12

THE NEUTRAL DIVERGENCE OF QUANTITATIVE TRAITS

From the perspective of the evolutionary biologist, missing or incomplete null or neutral models for many omics data (for example, transcriptome and metabolome data) limit our understanding of how selection has shaped their evolution — Leinonen et al. 2013

Draft 16 May 2015

In the preceding chapter, we learned how the opposing forces of random genetic drift and mutation lead to an equilibrium level of within-population genetic variance. In contrast, the phenotypic variance among isolated populations may continue to increase nearly indefinitely for neutral characters, as isolated demes or species recurrently acquire and become fixed for independent mutations. Here, we explore neutral factors that can drive the evolutionary dynamics of the among-population variance. As in Chapter 11, we will start with the situation in which the time span is short enough that most of the change in population-mean phenotypes is driven by drift acting on existing variation rather than by new alleles introduced by mutation. We then explore the consequences of longer-term divergence, with mutation playing an increasingly dominant role, showing that eventually the rate of divergence for neutral characters may become essentially independent of local effective population sizes. We conclude by using this theory to develop statistical tests of whether an observed pattern of phenotypic divergence is consistent with model of strict neutral drift and mutation.

Although few quantitative traits may actually evolve in a purely neutral fashion, a more compelling case for selection can always be made if the hypothesis of neutrality can be formally rejected. For example, an observed divergence of isolated lines that is significantly less than the neutral expectation provides evidence of stabilizing selection, whereas the reverse supports a role for diversifying selection. In addition, as populations become diminishingly small in size, drift begins to overwhelm selection, promoting nearly neutral patterns of evolution. Tests for departures from neutral divergence come in several different forms. First, one could compare the observed rate of divergence with that expected given estimates of the effective population size, time, and genetic variation. Second, one may have a time series of data (such as a fossil record sequence) and test whether the observed pattern is more consistent with a neutral random walk, a biased random walk, or stasis. Third, one can compare the within- and between- population structure of the genetic variance of a candidate trait (measured by Q_{ST}) against a genome-wide estimate based on presumed neutral markers (F_{ST}). Finally, tests have been proposed based on the distribution of QTL effects detected by crossing two divergent lines, such as whether one line contains an excess of plus alleles. This highly diverse collection of neutral divergence tests have been applied to an equally diverse collection of quantitative traits, ranging from morphological changes in the fossil record to examining which evolutionary forces shape omics data. To highlight the latter, we conclude the chapter by applying several of these approaches to

examine if (and if so, how) divergence in gene expression departs from neutrality. Chapters 8-10 considered the complementary topic of tests for departures from neutrality at specific *loci*, as opposed to specific *traits*, which is our focus here.

SHORT-TERM DIVERGENCE

We start with the special case in which all gene action is additive and random genetic drift is the only evolutionary force. Most of the predictions of this model can be expressed in terms of two observable quantities: the additive genetic variance in the base population $\sigma_A^2(0)$, and the effective population sizes N_e of the isolated lineages. The expected among-population genetic variance, $\sigma_B^2(t)$, under neutrality is obtained by noting that the mean genotypic value at a diallelic locus i is $2a_ip_i$ (there being two genes per locus, each with additive effect a_i with probability p_i, and effect 0 with probability $1-p_i$). The variance among populations for this locus is (from the definition of the variance) the expected value of the square of additive effects minus the square of their expected values, or $E\{[2a_ip_i(t)]^2\} - \{E[2a_ip_i(t)]\}^2$. This simplifies to

$$4a_i^2E\{[\,p_i(t)]^2\} - 4a_i^2\{E[\,p_i(t)]\}^2 = 4a_i^2\left(E\{[p_i(t)]^2\} - \{E[\,p_i(t)]\}^2\right) = 4a_i^2\sigma_{p_i}^2(t),$$

where $\sigma_{p_i}^2(t)$ is the expected among-population variance in allele frequency. Summing over all loci, assuming negligible gametic-phase disequilibrium, and substituting from Equation 2.14a gives

$$\sigma_B^2(t) = 4\sum_{i=1}^{n} a_i^2p_i(0)[1-p_i(0)]\left\{\frac{1}{N_{fo}} + \left[1-\left(1-\frac{1}{2N_e}\right)^t\right]\right\} \tag{12.1a}$$

$$= \left(\frac{1}{N_{fo}} + 2f_t\right)\sigma_A^2(0) \tag{12.1b}$$

where N_{fo} is the effective number of founders per line, the inbreeding coefficient f_t follows from Equation 2.4c, and the time index is defined such that $t = 0$ denotes the final generation of the base population and $t = 1$ denotes the founding generation for the isolated lines. Equation 12.1 shows that, under the assumptions of this ideal model, the expected variance among genotypic means of isolated populations increases linearly with the inbreeding coefficient, asymptotically approaching a limit (as $f_t \to 1$) that is very close to twice the additive genetic variance in the base population (Wright 1951). Equation 12.1 describes how any initial variation is partitioned by drift during the random (and differential) fixation of these initial alleles in the diverging populations. Under the assumption of additivity, Equation 12.1b holds regardless of the number of alleles at the underlying loci.

Ignoring the generally minor contribution (N_{fo}^{-1}) from the baseline founder effect, this limiting result may be obtained in a simpler manner. Because the probability of fixation of a neutral allele is equal to its initial frequency, when the process of random drift is completed, a proportion $p_i(0)$ of the populations will have genotypic value $2a_i$, while the remaining proportion, $1-p_i(0)$, will have genotypic value 0. The mean genotypic value is therefore $2a_ip_i(0)$ and the mean squared value is $(2a_i)^2p_i(0)$, which yields the among-population variance $4a_i^2p_i(0)[1-p_i(0)] = 2\sigma_{A_i}^2(0)$.

The expression for $\sigma_B^2(t)$ give by Equation 12.1 only considers the true genetic divergence among lines (the **evolutionary variance**), which can in principle be obtained by an analysis of variance of phenotypic variation within and among lines. If, however, one simply focuses on the raw variance of the observed means, additional sources of variation,

associated with finite sample sizes, also contribute to the observed divergence (Hill 1972; Lynch 1988). For example, when the mean phenotype of each line is determined using n progeny from $N/2$ matings (involving $N/2$ males and females, for a total parental sample size of N), there can be three additional sources of variance to add to Equation 12.1:

i) The segregational variance $(1 - f_{t-1})\sigma_A^2(0)/(Nn)$ of the mean offspring value about the mean breeding value of their parents resulting from the sampling of $Nn/2$ individuals. This follows as the segregational variance (in the absence of linkage disequilibrium) equals half the total variance (Chapters 16 and 24);

ii) The sampling variance $\sigma_{E_m}^2/(N/2)$ associated with any maternal effects resulting from the sampling of $N/2$ mothers;

iii) The residual variance $\sigma_{E_s}^2/(Nn/2)$ associated with special environmental effects averaged over the entire progeny pool.

Finally, the among-line variances in consecutive generations will be correlated as a consequence of shared ancestry,

$$\sigma_B(t,t') = \left(\frac{1}{N_{fo}} + 2f_t\right)\sigma_A^2(0) \qquad \text{for } 0 < t < t' \tag{12.2}$$

Equation 12.2 assumes no transmission of maternal effects across generations, which if present would further inflate this covariance.

A few words should also be said about the potential importance of nonadditive gene action. From Table 11.3, it can be seen that in the presence of dominance, the among-population variance (in the absence of any new mutation) eventually asymptotes at $\sigma_B^2 = 2\sigma_A^2 + 2\sigma_{ADI} + \sigma_{DI}^2$. Thus, dominance can magnify or reduce the among-population variance depending upon the magnitudes of σ_{DI}^2 and σ_{ADI} and on the sign of the latter. In addition, the asymptotic contribution from epistatic interactions involving additive effects is equal to $2^n\sigma_{A^n}^2$ for n-locus epistasis, i.e., $4\sigma_{AA}^2$ for additive $\times$ additive epistasis, and $8\sigma_{AAA}^2$ for additive $\times$ additive $\times$ additive epistasis. Thus, epistasis involving large numbers of loci can, in principle, greatly magnify the among-population variance, even if it appears to be of relatively minor importance within populations.

Sampling Error

We now consider the sampling properties of the among-population genetic variance by reference to a particular experimental design, again assuming a character with a strictly additive basis (Hill 1972; Lynch 1988). Starting from a base population with additive genetic variance $\sigma_A^2(0)$, L replicate lines are isolated and subsequently maintained each generation with $N/2$ random monogamous matings. Due to the fact that only a finite number of lines is studied, the among-population variance that actually develops in any particular experiment (the **realized varaince**), $\widehat{\sigma}_B^2(t)$, will deviate from the expectation $\sigma_B^2(t)$ given by Equation 12.1b. Moreover, due to finite sample sizes within populations, the among-population variance estimated by the investigator, Var(B,t) (which we will denote $V_B(t)$ for brevity) will further deviate from $\widehat{\sigma}_B^2(t)$. This first source of variation, $\sigma^2[\,\widehat{\sigma}_B^2(t) - \sigma_B^2(t)]$, is a function of population-genetic structure and, for a fixed system of mating, is largely beyond the control of the investigator. The second source of variation, the **sampling variance**, $\sigma^2[V_B(t) - \widehat{\sigma}_B^2(t)]$, arises in estimating $\widehat{\sigma}_B^2(t)$ from the among-line sample variance $V_B(t)$. Its contribution can be minimized by the use of large sample sizes.

