616	223
1943	4	0.056	0.478	269
1947	8	0.037	0.387	1341

Assuming $N_e = 1000$, the resulting covariance matrix, $\mathbf{V}$ (on the transformed scale), becomes

$$\mathbf{V} = \begin{pmatrix} V_{0,0} & V_{0,4} & V_{0,8} \\ V_{4,0} & V_{4,4} & V_{4,8} \\ V_{8,0} & V_{8,4} & V_{8,8} \end{pmatrix} = \frac{1}{2000}\begin{pmatrix} \frac{2000}{2\cdot 223} + 0 & 0 & 0 \\ 0 & \frac{2000}{2\cdot 269} + 4 & 4 \\ 0 & 4 & \frac{2000}{2\cdot 1341} + 8 \end{pmatrix}$$

$$= \frac{1}{2000}\begin{pmatrix} 4.484 & 0 & 0 \\ 0 & 7.717 & 4 \\ 0 & 4 & 8.745 \end{pmatrix}$$

In addition,

$$\mathbf{y} = \begin{pmatrix} 0.616 \\ 0.478 \\ 0.387 \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \text{yielding} \quad \widehat{y}_0 = \frac{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{y}}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}} = \frac{418.851}{774.701} = 0.541$$

Using this estimate for y_0, the vector of deviations from the initial value becomes $\boldsymbol{\delta}_{\mathbf{y}} = \mathbf{y} - 0.541 \cdot \mathbf{1}$, returning a test statistic value of $\boldsymbol{\delta}_{\mathbf{y}}^{T} \mathbf{V}^{-1} \boldsymbol{\delta}_{\mathbf{y}} = 7.964$, which when compared to a χ_2^2 distribution, returns a significance level of 0.0186. For $N_e = 500$, Equation 9.6c returns a value of 5.398, for a significance of 0.067, so the hypothesis that drift alone accounts for the observed pattern of change cannot be rejected under this smaller value of N_e.

Schaffer's Linear Trend Test

A variation of the Fisher-Ford test was suggested by Schaffer et al. (1977), who noted that power might be improved by going beyond a simple lack of fit test against the model $y_t = \mu + e$ (where μ is the transformed initial allele frequency), by asking if a *significant linear trend* is present. The model now becomes

$$y_t = \mu + \beta t + e \tag{9.7a}$$

where a trend is indicated if β is significantly different from zero (the Fisher-Ford test assumes $\beta = 0$). Such a linear trend is not expected under drift but would be expected under directional selection, assuming that the direction of selection is not changing (migration from a population with a different allele frequency could also generate a linear trend). In general-linear-model form (LW Chapter 8), Equation 9.7a becomes $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where

$$\mathbf{X} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_k \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \beta \end{pmatrix}, \qquad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{9.7b}$$

where the elements of $\mathbf{V}$ are given by Equations 9.5c and 9.5d. For the data in Example 9.2, the resulting $\mathbf{X}$ matrix and the GLS estimate, $\widehat{\boldsymbol{\beta}}$, of the vector of parameters becomes

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \end{pmatrix}, \qquad \widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0.609 \\ -0.028 \end{pmatrix}$$

Applying LW Equation 8.35, the standard error on the slope is found to be 0.0086, showing that it is highly significant. Stuber et al. (1980) used this approach to infer selection at sites linked to several allozyme markers in a series of selected maize lines. One advantage of the

Between-population divergence

- As we have just seen, divergence in allele frequencies could be measured as the change in a population over time
- It would equally well be measured as the observed divergence between two populations separated form an ancestral population at some time in the past
  - For example, dairy vs. meat breeds of cattle
  - The machinery just used can be applied

DIVERGENCE BETWEEN POPULATIONS: TWO-POPULATION COMPARISONS

While most of the analysis of divergence data in structured populations is based on F_{ST} statistics (Chapter 2), we start with a few comments on the simple situation in which one is comparing a biallelic locus between two populations. As in the case of the divergence of a single population measured at starting and ending time points, divergence can be measured as the squared allele-frequency difference,

$$\widehat{\delta}_t = \left(\widehat{p}_{t,1} - \widehat{p}_{t,2}\right)^2 \tag{9.8}$$

namely, the squared difference between the frequency in the two populations at some sample time, t, following their isolation from a common ancestor in generation 0. Whether $\widehat{\delta}_t$ is too large, or too small, relative to drift can be evaluated using a simple modification of the Waples test, wherein the denominator in Equation 9.2d is replaced by $\sigma^2\left(\widehat{p}_{t,1}\right) + \sigma^2\left(\widehat{p}_{t,2}\right)$, the sum of the allele-frequency sampling variances for each population (defined as in Equation 9.2b). This expression requires estimates of the divergence time, t, as well as the average effective sizes for both populations. More generally, because $E\left[\widehat{p}_{t,i}\right] = p_0$, in theory one could sample the two populations at different time points (t_1 and t_2), but now using $\sigma^2\left(\widehat{p}_{t_1,1}\right) + \sigma^2\left(\widehat{p}_{t_2,2}\right)$ in the denominator of the test statistic.