Since our concern here is variation in divergence due to genetic changes generated by random drift, we focus on the situation in which the among-line divergence has been measured in such a way as to eliminate nongenetic causes (such as a common-garden experiment to removal any environmental trends). Suppose that the same experiment has been repeated many different times, on each occasion starting with L lines from the same base population. Due to the variation in the drift process and the finite number of observed lines, each set of experimental lines will develop its own temporal pattern of realized among-population variance. The expected variation in the realized variance among these hypothetical replicate experiments provides a measure of confidence that one can have in the results of any single experiment. Letting $\widehat{\sigma}_B^2(t)$ be the realized among-population variance at generation t for a particular experiment, the expected variance of this quantity among replicate experiments is

$$\sigma^2[\,\widehat{\sigma}_B^2(t)\,] \simeq \frac{4\sigma_A^4(0)}{L-1}\left[\frac{1}{2N_{fo}^2} + 2\left(1+\frac{1}{N_{fo}}\right)f_t^2 + \sigma_f^2(t)\right] \tag{12.3}$$

Although, in practice, one generally performs such a divergence experiment only once, the utility of Equation 12.3 is that it is entirely expressed in terms of observable parameters, so that some idea of the reliability of estimates of $\sigma_B^2(t)$ can be determined in advance. In most situations, the terms in Equation 12.3 involving the founder number (N_{fo}) will be of second or third order and can be ignored.

The variance $\sigma_f^2(t)$ in the amount of actual inbreeding between individuals in the population requires additional comments. This has been examined in detail in Lynch (1988), drawing heavily from the results of Weir et al. (1980) and Cockerham and Weir (1983). For freely recombining loci, σ_f^2 is zero when the pedigree structure is fixed, e.g., obligate selfing, full-sib mating, the maximum avoidance systems of Wright (1921), and the circular systems of Kimura and Crow (1963a); and even with fairly tightly linked loci, $\sigma_f^2(t)$ is generally negligible in any generation under selfing or full-sib mating. However, under most natural mating schemes, some individuals mate by chance with closer relatives than do others. This results in variation in f among members of the same population, which because of sampling, accumulates as among-population variance in f (different lines stochastically accumulate different amounts of inbreeding). The theoretical value of $\sigma_f^2(t)$ under different systems of mating is of special interest because empirical studies usually do not record the essential pedigree information for its computation. For larger population sizes, even with unlinked loci, if the sexes are separate and matings are monogamous, its squared coefficient of variation $[\mathrm{CV}(f_t)]^2 = \sigma_f^2(t)/f_t^2$ can attain values of 0.1 to 1.0 in the first two to four generations of isolation, which is enough to contribute significantly to $\sigma^2[\widehat{\sigma}_B^2(t)]$. However, after six or so generations have passed, $\sigma_f^2(t)$ can be safely ignored regardless of the population size, even with tightly linked loci.

Ignoring the initial founder effect, these results indicate that the coefficient of variation of the among-population variance is $\sqrt{2\{1+[\mathrm{CV}(f_t)]^2\}/(L-1)}$, which is generally on the order of $\sqrt{2/L}$, although in some cases being as high as $2/\sqrt{L}$. Thus, studies of phenotypic divergence *need to have very large number of replicates to be statistically reliable*. For example, if it is desirable to reduce the standard error of the among-line variance to 10% of the expectation under the null hypothesis of neutrality and additivity ($\sqrt{2/L} = 0.1$), a minimum of 200 lines should be studied.

One can assess the fit of the additive theory to actual data under two different settings. In the first, we have a single estimate of the among-line variance and we compare this result to the value expected from theory (see Example 12.2). In the second, we have a series of

among-line estimates at different time points, allowing us to consider the temporal pattern of increase in σ_B^2, which as noted above, should eventually reach a constant as $f \to 1$. When such a temporal sequence of $V_B(t)$ is available, these may be regressed on f_t. Under the null hypothesis of neutral additive genes, from Equation 12.1b the expected slope of such a regression is $2\sigma_A^2(0)$. However, because of shared ancestry, consecutive estimates of mean phenotypes obtained from the same lines are nonindependent (Equation 12.2), violating a fundamental assumption of ordinary least-squares (OLS) regression analysis, and generalized least squares (GLS) must be used instead (Chapter 18, LW Chapter 8). For example, once the lines have become completely inbred (and ignoring mutation), all future values of $\widehat{\sigma}_B^2(t)$ must be fixed, and therefore should not be given equal weight in the regression analysis. The expected covariance of $\widehat{\sigma}_B^2$ between generations with inbreeding levels f_t and $f_{t'}$ is

$$\sigma[\,\widehat{\sigma}_B^2(t),\,\widehat{\sigma}_B^2(t')\,] \simeq \frac{4\sigma_A^4(0)}{L-1}\left[\frac{1}{2N_{fo}^2} + 2\left(1+\frac{1}{N_{fo}}\right)f_t f_{t'} + \lambda_1^{t'-t}\sigma_f^2(t)\right] \quad \text{for } t < t' \quad (12.4)$$

where $\lambda_1 = 1 - 1/(2N)$. Lynch (1988) provides approximate expressions for the standard errors of the slope and intercept that account for the intrinsic correlations in the data, assuming measurements of $V_B(t)$ in progressive generations. Chapter 18 also considers the same problem, but in the context of response in a selection experiment and frames the solution in a GLS framework. The variance of the regression coefficient increases with the duration of the experiment, but is essentially constant after the fourth generation of inbreeding. At that point, the standard error ranges from approximately $4\sigma_A^2(0)/\sqrt{L}$ under obligate self-fertilization to $3\sigma_A^2(0)/\sqrt{L}$ with larger N_e, implying coefficients of variation in the range of $1.5/(f\sqrt{L})$ to $2/f\sqrt{L}$. For large f, these are not greatly different from the sampling variances of single-point estimates noted above.

Sample Variance Confidence Intervals

Given the critical role played by the sample variance in empirical tests of the additive-drift model, we digress here to briefly consider a few statistical issues related to estimating a variance from a sample, which are used throughout the rest of the chapter. Provided individual observations used to estimate a sample variance are uncorrelated with $y_i \sim N(\mu, \sigma^2)$, then (LW Equation A5.14c) for a sample of size n we have for $\text{Var} = \sum(y_i - \overline{y})^2/(n-1)$ that

$$(n-1)\text{Var} \sim \sigma^2\chi_{n-1}^2 \quad (12.5a)$$

As a result, confidence intervals for the true variance σ^2 based on the observed sample variance Var follow from critical values for a χ^2 distribution. Letting $X_{p,n}$ satisfy $\Pr(\chi_n^2 \le X_{p,n}) = p$, then

$$\Pr(X_{\alpha/2,n} \le \chi_n^2 \le X_{1-\alpha/2,n}) = 1-\alpha \quad (12.5b)$$

From Equation 12.5a, substituting $(n-1)\text{Var}/\sigma^2$ for χ_{n-1}^2, we have

$$\Pr\left(X_{\alpha/2,n-1} \le \frac{(n-1)\text{Var}}{\sigma^2} \le X_{1-\alpha/2,n-1}\right) \quad (12.6a)$$

$$= \Pr\left(\frac{1}{X_{\alpha/2,n-1}} \ge \frac{\sigma^2}{(n-1)\text{Var}} \ge \frac{1}{X_{1-\alpha/2,n-1}}\right) = 1-\alpha \quad (12.6b)$$

giving

$$\Pr\left[\left(\frac{n-1}{X_{1-\alpha/2,n-1}}\right)\text{Var} \le \sigma^2 \le \left(\frac{n-1}{X_{\alpha/2,n-1}}\right)\text{Var}\right] = 1-\alpha \quad (12.7)$$

This motivates a $(1-\alpha)100\%$ confidence interval for the true variance σ^2 given the observed sample variance Var. As shown in Figure 12.1, confidence intervals for σ^2 are asymmetrical about Var, and tend to be quite large even for modest sample sizes.

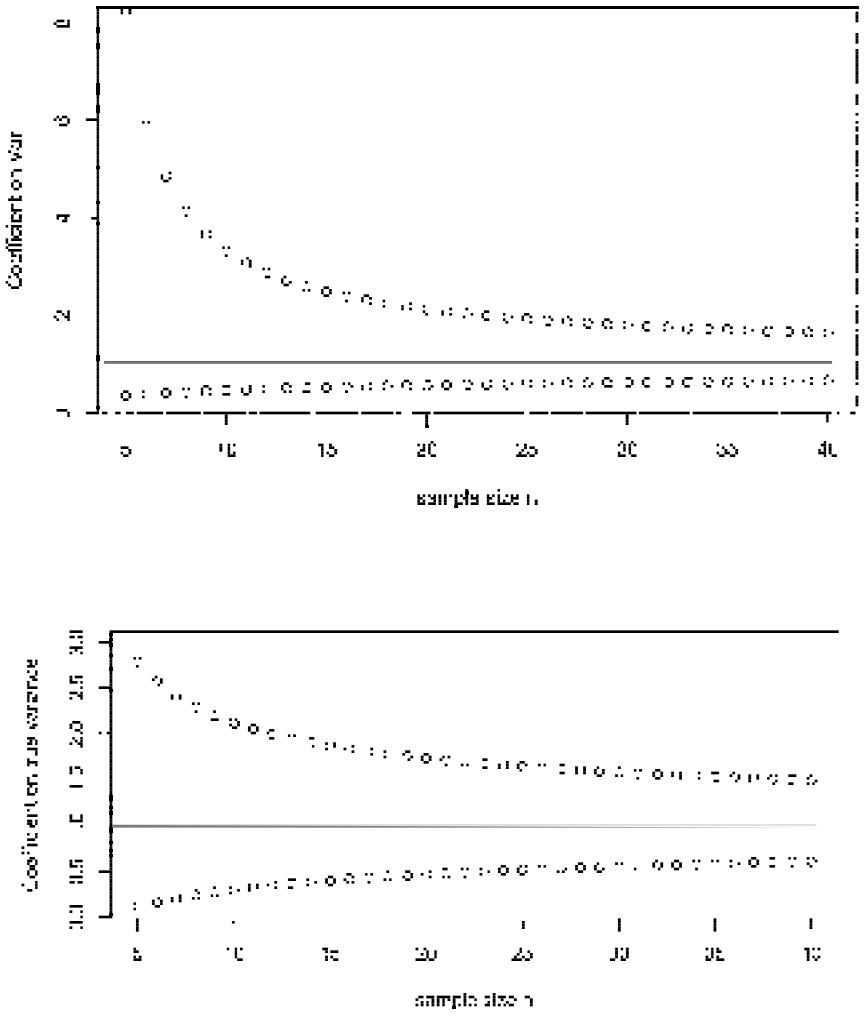

Figure 12.1. Confidence limits and critical values for σ^2 estimated from a sample of n observations. **A (Top):** Upper and lower values for the 95% confidence interval in σ^2 based on an observed sample variance Var. For example, for $n = 10$, the 95% confidence interval for σ^2 is 0.44·Var to 3.33·Var. **B (Top):** Upper and lower 5% critical values for an observed sample variance given an assumed variance σ_0^2. For example, for $n = 10$, 95% of the values of Var are expected to fall within the interval $0.30{\cdot}\sigma_0^2$ to $2.11{\cdot}\sigma_0^2$.

If the true variance is really $\sigma_1^2 \neq \sigma_0^2$, then the **power** (LW Appendix 5) for this parameter value is just the probablity that a sample variance falls outside of the interval given by

Equation 12.7, which is a function of the sample size n and the assigned significance α for the test. Letting β denote the probability of a type II error (failing to declare a test significant when the null is false, i.e., $\sigma^2 = \sigma_1^2 \neq \sigma_0^2$), we can obtain this from Equation 12.7 by noting that now $[(n-1)/\sigma_1^2]\,\mathrm{Var} \sim \chi^2_{n-1}$. Multiplying all terms of Equation 12.7 by $(n-1)/\sigma_1^2$ gives the probablity β of a sample variance failing to be declared significant as

$$\begin{aligned}\beta &= \Pr\left[\left(\frac{\sigma_0^2}{n-1}\right)\left(\frac{n-1}{\sigma_1^2}\right) X_{\alpha/2,n-1} \leq \chi^2_{n-1} \leq \left(\frac{\sigma_0^2}{n-1}\right)\left(\frac{n-1}{\sigma_1^2}\right) X_{1-\alpha/2,n-1}\right] \\ &= \Pr\left[\left(\frac{\sigma_0^2}{\sigma_1^2}\right) X_{\alpha/2,n-1} \leq \chi^2_{n-1} \leq \left(\frac{\sigma_0^2}{\sigma_1^2}\right) X_{1-\alpha/2,n-1}\right] \qquad (12.8a)\end{aligned}$$

Hence, the power $1-\beta$ is

$$\Pr\left[\chi^2_{n-1} \leq \left(\frac{\sigma_0^2}{\sigma_1^2}\right) X_{\alpha/2,n-1}\right] + \Pr\left[\chi^2_{n-1} \geq \left(\frac{\sigma_0^2}{\sigma_1^2}\right) X_{1-\alpha/2,n-1}\right] \qquad (12.8b)$$

Example 12.1. Consider a sample variance estimated from $n = 10$ observations (e.g., the between-group variance estimated from the means of ten replicate lines). Since $\Pr(\chi^2_9 \leq 2.700) = 0.025$ and $\Pr(\chi^2_9 \leq 19.023) = 0.975$, Equation 12.7 gives the 95% confidence interval ($\alpha = 0.05$) for the true variance σ^2 as between $(9/19.023)$Var and $(9/2.7)$Var, or 0.473·Var to 3.333·Var, for an uncertainty in σ^2 spanning almost a full order of magnitude.

What observed values of the sample variance are unlikely given an assumed variance of σ_0^2? From Equation 12.7, the upper and lower critical values (for a two-sided test with $\alpha = 0.05$) are $(2.700/9)\sigma_0^2 = 0.3 \cdot \sigma_0^2$ and $(19.023/9)\sigma_0^2 = 2.11 \cdot \sigma_0^2$. Finally, what is the power of this design (again taking $\alpha = 0.05$) when $\sigma_1^2 = 2\sigma_0^2$? Equation 12.8b gives the power as

$$\Pr\left(\chi^2_9 \leq \frac{2.700}{2}\right) + \Pr\left(\chi^2_9 \geq \frac{19.023}{2}\right) = 0.39$$

and hence a type II error rate of 61% when the true variance is twice the assumed variance. A similar calculation assuming $\sigma_1^2 = \sigma_0^2/2$ gives a power of 0.20, or a type II error rate of 80%. Useful `R` commands for these calculations are `pchisq(x,n)`, which returns $\Pr(\chi^2_n \leq x)$, and hence `1- pchisq(x,n)` returns $\Pr(\chi^2_n \geq x)$, while `qchisq(p,n)` returns $X_{p,n}$.

Empirical Observations

As an example of the application of the preceding theory, consider the results from a large drift experiment with laboratory cultures of the flour beetle *Tribolium castaneum* (Rich et al. 1984). The authors followed twelve replicate populations at four population sizes (1:1 sex ratio, random mating) over 20 consecutive generations. Each generation, the mean pupal weight (in μg) of each population was obtained from a bulk sample of 100 random individuals. The additive genetic variance was estimated to be 460 in the base population. The observed $V_B(t)$ are plotted as a function of f_t in Figure 12.2, along with the expected divergence $2\sigma_A^2(0)f_t = 920 f_t$ (solid lines). The dashed lines, obtained by using Equation

12.3 for the expected variance and substituting this into Equation 12.7 (using $\alpha = 0.05$ and $n = 12$), give the limits of the among-population variance beyond which there is less than a 5% chance for the realization of the drift process under the null to generate these values. Since these bounds ignore measurement error, they may be regarded as conservative confidence limits (as they assumed a smaller variance and hence are too narrow). Nevertheless, almost all of the observations, with the exception of the clusters of the late generations at $N = 10$ and 20, lie within these limits. The least-squares regressions of the data are given by the dotted lines (more formally, a GLS regression would be used to account for correlated and heteroscedastic residuals, see Chapter 18). The slope of each regression is less than the expected 920, but all are within two standard errors of the expectation. The observed patterns are fairly consistent with a hypothesis of random drift of neutral additive genes. The observed declines in $V_B(t)$ late in the experiment at the two smallest population sizes may have simply arisen by chance and remained there due to intergenerational correlations (Equation 12.2).

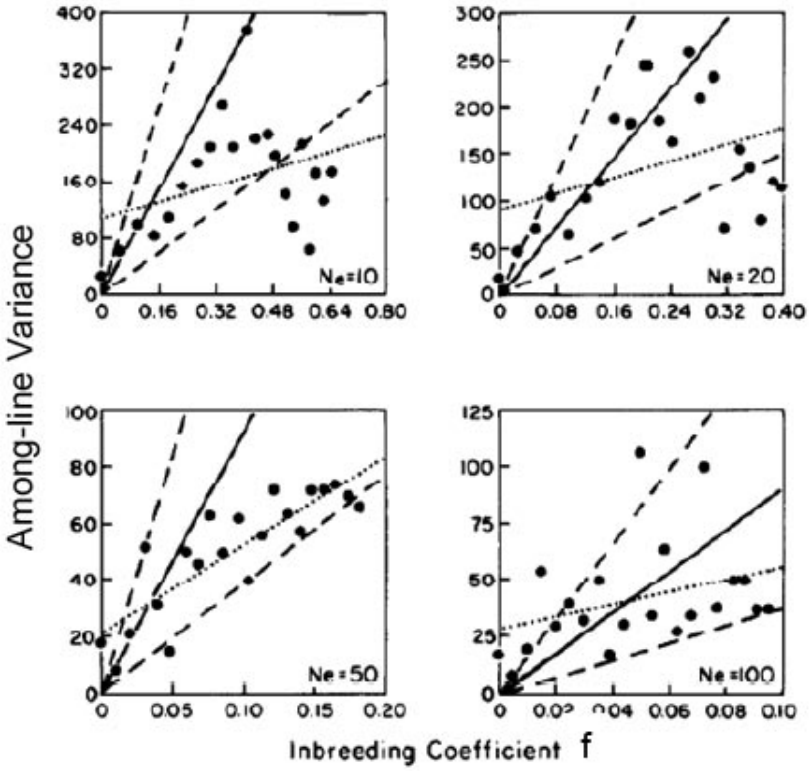

Figure 12.2. Observed and expected levels of the among-population variance for pupal weight in a divergence experiment with the flour beetle *Tribolium*. Solid lines are the expected divergence, dotted lines the least-squares regression of the observations, and the paired dashed lines give the approximate 95% confidence interval. Data from Rich et al. (1984).