Finally, a very simple statistic that often appears in comparisons of selected versus control populations is Grossman et al.'s (2010) ΔDAF statistic. This metric is a natural outgrowth of the type of comparisons shown in Figure 9.1, which focuses on the difference in the **derived allele frequency** (DAF) between a control and a selected population. For a candidate SNP, let $\overline{D}_{NS}$ denote the frequency of the derived allele in a nonselected control population (or its average frequency if multiple control populations are used) and its frequency, D_S, in the putatively selected population, with $\Delta DAF = D_S - \overline{D}_{NS}$. This statistic ranges between plus one and minus one, and standard outlier approaches are used to highlight SNPs with excessive values.

F_{ST} - based tests

- With more than two populations, F_{ST} provides a natural metric for divergence
- F_{ST}-based tests examine where the amount of between-population divergence is too large, or too small, relative to the pure drift hypothesis
  - Landscape genetics

Recall: F_{ST} is the fraction of genetic variation due to between-population differences

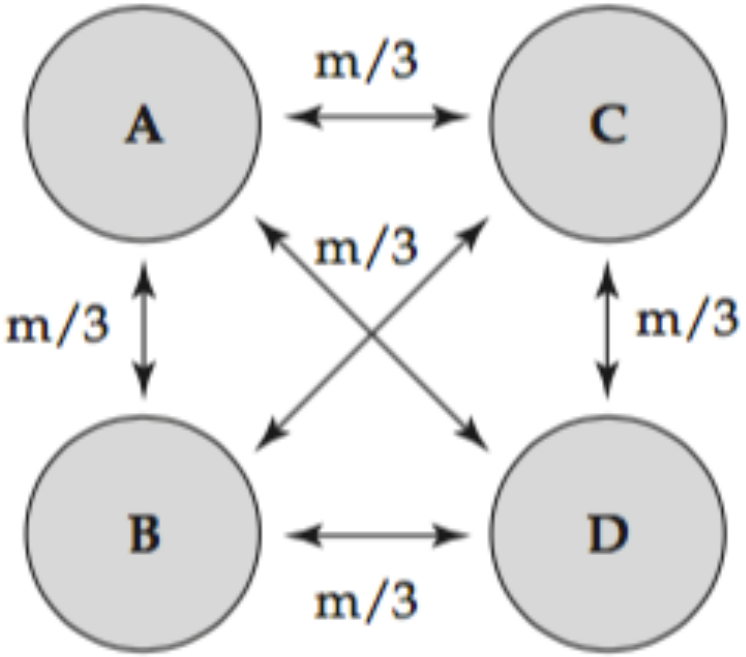
$$F_{ST} = \frac{\sigma^2(p)}{p_0(1-p_0)}$$

Under pure drift, this is roughly a linearly-increasing function of divergence time

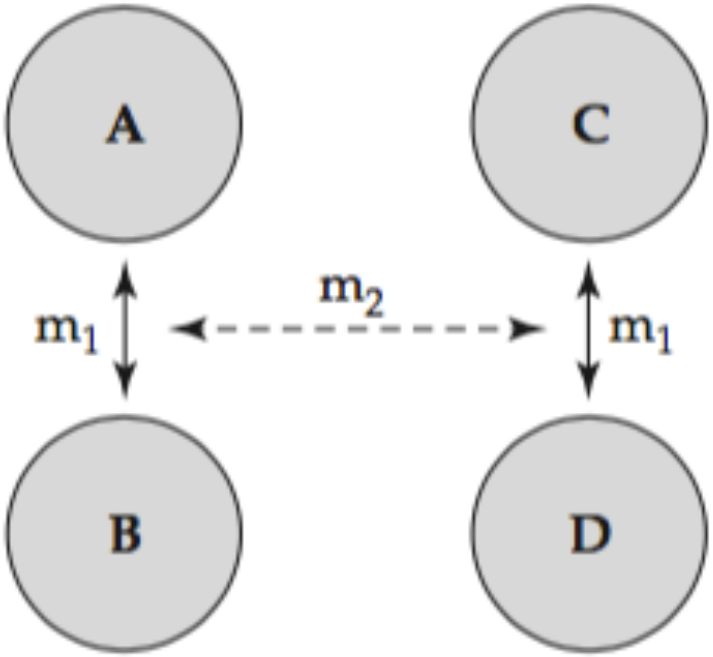
$$F_{ST} = \left[1 - \left(1 - \frac{1}{2N_e}\right)^t\right] \simeq \frac{t}{2N_e} \quad \text{for} \quad t \ll N_e$$

With migration and mutation, the equilibrium value of F_{ST} is very model-dependent

Island model



Hierarchical model



$$F_{ST} = \frac{1}{1 + 4N\dfrac{md^2}{(1-d)^2}} \simeq \frac{1}{1+4Nm} \quad \text{for} \quad d \gg 1$$

Example 9.4. The effectiveness of F_{ST} to detect selection was examined by Taylor et al. (1995), using a putative target of selection in the tobacco budworm (*Heliothis virescens*), a noctuid moth and a major cotton pest in the United States. Pyrethroid insecticides have been used in control efforts, and these act on voltage-gated sodium channels in the nervous system. The historical usage patterns of these insecticides, and hence the putative selection pressures on sodium channel genes, differed over the sampled populations examined by the authors. As a result, they predicted that F_{ST} values at the sodium channel *Hpy* gene should be significantly higher than for background loci, reflecting this differential selection over the sampled subpopulations. Samples of adults from widely spaced locations in the United States revealed an F_{ST} value of 0.041 ± 0.005 at the *Hpy* marker, in contrast to values of 0.002 ± 0.001 at 14 other loci, with the latter result indicating fairly weak population structure in this species.

Outlier Approaches

The underlying premise for most F_{ST}-based tests of selection was the suggestion by Cavalli-Sforza (1966) that all neutral loci should have the same expected value of F_{ST}, reflecting the genome-wide impact of common demographic and population-structure forces. Thus, one can (in theory) use a large number of marker loci to estimate the baseline F_{ST} value for the set of populations being compared, and then search for outlier loci. This approach is easily modified to look for specific loci being outliers in specific populations (e.g., Akey et al. 2002; Kayser et al. 2003; Akey et al. 2010). Loci with excessively high values indicate more divergence than expected under drift, and the possibility that the marker is linked to a site that is under differential selection over the demes. Likewise, excessively low values indicate less divergence than expected under drift, and hence the potential for a site that is under balancing selection near the marker. While the historical interpretation of F_{ST} data follows from these last two statements, results from Chapter 8 on sweeps under *uniform* selection in structured populations suggest that a more nuanced view is needed. Recall from Figure 8.8 that uniform selection over the entire metapopulation can generate excessive divergence (Figure 8.8A) during a hard sweep of a single allele when it is still restricted to a subset of the demes. Similarly, a soft sweep under uniform selection can also generate excessive divergence. Conversely, a completed hard sweep through the sampled demes generates a reduction in divergence relative to background levels of F_{ST} (Figure 8.8B).

The outlier strategy makes two assumptions: the vast majority of scored loci are neutral, and all neutral sites reflect the same underlying population demography. As discussed in the introduction to this chapter, new alleles arising on the leading wave of a population expansion can "surf" to high frequencies, generating excessive values over the expected background. Likewise, differences in the ratio of gene density to recombination rate in different parts of the genome change the expected pattern of background selection, potentially creating outliers even among neutral markers.

A final complication is that when the population structure departs from the island model (equal divergence is expected between all demes; Chapter 2), the variance in F_{ST} is inflated, generating an excess of outliers. An interesting example of this phenomenon appears in the work of Fourcade et al. (2013), who found that river fishes showed an unusually high number of outlier loci. While such an observation might be taken as evidence that river species have higher rates of local adaptation, simulations by these authors showed that species with a **fractal** (highly branching) **population structure** have a greatly inflated variance in F_{ST} relative to the island model. This arises because migration on fractal structures (such as rivers or valleys) generates a complex pattern of correlated allele frequencies. Other types of population structures, such as hierarchical island models (Figure 2.11), population expansions from refugia, and allelic surfing, can all inflate the number of outliers (Excoffier et al. 2009a; Bierne et al. 2013).

Tests Based on F_{ST}-generated Branch-lengths

When migration and new mutation can be ignored, F_{ST} provides an estimate of the divergence time, T (scaled in $2N_e$ generations), between two populations. Rearranging Equation 2.43, taking the log of both sides, and recalling that $\ln(1-x) \simeq -x$ (for $|x| \ll 1$) yields

$$\ln(1-F_{ST}) = t \ln\left(1 - \frac{1}{2N_e}\right) \simeq -t/2N_e \tag{9.9}$$

Hence $T = -\ln(1-F_{ST}) \simeq t/2N_e$, and one can recast an excessive F_{ST} value as an excessive separation time required for drift to account for the observed divergence. These estimated times are called **branch lengths** and (following the Cavalli-Sforza premise) should have the same expected value over all neutral genes. An excessive branch length for a candidate gene relative to some reference set of genes suggests excessive change relative to drift (Vitalis et al. 2001; Rockman et al. 2003), and is the basis of the **population branch statistics** (**PBS**) of Yi et al. (2010); see Figure 9.2.