The results of some other short-term divergence experiments previously given in Figure 11.3 show no evidence for nonlinear increases in the among-population variance with inbreeding. Eisen and Hanrahan (1974) have argued that the divergence of growth and

reproductive rates in inbred lines of mice is more rapid than can be accounted for by the additive genetic variance in the base population, and Bryant et al. (1986) suggested the same for morphological traits in bottlenecked housefly lines. The implication of these authors is that some nonadditive variance is converted by inbreeding into σ_A^2 (Chapter 11), leading to a faster between-line divergence. In neither case was it verified that the departures from expectations were significant, but this lack of significance is tempered by the low power of their designs.

Lande's Constant Variance F_{CV} Test

Is an observed divergence over a modest amount of time significantly different than expected by drift? For the case in which one has only a single estimate of the among-population divergence, Lande (1977) suggested the statistic

$$F_{CV} = \frac{V_B(t)}{t \cdot \text{Var}(A,0)/N_e} \tag{12.9a}$$

as a test for neutrality. This is the **constant variance** version of the test, as it assumed that the additive variance remains unchanged over the period being followed (Turelli et al. 1988). As noted by Lande, under approximate assumptions, this follows an F distribution, which we can show as follows. Assuming the trait is normally distributed, the sample mean $\overline{y}_i \sim N[\mu(0), \sigma_B^2(t)]$, where we have assumed that sampling variance terms are small enought to be ignored. With L independent lines (i.e., they have a star phylogeny), Equation 12.5a gives

$$V_B(t) = \frac{1}{L-1}\sum_{i=1}^{L}(\,\overline{y}_i - \overline{y}_{\cdot}\,)^2 \sim \frac{\sigma_B^2(t)}{L-1}\chi_{L-1}^2 \tag{12.9b}$$

Ignoring the (usually) small founder effect, Equation 12.1b gives

$$\sigma_B^2(t) = 2f_t\sigma_A^2(0) = 2\left[1 - \left(1 - \frac{1}{2N_e)}\right)^t\right]\sigma_A^2(0) \simeq t\sigma_A^2(0)/N_e \quad \text{for } t \ll N_e \tag{12.9c}$$

and hence

$$V_B(t) \sim \frac{t\sigma_A^2(0)/N_e}{L-1}\chi_{L-1}^2 \tag{12.9d}$$

Assuming that $\text{Var}(A,0)$ is a good estimate of $\sigma_A^2(0)$, substitution into Equation 12.9a gives

$$F_{CV} \sim \frac{\chi_{L-1}^2}{L-1} \sim F_{L-1,\infty} \tag{12.9e}$$

The last step follows from the definition of an F distribution (LW Appendix 5). Hence Lande's F_{CV} statistic follows an F distribution with $L-1$ numerator and infinite denominator degrees of freedom. More generally, since $\sigma_A^2(0)$ is estimated by $\text{Var}(A,0)$, the denominator degrees of freedom are those associated with this estimate. Finally, as suggested by Turelli et al. (1988) and Lynch (1988, 1990), given the sampling variance associated with the means, estimating the between-group variance from a standard one-way ANOVA is preferable, with

$$V_B(t) = \frac{MS_B - MS_W}{n_0} \tag{12.9f}$$

using the between- and within-group mean squares, and n_0 is a measure of the average number per group (see LW Table 18.1).

A couple of approximations were required to reach Equation 12.9e. One check of their validity is that if $x \sim \sigma^2\chi_n^2/n$, then $\sigma^2(x) = 2\sigma^4/n$ (as $\sigma^2(\chi_n^2) = 2n$, LW Equation A5.15b). Hence, the numerator should have a variance approximately equal to $2[2f_t\sigma_A^2(0)]^2/(L-1)$. Ignoring the added contribution from sampling error, this can be seen to be approximately true for large N_{fo} by reference to Equation 12.3. However, with selfing and full-sib mating, the expected variance is about twice and 1.5 times too high respectively. Thus, Lande's approach should be restricted to lines with at least moderate effective size. Moreover, as we will see below, all of the proceeding formulae for σ_B^2 become questionable for $t > N_e$, because they ignore the contribution from new mutations. Hence, Lande's F_{CV} test is best thought of as one for short-term divergence, such as would be seen in a laboratory experiment or at most a modest amount of time in a set of natural populations.

Example 12.2. Lande (1977) used Equation 12.9a to evaluate the results of a 12-year divergence experiment involving five populations of *Drosophila pseudoobscura* (Anderson 1973). Two of the populations had been maintained at 25°C, two at 27°C, and one at 16°C. They were then raised in two common environments (16 and 25°C) and measured for wing length. Estimates of the additive genetic variance for these two environments were 0.88 and 0.77, while the among-population variances were approximately 6.62 and 4.37 respectively. An approximate upper bound for the number of generations of divergence is $t = 150$, whereas the effective population size probably always exceeded $N_e = 1000$. The use of these extreme bounds gives conservative estimates of F_{CV}, making it more difficult to demonstrate diversifying selection on wing length. The resulting values (for the two environments) are

$$\mathrm{F}_{CV,1} = \frac{6.62}{150 \cdot 0.88/1000} = 50.15, \quad \text{and} \quad \mathrm{F}_{CV,2} = \frac{4.37}{150 \cdot 0.77/1000} = 37.84$$

Given that $\Pr(F_{4,\infty} > 4.6) = 0.001$, both values are highly significant. Thus, the hypothesis that the observed line divergence is solely attributable to random genetic drift can be rejected confidently. More likely, the different thermal conditions resulted in selection for different wing lengths.

LONG-TERM DIVERGENCE

Our previous results were simply concerned with how any initial variation is partitioned among lines during drift/inbreeding. While this is occurring, new variation is constantly being generated by mutation, further driving divergence (Haldane 1949). Polygenic mutation was first incorporated into the theory of population divergence by Dempster (appendix in Bailey 1959) and was subsequently studied by Lande (1976), Chakraborty and Nei (1982), and Lynch and Hill (1986). As in Chapter 11, the mutational model generating the new mutational effect a_m given the ancestral effect a_0 is critical. The standard model is a version of infinite-alleles, where each new mutation gives rise to a new allele, with $a_m = a_0 + e_a$, namely the ancestral value plus a random increment e_a. To distinguish this from the standard infinite-alleles model of Chapter 2 (where allelic effects are not a concern), this is also known as the **incremental** or **Brownian motion** mutational model in the literature and is examined more fully in Chapter 27.

Again focusing on a character with a purely additive genetic basis, starting with an ancestral-population genetic variance of $\sigma_A^2(0)$, and assuming the infinite-alleles model, the expected variance of genotypic means for replicate populations isolated t generations in the past is

$$\sigma_B^2(t) = 2\sigma_m^2 t + 2[\sigma_A^2(0) - 2N_e\sigma_m^2][1 - e^{-t/(2N_e)}] \tag{12.10}$$

where σ_m^2 is the per-generation mutational rate of input of genetic variance, as described in Chapter 11. This expression shows that as t becomes large, the expected rate of increase of the among-population variance for a neutral quantitative trait becomes a constant $2\sigma_m^2$ per generation. The same formulation applies to the among-species genetic covariance for a pair of traits, if the mutational rate of production of covariance between the traits is substituted for σ_m^2 (Lande 1979).

Thus, under the infinite-alleles model, the asymptotic divergence rate is independent of the population size, just as it is in the neutral theory of molecular evolution (Chapter 7; Kimura 1983). Although the expected number of new mutations entering a population each generation is $2Nu$ per locus, the probability of fixation of a new mutation is its initial frequency $1/(2N)$, so (at equilibrium) the expected number of mutations fixed per locus per population per generation is simply u. For a set of L populations, with each fixed mutation causing an increase in expected among-population variance of $\sim E[(2a)^2]$, and n loci contributing, the asymptotic divergence rate is $nuLE[(2a)^2/L] = 2\sigma_m^2$.