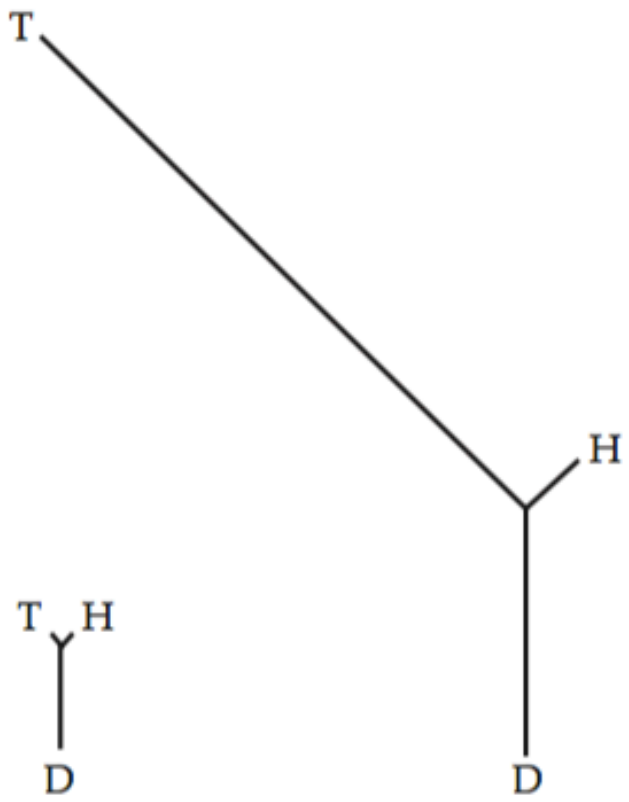
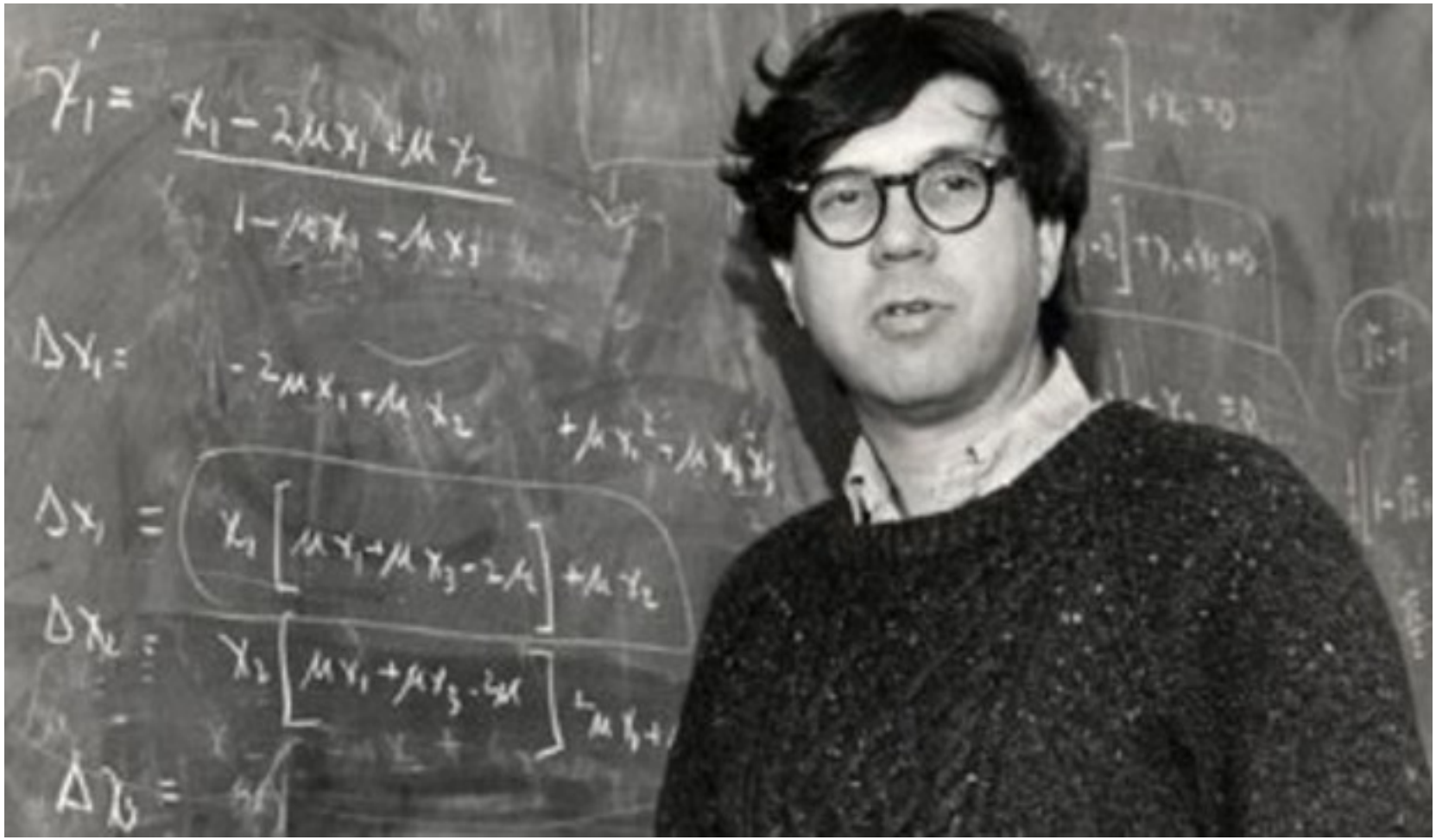


Figure 9.2 F_{ST}-based branch lengths for Tibetan (T), Han (H), and Danish (D) populations. (**Left**) Lengths based on the average F_{ST} values for all sampled markers. (**Right**) The tree for the *EPAS1* gene. While the D and H branches show increased divergence relative to the average F_{ST}, the divergence along the T lineage is far more dramatic. This is consistent with excessive allelic divergence due to selection for living at high altitude (or perhaps other features, such as allelic surfing). (After Yi et al. 2010.)