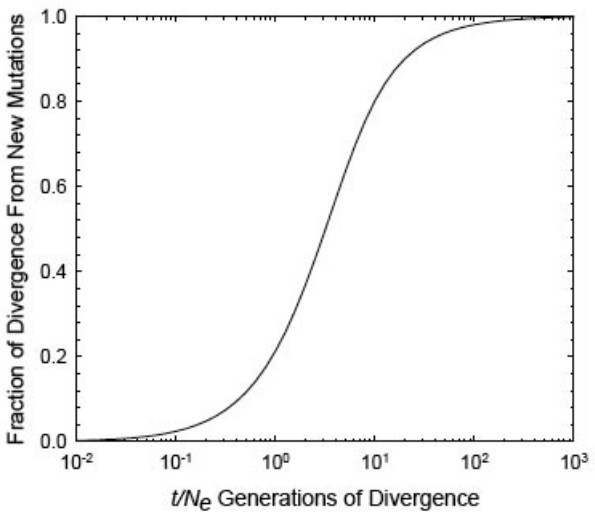

Figure 12.3. The expected fraction of among-population variance attributable to mutations arising subsequent to the isolation event. It is assumed that the base population is in drift-mutation equilibrium, $\sigma_A^2(0) = 2N_e\sigma_m^2$, with the same effective size as the daughter species, so that from Equation 12.10, the divergence due to base-population variance is $2\sigma_A^2(0)[1 - e^{-t/(2N_e)}]$. To obtain the actual number of generations of divergence for any population size, multiply the horizontal axis by N_e.

Under the assumptions of the infinite-alleles model, the asymptotic divergence rate of $2\sigma_m^2$ is a fairly general result. It is independent of the degree of dominance of new mutations,

of the linkage relationships of the constituent loci, and of the mating system (Lynch and Hill 1986). This is because both dominance and gametic-phase disequilibrium are transient properties of alleles en route to loss or fixation, and not cumulative phenomena, and because the probability of fixation of a new neutral mutation is equal to its initial frequency regardless of the breeding system.

How long should populations be isolated before one should start to worry about the contribution of new mutations to their divergence? From Equation 12.10, it can be seen that this depends on the initial level of genetic variance and on the effective sizes of the derived isolates. In Figure 12.3 it is assumed that the initial base population is in drift-mutation equilibrium, so that $\sigma_A^2(0) = 2N_e\sigma_m^2$, and that the isolated lineages have rapidly attained the same effective sizes (N_e). Under these circumstances, by the time N_e generations have elapsed, polygenic mutation subsequent to the isolation event has caused about 20% of the divergence, whereas for $t > 3N_e$ generations, the majority of the divergence is due to new mutations.

As emphasized in the preceding chapter (and in Chapter 27), alternatives exist to the infinite-alleles model, raising questions about the appropriate structure of a neutral null model. For example, Cockerham and Tachida's (1987) model, which assumes a finite number of alleles with each new mutational effect being independent of the prior allelic state (the house-of-cards model), yields an equilibrium among-population variance

$$\sigma_B^2 = 2[1 - E(H)]\sigma_A^2(\infty) \tag{12.11}$$

where $E(H)$ is the expected heterozygosity per locus, and $\sigma_A^2(\infty) = 2nE(a^2)$ is the expected additive genetic variance in a population of infinite size (Chapter 11). Note that under this model, not only does the among-population variance not build up indefinitely, but as $4N_eu \rightarrow \infty$, driving the heterozygosity to 1.0, the among-population component of variance asymptotically approaches zero. This is because under the house-of-cards model, replicate populations that are each effectively infinite in size will individually harbor the same alleles with the same frequency spectrum defined by the mutational interconversion rates.

If nothing else, these dichotomous results indicate that although neutral models are essential to demonstrating the necessity of invoking natural selection to explain an observed pattern of divergence, the actual *construction* of the null model depends on unresolved biological issues. Recall from Chapter 11 that Zeng and Cockerham (1993) proposed a model bridging infinite alleles (incremental) and house-of-cards, where the effect of a mutant allele is given by $a_m = \tau a_o + e_a$. Taking $\tau = 1$ covers the incremental model, while $\tau = 0$ the house-of-cards. Since τ gives a measure of the dependency on the current value, this is often called the **regression mode**l (see Chapter 27 for more details). Using their model, the equilibrium among-population variance becomes

$$\sigma_B^2 = \frac{4E(a^2)}{(1-\tau)^2[1 + 4N_eu(1-\tau)]} \tag{12.12}$$

where $\tau = 1$ under the Lynch-Hill model (infinite alleles) and 0 under the Cockerham-Tachida (house-of-cards) model. For $\tau < 1$, the approach to the equilibrium level of divergence is defined by the mutation rate (u), assuming an identical N_e in the base and descendant populations,

$$\sigma_B^2(t) = [1 - (1-u)^{2t}]\sigma_B^2 \tag{12.13}$$

and hence is quite slow (approximately $2u$ per generation).

Finally, we note that the expression for the variance of the among-population variance (i.e., the variance of $\sigma_B^2(t)$ among replicate experiments with mutational input) is

algebraically complex, and has only been worked out for the infinite-alleles model (Lynch and Hill 1986). However, if it is assumed that the number of loci is large and the distribution of mutational effects is normal with mean zero, the variance of the realized among-population variance approaches $2(2\sigma_m^2 t)^2/L$ for large t. This is simply twice the square of the expected among-population variance. Thus, for large t, the coefficient of variation of a realized among-population variance based on L lines is expected to be on the order of $\sqrt{2/L}$, so as we have noted before, unless L is quite large, estimates of σ_B^2 can deviate quite far from their expectation.

Effectively Neutral Divergence and the Estimation of Rates of Mutational Variance

As discussed in detail in LW Chapter 12, the theoretical expectations of the neutral model provide the basis for estimating the rates of polygenic mutation. Starting from an inbred base population, experimental lines with known times of divergence can be used to estimate the amount of polygenic mutation that is necessary to account for the distribution of the resultant mean phenotypes. In one of the earliest endeavors of this sort, Russell et al. (1963) started with several lines of maize that had been maintained by prolonged self-fertilization. They then performed a dichotomous branching experiment for five generations in which each plant was self-fertilized to produce two new daughter sublines. Seed was saved from each generation, so that at the end of the experiment members of all generations could be assayed simultaneously in a common environment, and then sib analysis was used to estimate the additive genetic variance for the *total* population each generation. Assuming the within-population variance to be in drift-mutation equilibrium, this type of population expansion should give rise to an average rate of increase in the total genetic variance of $2\sigma_m^2$ per generation. In accordance with this prediction, the regressions of the genetic variance on time were positive for every character investigated (Figure 12.4). The rate of polygenic mutation for each of the traits is thus estimated by one-half the slope.

Results from many other experiments of this sort were reviewed in LW Chapter 12. Although a number of additional results have emerged since then, most of these are confined to a small number of model systems, and the conclusions reached in our earlier review remain unaltered. Here, we simply give a brief update, providing references only to post-1998 papers. Most estimates are framed in terms of the mutational heritability, $h_m^2 = \sigma_m^2/\sigma_e^2$. Estimates of h_m^2 for a diversity of morphological, physiological, and life-history traits in *D. melanogaster* are consistently in the range of 0.001 to 0.005. Mutational heritabilities for body size and life-history traits in nematodes fall in the range of 0.001 to 0.008 (Vassilieva et al. 2000; Baer et al. 2006; Ostrow et al. 2007), and the same is true for life-history traits in the microcrustacean *Daphnia pulex* and in the grape phylloxera *Daktulosphaira vitifoliae* (Downie 2003). Thus, essentially all studies with invertebrates imply $0.001 < h_m^2 < 0.01$ for complex traits.

Although the numbers of studies are still rather limited, estimates of h_m^2 for some land plants and vertebrates appear to be several-fold higher than those noted above. Mutational heritabilities for growth and reproductive traits in *Arabidopsis thaliana* are in the range of 0.001 to 0.008 (Schultz et al. 1999; Shaw et al. 2000; Chang and Shaw 2003; Kavanaugh and Shaw 2005), but the average h_m^2 for maize, from the study of Russell et al. (1963), is 0.0092. In addition, mutational heritabilities for morphological and reproductive traits in mice fall in the range of 0.003 to 0.023 (Casellas and Medrano 2008). Thus, there is at least a rough indication that mutational heritabilities are increased in organisms with longer life spans, which might in principle be a consequence of elevated rates of mutation per generation

(Chapter 4).