Dick Lewontin

Lewontin-Krakauer tests

- Lewontin and Krakauer (1973) showed that F_{ST} is roughly chi-squared distributed under neutrality and the island model
- Third-generation versions of this test allow for rather arbitrary covariance structures among subpopulations and can be rather powerful

The Lewontin-Krakauer Test: Basics

The above outlier methods (for either F_{ST} or branch lengths) are rather *ad hoc*, and best viewed as *enrichment* methods, distilling down a reduced set of markers that is likely enriched for selected sites. The critical missing element in these methods is the expected *distribution* of F_{ST} values for a random marker, allowing p values to be placed on outliers. Formal distribution-based tests were introduced by Lewontin and Krakauer (1973), who considered the distribution of F_{ST} values for a random biallelic locus sampled over n populations under an island model (Figure 2.11). If we assume that the distribution (over populations) of the frequency of an allele is roughly normal, the expected large-sample distribution of F_{ST} values approximately follows a $\lambda\chi^2_{n-1}$ distribution, with a scaling factor of $\lambda = E(F_{ST})/(n-1)$. Given Cavalli-Sforza's assumption that, on average, population structure influences all neutral loci equally, Lewontin and Krakauer estimated $E(F_{ST})$ from the average $\overline{F}_{ST}$ over all scored loci, giving the distribution for a random realization F_{ST} as

$$\frac{1}{\lambda}F_{ST} = \frac{(n-1)F_{ST}}{\overline{F}_{ST}} \sim \chi^2_{n-1} \tag{9.10a}$$

There are a number of additional potential problems with this approach of using $\overline{F}_{ST}$ to provide an estimator of λ. First, this estimate can be biased by skew resulting from a few excessive F_{ST} values. Specifically, if $F \sim \lambda\chi^2_{n-1}$, estimating λ by comparing means yields the method-of-moments estimator, $\widehat{\lambda} = \overline{F}/(n-1)$, as $E[\chi^2_n] = n$. However, even just a few loci that are under selection—and hence with extreme large values of F_{ST}—inflate $\overline{F}$ and bias the estimate of λ under the null. A more robust approach is to replace the usage of the means with **medians**, the 50% values of the two distributions (Devlin and Roeder 1999). Specifically, $\text{med}(F) = \text{med}(\lambda\chi^2_{n-1})$, or

$$\widehat{\lambda} = \frac{\text{med}(F)}{\text{med}(\chi^2_{n-1})} \tag{9.10c}$$

For example, suppose the median for single-locus F_{ST} values among a collection of loci sampled over five populations is 0.127. Because $\Pr(\chi^2_4 \leq 3.357) = 0.5$, the median value of a χ^2_4 is 3.357, yielding

$$\widehat{\lambda} = \frac{\text{med}(F)}{\text{med}(\chi^2_{n-1})} = \frac{0.127}{3.357} = 0.038$$

as a more robust estimate of λ under the null (drift) relative to that based on the mean, $\overline{F}_{ST}$, because the median-based estimate is not biased by the presence of a modest number of loci under selection.

A third, and deeper, problem is the implicit assumption of Lewontin and Krakauer that neutral allele frequencies are *independent among demes*. This is correct under the standard island model (Figure 2.11), which yields equal expected divergence among any pair of demes, and the same amount of variation within any deme (assuming no among-deme differences in N_e). However, this assumption fails under more complex population structures, such as unequal migration between demes (e.g., the **isolation by distance** model, wherein closer demes exchange migrants at higher rates) or hierarchical structure among demes generated by their founding. These population-structure issues create correlations among allele frequencies from different demes, inflating the variance of F_{ST} relative to the expectations under the island model, which impacts the χ^2 assumption (Nei and Maruyama 1975; Robertson 1975a, 1975b; Tsakas and Krimbas 1976).

As a result of these concerns (and others; see Nicholas and Robertson 1976), the original version of the Lewontin-Krakauer test quickly languished. However, its basic simplicity, coupled with its requirement of only the type of data routinely gathered by ecological geneticists (estimates of locus-specific F_{ST} values), fueled the search for ways to correct these initial flaws.

Whitlock and Lotterhos (2015) recently suggested a potentially simple work-around for many of these issues, going by the name of ***OutFLANK***. They noted through extensive simulations of very different population structures that the distribution for F_{ST} values (provided heterozygosity levels were not too small) was very close to χ^2, but with different degrees of freedom from the Lewontin-Krakauer value of $(n-1)$. This difference in the degrees of freedom makes sense, given a lack of independence among demes, and they recommended a two-step approach for obtaining approximate p values. First, the upper and lower 5% of the empirical F_{ST} values are trimmed. The logic being that loci under uniform selection (generating excessive low values) and under divergent selection (generating expressive high values) are expected to be only a tiny fraction of all tested sites. The remaining trimmed distribution, representing the core 90% of the values, is then used in a ML setting to estimate the appropriate degrees of freedom for such a doubly truncated χ^2. (More generally, Table A2.1 shows that the χ^2 distribution is a special case of the gamma distribution, and fitting the latter allows for what amounts to fractional degrees of freedom, which might further improve the fit.) With the corresponding null density now estimated, appropriate p values for outliers can be obtained. Their simulations showed that this approach worked well for excessively high values (i.e., the right-hand tail of the distribution), but very poorly for the left-hand tail (those loci showing small F_{ST} values than expected).