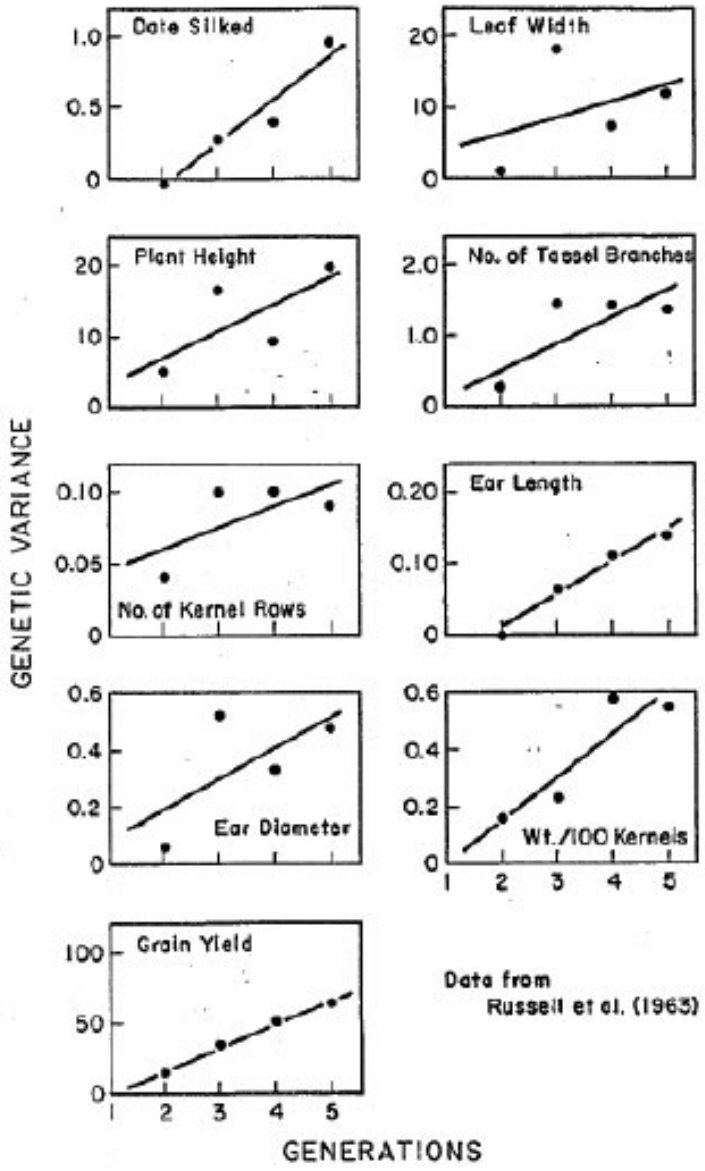

Figure 12.4. The increase in additive genetic variance (within- plus among-population components) in an expanding set of lines of corn. From Russell et al. (1963).

Finally, it should be emphasized that in all mutation-accumulation experiments, fitness declines in the vast majority of lines, indicating that mutations are on average deleterious, although the fraction of mutations that are beneficial remains unclear (Shaw et al. 2002; Keightley and Lynch 2003; Charlesworth and Eyre-Walker 2007; Eyre-Walker and Keightley

2007; Dickinson 2008; Hall et al. 2008). Equally importantly, for characters that influence fitness only indirectly (e.g., morphology), the fraction with negative pleiotropic effects on fitness remains unclear. Hence, estimates of h_m^2 from mutation-accumulation experiments with their very small effective population sizes may overestimate, perhaps significantly, the actual *usable* amount of h_m^2 for most populations. Further, if one imagines that deleterious pleiotropic effects are often small, as N_e increases, the fraction of all new mutations which are effectively neutral decreases, so that the effective value of h_m^2 is likely a decreasing function of N_e. What is unclear is whether this plateaus fairly quickly or continues to decrease over a large range of N_e. Resolving these issues are critical to any attempts to utilize estimates of mutational heritability to infer long-term mechanisms of evolution, as illustrated in the following section.

TESTING THE NULL HYPOTHESIS OF NEUTRAL PHENOTYPIC DIVERGENCE: RATE-BASED TESTS

One of the enduring problems in evolutionary biology is the struggle to demonstrate that various aspects of biodiversity are products of diversifying selection. It is one thing to concoct plausible adaptive scenarios to explain patterns of morphological, physiological, or behavioral divergence, but quite another to formally demonstrate that an observed level of divergence cannot simply be explained by a null model of random genetic drift. This is especially the case with -omics data.

Three broad classes of tests for departures from neutral divergence have been proposed, each requiring successively more difficult-to-obtain information. The first (**rate tests**) simply require the means of divergent populations, and ask (given N_e, t, and either h^2 or h_m^2) whether the rate of divergence is consistent with drift. A second approach is based entirely on the internal information in a temporal sequences of trait values, typically from the fossil record. A third class requires information on the additive variation of a trait and the frequencies of a set of markers in a subdivided population. Such $Q_{ST} - F_{ST}$ tests compare the within- and between-population structure of the additive variation of a trait (measured by Q_{ST}) with that from some presumably neutral set of markers (measured by F_{ST}, Chapter 2). The final type of tests is based on the pattern of QTL substitutions, and either requires population crosses or GWAS analyses. Under neutrality, the pattern of fixed QTL effects should be largely random, i.e., no excess of plus (increasing) QTL alleles in one line. We examine all these approaches in turn, starting with rate tests.

The molecularly-focused reader may (incorrectly) get the impression that despite the use of marker information in the last two classes of tests, none of these approaches are not relevant to their work. Nothing could be further from the truth. Methods initially developed to test for neutrality in changes in morphology (often over a fossil record) are now widely applied to omics data. To highlight this transition from morphology to molecules, we conclude by applying these tests to gene expression data to examine if their divergence is largely governed by neutral drift, stabilizing selection, or directional selection.

Rate Tests

Lande's F_{CV} (Equation 12.9) is one example of a rate test, wherein one compares the amount of divergence across a set of lines/populations with the expected between-population variance (σ_B^2). As mentioned, this test is best applied over short time scales ($t \ll N_e$), such as might occur in a selection experiment or over a short to modest amount of time in nature.

Here we focus on tests potentially over much longer time scales and ask whether an observed amount of total divergence $d = \mid \mu(t) - \mu(0) \mid$ within a single lineage is excessively

large (or small) relative to drift. As with comparisons based on the amount of divergence over a set of lines or populations, the expected total divergence within a single lineage is also a function of the between-population variance σ_B^2. Tests for unusual amounts of divergence (large or small) are framed by asking what critical values for an effective population size N_e or mutation variance σ_m^2 are consistent with the amount of divergence and whether these values are biological reasonable. If they are, the hypothesis of drift alone accounting for the pattern is not rejected. It should be stressed that considerable selection may have shaped the observed pattern, and yet we can still fail to reject the drift model.

While these tests are widely used, they have several important caveats. For any analysis of this sort to be meaningful, one must be confident that the majority of population divergence is genetic, and not inflated by environmental effects on phenotypes. This is clearly problematical when populations cannot be assayed in a "common-garden" environment, such as data from the fossil record. In addition, the divergence of means is best estimated by a formal analysis of variance (Equation 12.9f) so as to eliminate the inflation of the divergence by sampling error of the population means (Lynch 1988, 1990; Turelli et al. 1988; Björklund 1991; Savalli 1993). A further complication is that expressions for critical effective population sizes or mutation rates ignore the sampling error of all other terms. Finally, as we have discussed, the infinite-alleles model for neutral trait evolution (where σ_B^2 is an ever-increasing function of time) may be too extreme, as the usable amount of σ_m^2 likely decreases with increasing N_e. Given all these considerations, it should be clear that the following tests for neutrality cannot be regarded as being very rigorous in a statistical sense. As was the case of most of the tests for selection on specific genes developed in Chapters 9 and 10, they are best employed as diagnostic guides to prioritize traits for further study.

Lande's Brownian Motion Model of Neutral Trait Evolution

The basic structure of tests for neutral trait divergence is that $\mu_t \sim (\mu_0, \sigma_B^2[t])$, namely the mean at time t has as expected value equal to the initial mean μ_0 and has variance $\sigma_B^2(t)$. To proceed further, we need additional assumptions about the actual *distribution* from which the means are sampled, which is generally assumed to be Gaussian (normal). Support for this assumption traces back to Lande (1976), who framed mean divergence in terms of a **Brownian motion process** (Appendix 1). Under the simplest Brownian motion model, the expected change in value over a small time interval is zero with constant variance b. Under this model, the distribution of values at time t is normal with mean x_0 (the initial value) and variance $\sigma_t^2 = bt$ (Appendix 1). Assuming a strictly additive model with no environmental trends, Lande noted that if we sample N_e individuals, their mean breeding value (i.e., the trait mean) has a sampling variance each generation of σ_A^2/N_e, which is used for b. Hence, at generation t, the distribution of phenotypic means is approximately normal with mean μ_0 (the initial mean) and variance

$$\sigma_t^2 = t\sigma_A^2/N_e \tag{12.14a}$$

This assumes a constant additive variance as well as a constant effective population size during the period of divergence being considered. Since drift can also change σ_A^2, the assumption of a constant σ_A^2 is reasonable only for $t \ll N_e$, unless the initial variance is close to its mutation-drift equilibrium value. More generally, if the additive variance and N_e are both changing each generation, then under the Brownian motion model (taking the first time point at $i = 0$ and the last at $i = t - 1$ bringing us up to generation t),

$$\sigma_t^2 = \sum_{i=0}^{t-1} \sigma_A^2(i)/N_e(i) \tag{12.14b}$$

For example, assuming a constant effective population size, the additive genetic variance at time t under drift and mutation is given by Equation 11.20b. Substituting this into Equation 12.14b gives

$$\begin{aligned}\sigma_t^2 &= \frac{1}{N_e}\sum_{i=0}^{t-1}\left[2N_e\sigma_m^2 + \left(\sigma_A^2(0) - 2N_e\sigma_m^2\right)\exp(-i/2N_e)\right] \\ &= 2\sigma_m^2 t + \left(\sigma_A^2(0) - 2N_e\sigma_m^2\right)\left(\frac{1}{N_e}\sum_{0=1}^{t-1}\exp(-i/2N_e)\right) \qquad (12.14c)\end{aligned}$$

This recovers our previous expression (Equation 12.10) for σ_B^2 under drift and mutation by noting that

$$\frac{1}{N_e}\sum_{i=0}^{t-1}\exp(-i/2N_e) \simeq 2(1-\exp[-t/2N_e)] \qquad (12.15a)$$

This useful identity follows by recalling that the partial sum of a geometric series is

$$\sum_{i=0}^{k-1} x^i = \frac{1-x^k}{1-x}. \qquad (12.15b)$$

Taking $x = \exp(-1/2N_e)$ and noting from a first-order Taylor series that

$$1-\exp(-1/2N_e) \simeq 1-\left(1-\frac{1}{2N_e}\right) = \frac{1}{2N_e} \qquad (12.15c)$$

returns Equation 12.15a.