Second-generation versions

- The next wave of LK-type tests were model-based, moving away from the island model
- *Bayes* F_{ST} (Beaumont and Balding 2004)
  - All demes drawn from a common ancestor, but with no further migration
  - Foll and Gaggiotti's (2008) *Bayescan* method

Third-generation tests

- Use marker data to estimate a general covariance structure (or phylogeny) among dens, use this to adjust tests
  - *FLK* test (Bonhomme et al. 2010)
  - *hapFLK* test (Fariello et al. 2013)
  - *Bayenv/Bayenv2* (Coop et al. 2010; Gunther and Coop 2013)
  - *PCAdapt* (Duforet-Frebourg et al. 2014)

The F_{ST} extension (*FLK*) of Bonhomme et al. (2010) uses a set of neutral loci together with an outgroup to construct a kinship matrix, $\mathcal{F}$, of populations, based on branch lengths of the estimated phylogenetic tree among the sampled populations. The assumption is that some pattern of evolution (described by $\mathcal{F}$) unfolds from an ancestral population with an allele frequency of p_0, but with no further migration between subpopulations. For n populations, the *FLK* test statistic is given by

$$T_{FLK} = \frac{(\mathbf{p} - \widehat{p}_0\mathbf{1})^T\mathcal{F}^{-1}(\mathbf{p} - \widehat{p}_0\mathbf{1})}{\widehat{p}_0(1-\widehat{p}_0)}, \quad \text{with} \quad \widehat{p}_0 = \frac{\mathbf{1}^T\mathcal{F}^{-1}\mathbf{p}}{\mathbf{1}^T\mathcal{F}^{-1}\mathbf{1}} \tag{9.12}$$

where $\mathbf{p}$ is a vector of the allele frequencies for one particular locus over the n sampled demes and $\mathbf{1}$ is a column vector of n ones. Bonhomme et al. showed that T_{FLK} follows a χ^2 distribution under the null model of no selection, provided allele frequencies are not too extreme, with outliers deemed to be candidates for loci under selection. Note that $\widehat{p}_0$ is of the same form as the GLS estimators for the initial frequency (Equations 9.4 and 9.6a), and that T_{FLK} has the same general form as the test statistic for the Fisher-Ford test for excessive allele-frequency change (Equation 9.6c).

ALLELE-FREQUENCY CORRELATIONS WITH ENVIRONMENTAL VARIABLES

A final approach for comparing allele frequencies over a set of populations was introduced in Chapter 8, namely to search for correlations between allele frequencies and environmental factors. This approach is often referred to as **environmental association analysis** (**EAA**) or **genetic-environmental analysis** (**GEA**), although our preference is for the former to avoid confusion of the latter with the analysis of genotype × environment interactions. In such studies, typically, a large number of potential factors are initially considered, and then the method of principal components (Appendix 5) is used to extract a smaller set of environmental features. If polygenic adaptation is the norm, classic hard-sweep (Table 8.2) or even soft-sweep signals will be unlikely, as the response is driven by modest allele-frequency changes over a number of small-effect loci. Hancock et al. (2010a, 2010b) suggested that such polygenic sweeps might be detected through subtle allele-frequency shifts that are concordant in populations experiencing similar environments but in different geographic regions.

Joost's Spatial Analysis Method (*SAM*)

The extension of testing for an association between a specified candidate gene and a single environmental factor to a more general genome scan over a set of environmental features starts with Joost et al. (2007). Their **spatial analysis method** (*SAM*) computes separate logistic regressions for each allele-environment combination. As discussed in Chapter 14, logistic regressions are commonly used to model how the probability of an event varies with some other variable, in this case predicting allele frequency as function of the environmental value. As with second-generation LK tests, *SAM* has a critical limitation in assuming that neutral alleles from different populations are uncorrelated. Failing to account for the natural correlation in neutral allele frequencies shaped by shared migrations and/or history will yield incorrect sampling errors. Further, populations in geographic proximity are expected to have both correlated allele frequencies (due to migration) *and* correlated environmental values, generating many false positives. While Poncet et al. (2010) extended *SAM* by allowing for small-scale correlations in allele frequencies within spatially proximate demes, their approach does not adjust for larger-scale correlations.

Accounting for Population Structure: Coop's *Bayenv* and Frichot's *LFMM*

Coop et al. (2010; Eckert et al. 2010; Günther and Coop 2013; also see Gautier 2015) attacked the problem of adjusting for unknown population structure by using marker data to estimate the expected correlation pattern among neutral alleles for the sampled populations. This is akin to the kinship matrix approach used by Bonhomme et al. (2010) to adjust for correlations among allele-frequency values from different demes. Example 9.5 sketches the basic structure of their ***Bayenv*** approach, which uses Bayes factors (Appendix 2) to gauge the support for an allele-environmental correlation after the effects of population structure have been removed. Formally, however, this is still an outlier method, as it generates an empirical distribution of Bayes factors for each SNP and uses this to assess the strength of association for a given locus. An alternative implementation to adjust for population structure, which is very closely related to Coop's method (as well as to Duforet-Frebourg et al.'s previously mentioned *PCAdapt* approach), is the **latent factor mixed model** (***LFMM***) approach of Frichot et al. (2013), which is also outlined in Example 9.5.

Coop's base model (Equation 9.13c) is extended to account for environmental factors that influence the allele frequencies as follows. Consider a vector, $\boldsymbol{\beta}$, of potential regression coefficients for the impact of environmental factors on allele frequencies, and a matrix, $\mathbf{X}$, whose values in row i correspond to the environmental parameters measured for the ith population (this is simply a GLS linear model; see LW Chapter 8). The null mean p_0 for an allele (Equation 9.13c) is augmented by the environmental effect to give

$$\boldsymbol{\Theta} \sim \text{MVN}_n \left[p_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, p_0(1-p_0)\boldsymbol{\Omega} \right] \tag{9.13d}$$

where $\mathbf{1}_n$ is an n-dimensional vectors of ones. This model assumes that any relationships between allele frequencies and the environmental variables have some linear component. The addition of the vector $\mathbf{X}\boldsymbol{\beta}$ to account for environmental effects is an example of a **factorial regression** (e.g., Baril et al. 1992), which is discussed at length in Volume 3 in the context of analyzing genotype-by-environment interactions.

$$\frac{(\mathbf{p} - \hat{p}_0\mathbf{1} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{\Omega}^{-1} (\mathbf{p} - \hat{p}_0\mathbf{1} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\hat{p}_0(1-\hat{p}_0)} \tag{9.13e}$$

The *Bayenv* method of Coop et al. is a two-step approach: (i) $\boldsymbol{\Omega}$ is estimated from a presumed set of neutral markers, and (ii) the model is run with this matrix (or in a Bayesian framework, with draws of this matrix to generate a posterior accounting for the uncertainty in its estimation; Appendix 2).

Latent factor mixed model (*LFMM*, Frichot et al. 2013).

Tests for a pattern of reduced variation

- Visual scans
- Bottleneck ML models
  - Uses maximum-likelihood to test whether the data are better fit with a double-bottleneck model (see Chapter 9 for details)
- Formal test using the spatial pattern of variation
  - *CLRT-GOFT*
  - *Sweepfinder* (uses empirical SFS)
  - *XP-CLR* (divergence between a selected an unselected population)

Simple Visual Scans for Changes in Nucleotide and STR Diversity

The most basic approach is a simple plot of variation as a function of genomic location, looking for either peaks (long-term balancing selection) or valleys (a recent sweep); see, for example, Figures 8.1 and 8.2. With SNP data, variation is typically scored as average nucleotide diversity, π (Chapter 4), within a sliding window to smooth out the inherent noisiness from individual sites. With simple sequence repeats or microsatellite markers (also known as simple tandem repeats, or STRs, and simple sequence repeats, or SSRs), several different metrics of variation are available, such as copy-number variance, number of alleles, and probability of heterozygosity. With their large number of alleles per marker and high mutation rates, STRs provide a more consistent signal and are usually plotted on a per-marker basis (as opposed to a sliding-window analysis); see Figure 9.3 and Example 9.6. A point of caution is that mutation rates at STRs can be length dependent, with smaller arrays often expected to show less variation.