Thus, as expected, the variance σ_t^2 of the Brownian motion process corresponds to the between-group drift variance $\sigma_B^2(t)$ in population means. The notion that (under a pure drift model) the additive variance changes until it reaches a drift-mutation equilibrium value has resulted in different parameterizations of σ_t^2 in tests of drift (Lande 1976; Turelli et al. 1988). Lande assumed a constant variance, $\sigma_t^2 = t\sigma_A^2/N_e$, but since part of his concern was evolution in the fossil record, he replaced σ_A^2 by $h^2\sigma_z^2$, giving

$$\sigma_t^2 = h^2\sigma_z^2 t/N_e \qquad (12.16a)$$

His logic was that σ_z^2 could be estimated directly from a sample in the fossil record, while h^2 values for many morphological traits fall within a relatively narrow window. Hence, either a representative value for h^2 could be used, or different values tried to examine the robustness of any conclusions. This results in a test based on joint considerations of N_e and h^2. Conversely, Turelli et al. (1988) note that if the population has been at its current size sufficiently long enough so that additive variance is at its mutation-drift equilibrium value, then (assuming the infinite-alleles model) $\sigma_A^2 = 2N_e\sigma_m^2$, giving

$$\sigma_t^2 = 2tN_e\sigma_m^2/N_e = 2t\sigma_m^2, \qquad (12.16b)$$

Under this setting, N_e does not appear in σ_t^2 and tests are based on whether the required values of σ_m^2 to be consistent with drift are plausible. This is the quantitative-trait analog of the divergence $d = t\mu$ between two neutral sequences separated by t total generations (Chapter 7), with the *allelic* mutation rate μ replaced by twice the *trait* mutational variance $2\sigma_m^2$.

Tests Based on the Brownian Motion Model

Under the Brownian motion model, $\mu_t \sim N(\mu_0, \sigma_t^2)$, leading to tests of either excessive (or too little) divergence based on simple normal theory. Suppose an absolute divergence of $d = |\,\mu(t) - \mu(0)\,|$ is observed. The probability of this under drift alone is given by

$$\Pr\left(|\,\mu(t) - \mu(0)\,| \le d\right) = \Pr\left(\frac{|\,\mu(t) - \mu(0)\,|}{\sigma_t} \le \frac{d}{\sigma_t}\right) = \Pr\left(|\,U\,| \le \frac{d}{\sigma_t}\right) \tag{12.17}$$

where U is a unit normal random variable. Lande's (1976) original test (distinct from his 1977 F_{CV} test discussed above) was based on the constant variance assumption, $\sigma_t^2 = h^2\sigma_z^2 t/N_e$. Recalling that $\Pr(|\,U\,| \le 1.96) = 0.95$, Lande's critical effective population size below which there is a $< 5\%$ probability of a deviation as large as d satisfies

$$1.96 = \frac{d}{\sqrt{th^2\sigma_z^2/N_e}}, \quad \text{implying} \quad (1.96)^2 th^2\sigma_z^2 = N_e d^2 \tag{12.18a}$$

Equation 12.18a allows one to determine critical values for either divergence time t, heritability h^2, or N_e that are consistent with drift. For example, solving for N_e gives

$$\widehat{N_e} = \frac{t \cdot h^2 \cdot 1.96^2}{d_*^2} = 3.84 \cdot \frac{t\,h^2}{d_*^2} \tag{12.18b}$$

where $d_* = d/\sigma_z$ is the divergence scaled in phenotypic standard deviations. Drift with $N_e > \widehat{N_e}$ is unlikely to generate the observed amount of divergence. For a more general test of significance level α, one replaces 1.96 by $z_{1-\alpha/2}$, where z_p satisfies $\Pr(U \le z_p) = p$ for a unit normal U. Likewise, if one is comparing the means of two species that had a common ancestor τ generations ago, then $t = 2\tau$ and d is the absolute difference between their means. For historical reasons, the above discussion has been framed in terms of d^2. More generally, the ANOVA-based estimate V_B of the between-group variance (Equation 12.9f) is often used in place of d^2 (and V_B/σ_z^2 in place of d_*^2).

Example 12.3. Reyment (1982) observed a change of $1.49\sigma_z$ over roughly 5×10^5 generations in the size of a Cretaceous foraminifer. Taking a typical heritability value of 0.3, Equation 12.18b gives the largest population size consistent with this amount of divergence as

$$\widehat{N_e} = 3.84 \cdot \frac{t\,h^2}{d_*^2} = 3.84 \cdot \frac{5 \times 10^5 \cdot 0.3}{1.49^2} \simeq 260,000$$

However, paleontological data suggests that the effective population size was greater than 10^6, implying that drift was unlikely to account for such a rapid divergence. Assuming h^2 values of 0.5, 0.7, and 1 yields critical N_e values of 433,000; 607,000; and 867,000, so that only for assumed h^2 values close to one does the critical maximal size under drift approach the assumed size of $N_e > 10^6$.

As noted by Turelli et al. (1988), the population size test given by Equation 12.18b is really *two-sided*. Lande's original test examines whether N_e may be too large to account for

the observed divergence (as might occur if directional selection was changing the mean). However, one can also inquire as to whether the *stability* of population means is too great to be compatible with neutrality (too little divergence). For a two-tailed test for neutrality with a 5% overall significance level, we use a 2.5% probability cutoff for the observed divergence being too small to be consistent with drift and a 2.5% cutoff for excessively high divergence. Since $\Pr(|U| \geq 2.24) = 0.025$, the critical maximum population size in a test that evolution has been too fast for drift is

$$\widehat{N_e}(fast) \leq \frac{t \cdot h^2 \cdot 2.24^2}{d_*^2} = 5.02 \cdot \frac{t\, h^2}{d_*^2} \tag{12.19a}$$

Since populations with smaller N_e should show more drift (and divergence), Equation 12.19a gives the largest value of N_e consisted with drift generating the observed divergence. If our assumed N_e exceeds $\widehat{N_e}(fast)$, we reject the hypothesis that drift can account for this fast a divergence (using the values in Example 12.3 returns $N_e \simeq 340,000$). Likewise, since $\Pr(\,|\,U\,|\, < 0.03) = 0.025$, the critical minimal population size in a test that evolution has been too slow (support for stabilizing selection) is

$$\widehat{N_e}(slow) \geq \frac{t \cdot h^2 \cdot 0.03^2}{d_*^2} = 0.0009 \cdot \frac{t\, h^2}{d_*^2} \tag{12.19b}$$

If our assumed N_e is *less* than $\widehat{N_e}(slow)$, we reject the hypothesis that drift can account for this slow a divergence (applying Equation 12.19b with the values from Example 12.3 gives a critical minimal N_e of 61). More generally, for a two-sided test at overall significance level α ($\alpha/2$ for too much divergence and $\alpha/2$ for too little), 2.24 and 0.03 are replaced by $z_{1-\alpha/4}$ and $z_{0.5+\alpha/4}$.

When long time scales ($t > 4N_e$) are considered, Equation 11.20d shows that the additive variance approaches its mutation-drift equilibrium value $2N_e\sigma_m^2$. Under this condition, Equation 12.16b shows that the between-group variance becomes $\sigma_B^2 = t\sigma_m^2$ (with t the total divergence time), independent of N_e (Turelli et al. 1988). This gives the **MDE** (mutation-drift equilibrium) version of Lande's F test as

$$\mathrm{F}_{MDE} = \frac{V_B(t)}{2t\sigma_m^2} \tag{12.20a}$$

Since t is the total divergence, it is replaced by $2t$ when contrasting two lineages with a common ancestor t generations ago. As above, V_B is best estimated from the between-group variance in a one-way ANOVA (Equation 12.9f). When V_B is based on more than two lineages, Equation 12.20a assumes they have a star phylogeny (Chapter 8). If not, one has to place the lineage relationships into a phylogenetic framework to account for the covariance structure imparted by shared common ancestry (Felsenstein 1985, 2004, 2008; Lynch 1991; Gu 2004).