Large
Munsterlander

Dachshund

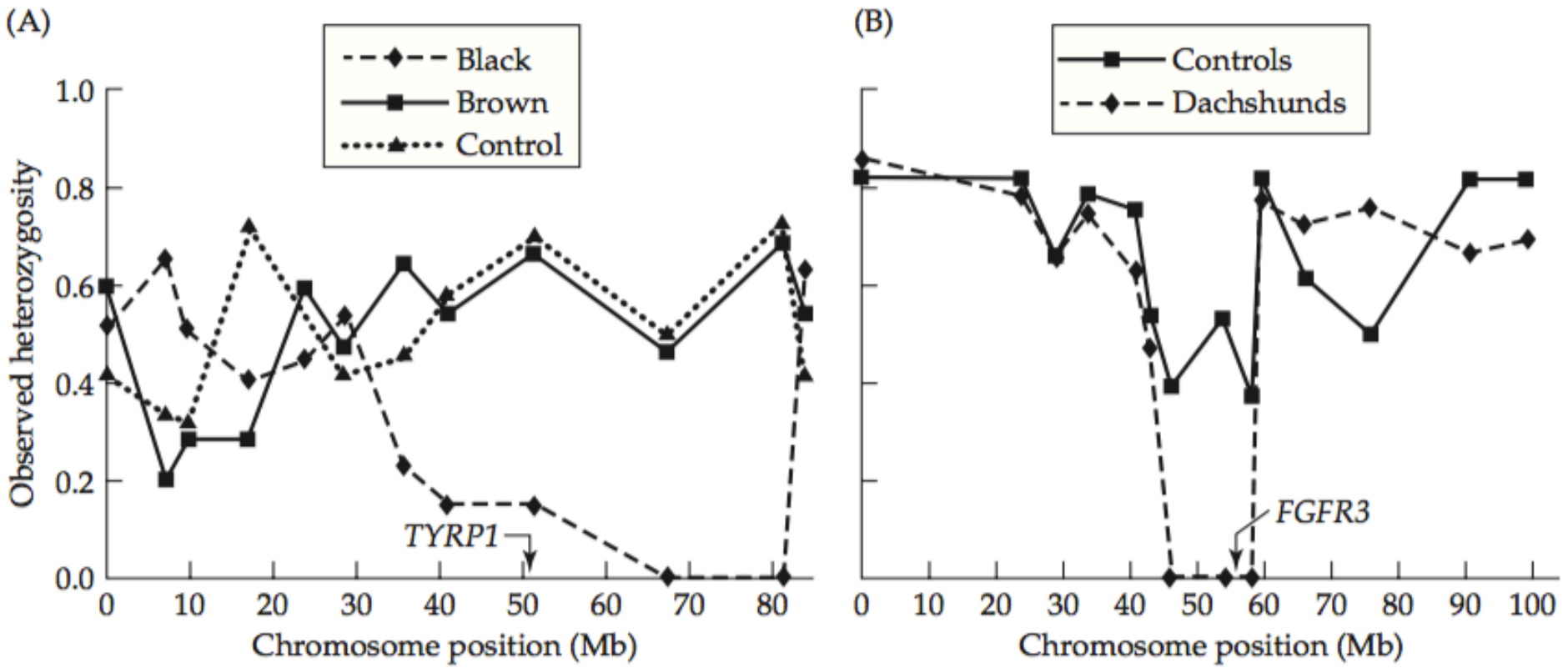


Figure 9.3 Using microsatellites in the search for dog domestication genes. (**A**) Large Munsterlanders have a black coat, suggesting the pigment gene *TYRP1* on chromosome 11 may be a possible domestication gene. A plot of variation for this breed (black) relative to both control (neither black or brown) and brown individuals shows depressed variation spanning this gene. (**B**) Dachshunds are characterized by shortened limbs, suggesting the *FGFR3* gene on chromosome 3 as a candidate. Dachshunds have an absence of variation at three microsatellites spanning this gene, while variation is present in controls (normal-limbed breeds). (After Pollinger et al. 2005.)

Likelihood-based tests

- Recall that a sweep generates a particular pattern around a selected site, where the diversity increases as one move away from the location

$$f_{s,i} = (4N_e s)^{-c_i/(2hs)} = e^{-c_i \lambda}$$

- A number of likelihood-based tests examine where the spatial pattern of diversity in a region fits this pattern. If so, it allows one to estimate s

Wolfgang Stephan

CLRT-GOF

- Stephan and Kim proposed the composite -likelihood ratio test (CLRT)
- Jensen et al. (2005) found that the *CLR* test is *not* robust to population structure or recent bottlenecks.
- To distinguish sweeps from false signals generated by demography and population structure, Jensen et al. proposed that any significant *CLR* result be subjected to an additional goodness-of-fit (*GOF*) test to see how well it fits a sweep model

Using spatial information (pattern of diversity along a chromosome) to detect sweeps

Likelihood of seeing k_i/n derived alleles at a site

$$\Pr(k_i \mid n, \boldsymbol{\Theta}) = \binom{n}{k_i} \int_{1/(2N)}^{1-1/(2N)} x^{k_i}(1-x)^{n-k_i}\phi_i(x \mid \boldsymbol{\Theta})\, dx$$

$$\phi(x) = \begin{cases} \theta\left(\frac{1}{x} - \frac{1}{f}\right), & \frac{1}{2N} \le x < f \\ 0, & f \le x \le 1-f \\ \frac{\theta}{f}, & 1-f < x \le 1 - \frac{1}{2N} \end{cases}$$

$$f_i = (4N_e s)^{-c_i/(2hs)} = e^{-c_i \zeta}$$

Key: varies in a defined way (i.e., with c) around the sweep

Sweepfinder

- The CLRT starts by assume the Watterson distribution, as modified by a sweep.
- Neilsen set al (2005) modified this approach to use the empirical SFS (need to adjust for BGS)
- They called this approach *Sweepfinder*

XP-CLRT

- Chen et al also proposed using the spatial pattern in a chromosome region, but they examined the expected allele frequency difference between two populations (one selected, the other not) descending from a common population
- This is their cross-population (XP) CLRT
- Often used in the search for domestication genes

Ascertainment Issues

Because many of these likelihood models exploiting genomic positional information are computationally demanding, they are typically employed *following* a general scan of a genome for some signature of selection, such as regions of depressed variation, or showing unusual site-frequency spectra (such as those with a negative Tajima's D or positive Fay and Wu's H values, which are discussed in the next section). Choosing the region or regions in which to perform the likelihood tests based on the appearance of these special features creates a strong ascertainment bias that dramatically shifts the null distribution. (Note that this is different from SNP ascertainment bias arising from the nonrandom choice of SNPs at the start of the analysis.) The coalescent process can be noisy, and regions with unusual underlying genealogies (such as strong compression of the nodes) can occur by chance even under the equilibrium neutral model. This is especially true when a large number of sites are sampled, presenting more draws from the same underlying process, some of which will be realizations that are extreme values.