Using the same logic leading to Equation 12.9e, with L lineages, Equation 12.5a gives $(L-1)V_B(t) \sim 2t\sigma_m^2\chi_{L-1}^2$, and Equation 12.7 gives

$$\Pr\left[\left(\frac{L-1}{X_{1-\alpha/2,L-1}}\right)V_B(t) \leq 2t\sigma_m^2 \leq \left(\frac{L-1}{X_{\alpha/2,L-1}}\right)V_B(t)\right] = 1-\alpha \tag{12.20b}$$

Thus, if an estimate of t is available, one can test the neutral hypothesis without an estimate of N_e by inquiring whether that has been too little, or to much, divergence given some

assumed value of σ_m^2 (Turelli et al. 1988). A slightly different formulation of this test is based in terms of the observed rate of divergence (Lynch 1990). Let $r = V_B/t$ be the estimated rate of divergence and $\Delta = r/\sigma_e^2$ this rate scaled in units of the environmental variance. Rearranging Equation 12.20b gives

$$\Pr\left[\left(\frac{(L-1)\Delta/2}{X_{1-\alpha/2,L-1}}\right) \leq \frac{\sigma_m^2}{\sigma_e^2} \leq \left(\frac{(L-1)\Delta/2}{X_{\alpha/2,L-1}}\right)\right] = 1-\alpha \tag{12.20c}$$

giving upper and lower bounds on the mutational heritability $\sigma_m^2/\sigma_e^2 = h_m^2$ consistent with drift. For $\alpha = 0.05$ and $L = 2$, Equation 12.20b becomes

$$\Pr\left(0.10 \cdot \Delta \leq h_m^2 \leq 509 \cdot \Delta\right) = 0.95 \tag{12.20d}$$

Implying

$$\begin{aligned} h_m^2 &< 0.10 \cdot \Delta && \text{too much divergence for drift} \\ 0.10 \cdot \Delta \leq h_m^2 &\leq 509 \cdot \Delta && \text{consistent with drift} \\ h_m^2 &> 509 \cdot \Delta && \text{too little divergence for drift} \end{aligned} \tag{12.20e}$$

Thus, the hypothesis of drift is rejected (at $\alpha = 0.05$) if the mutational heritability is too small to account for the observed divergence

$$h_m^2(fast) < 0.10 \cdot \Delta = 0.10\frac{d_*^2}{t} \tag{12.21a}$$

Just as a smaller N_e allows for more divergence (and hence we set a critical *upper* value in Equation 12.18a), so does a larger mutational heritability h_m^2, and we set a critical *lower* value. Above this value drift could account for the observed divergence. Given that a typical upper-range value is $h_m^2 = 5 \times 10^{-2}$ (LW Table 12.1), if $0.20 \cdot \Delta > 5 \times 10^{-2}$, there is too much divergence to reasonably be accounted for by drift.

Conversely, the divergence is too slow to be accounted for by drift if the assumed mutational heritability is greater than the critical value

$$h_m^2(slow) \geq 509 \cdot \Delta \simeq 509 \cdot \frac{d_*^2}{t} \tag{12.21b}$$

A mutational heritabilty above this value would lead to significantly more divergence than observed. One important caveat for such tests of stabilizing selection is that laboratory–based estimates of h_m^2 are measured under very small effective population sizes and hence most deleterious pleiotropic effects are likely effectively neutral in these settings (Chapter 7). As the population size increases, the actual usable amount of polygenic mutational variance in a trait is likely considerably less than the laboratory estimates, perhaps by orders of magnitude. Hence, observations of stabilizing selection based on the perceived polygenic mutation rate being to high to generate so little divergence may be very biased and this test should be used with caution.

Example 12.4. Let's return to Reyment's foraminifer data from Example 12.3. Using the original Lande model, we rejected the hypothesis that drift could have accounted for the divergence. Applying Equation 12.21a, the hypothesis of drift accounting for excessive divergence is not rejected when

$$h_m^2 > 0.10 \cdot \frac{1.49^2}{5 \times 10^5} = 4.4 \times 10^{-7}$$

Turelli et al. note that this is several orders of magnitude lower than typical values of the scaled mutation variance, and thus this pattern of divergence is not too excessive for drift.

Which of the two variance assumptions (Equation 12.16a or 12.16b) should one use in a test? In our view, the constant variance assumption (Equation 12.16a) is less problematic, as the usable amount of σ_m^2 and h_m^2 may *decrease* with N_e. In such cases, $2N_e\sigma_m^2/N_e$ may *not* be a constant over N_e, greatly complicating tests based on critical mutational variances. Conversely, most trait heritabilities typically fall within a modest window of values, and one can vary the assumed value of h^2 to examine its consequences.

Ornstein-Uhlenbeck Models

As developed in Appendix 1, the Ornstein-Uhlenbeck (OU) process provides a model of Brownian motion drift coupled with a resorting force back to some optimal value θ, as might be expected with drift and stabilizing selection. Under the OU model, the expected change in the mean value of a process at value x is $a(\theta - x)$, so that if $x < \theta$, it increases, while for $x > \theta$ it decreases. The parameter a which measures the strength of the restoring force is also a measure of the strength of stabilizing selection. Under the standard model of Gaussian stabilizing selection (Equation 16.16), where ω^2 measures the strength of selection (smaller equals strong selection), Example A1.13 gives

$$a = \frac{\sigma_A^2}{\sigma_z^2 + \omega^2} \tag{12.22a}$$

As with Brownian motion, the value of the process at time t is normally distributed (Equation A1.33b), but now with mean and variance

$$\mu_t = \theta + [x_o - \theta]\exp(-at) \tag{12.22b}$$

$$\sigma_t^2 = \frac{b}{2a}[1 - \exp(-2at)] \tag{12.22c}$$

where $b = \sigma_A^2/N_e$ under the constant variance model. For large t, the mean value approaches the optimal value θ while the divergence variance saturates at a value $b/(2a)$, giving an asymptotic variance in the divergence of means as

$$\frac{b}{2a} = \frac{\sigma_z^2 + \omega^2}{2N_e} \tag{12.22d}$$

Ornstein-Uhlenbeck processes have been used to model the divergence of traits over a phylogeny (Felsenstein 1988, 2004, 2008; Garland et al. 1993; Martins 1994; Hansen and Martins 1996; Martins and Hansen 1997; Hansen 1997; Butler and King 2004; Beaulieu et al. 2012), including gene expression data (Bedford and Hartl 2009; Kalinka et al. 2010; Perry et al. 2012; Rohlfs et al. 2013). A key observation is that while the between-population divergence increases linearly with divergence time t under pure drift (e.g., Equation 12.10), Equation 12.22c shows that while initially the between-lineage variance increases linearly, it will asymptote out at a fixed level of divergence after sufficient time under an OU model.

Divergence in Morphological Traits

Numerous attempts have been made to apply the above procedures, or variants of them, to data from the fossil record to test the hypothesis that levels of morphological divergence over geological time scales have been driven by directional selection. For example, in the first of such studies, Lande (1976) showed that change in tooth-size dimensions over a 42 million year period in early horse evolution are consistent with the hypothesis of random genetic drift if the heritabilities of the traits had been near 0.5 and the long-term effective population size was smaller than 60,000 or so individuals. Given the generally high levels of heritability observed for mammalian morphological traits (Lynch and Walsh 1998), an assumption of $h^2 = 0.5$ is not unreasonable, and the argument that the long-term N_e in such lineages could be smaller than the critical value $N_e^* = 60,000$ is also plausible (Chapter 4), which would imply that drift could have acted alone to cause the observed changes. Analyses of tooth morphometrics in two additional lineages of extinct mammals (condylarths and oreodonts) suggest critical effective sizes of 80,000 to 120,000, below which the observed changes would be compatible with a neutral hypothesis. Thus, only if the effective sizes of these ancient mammalian taxa were actually in excess of 10^5, a matter that remains unclear, would the observed changes require some mechanism of directional selection.

Several other studies of this nature have been applied to aspects of mammalian skull evolution. For example, by taking the upper and lower limits to mutational heritability, σ_m^2/σ_e^2, to be 10^{-2} and 10^{-4}, Lynch (1990) found that the rates of evolution of cranial morphology in a wide array of placental mammalian lineages are one to two orders of magnitude below the minimum neutral rate, and Lemos et al. (2001) observed a similar pattern in marsupials. The only exception to this general trend concerns the races of modern man, which appear to have diverged at a rate slightly above the minimum neutral expectation (Lynch 1990; Ackermann and Cheverud 2004; Roseman 2004). Although they leave many questions unanswered, these kinds of results put in perspective previous arguments that rates of morphological evolution are exceptionally high in mammals, and especially so in the great apes (e.g., Cherry et al. 1982; Wyles et al. 1983; Van Valen 1985). Clearly, the predominant mode of evolution in mammalian skeletal morphology has been one of stabilizing selection, not of strong diversifying selection. Similarly, Spicer (1993) found widespread evidence of stabilizing selection on a variety of morphological traits in *Drosophila*, but some caution is in order here as the tests were based on critical mutation variances (Equation 12.21b). As mentioned, this approach likely generates many spurious calls of too little divergence, and hence spurious calls of stabilizing selection.

TIME SERIES DIVERGENCE TESTS

The methods discussed above are based on *external comparisons*. One has information from two time points and examines whether the observed amount of divergence is consistent with some external information, such as (estimated/assumed) effective population size N_e, time of divergence t, and some measure of genetic variation (h^2, σ_m^2). Conversely, starting with Raup (1977; Raup and Crick 1981), a number of methods have been proposed (largely from paleontology) to look detect departures from drift based entirely on the *internal* characteristics of a time series of morphology over time. The sequence of trait data from a set of subsamples within a stratigraphic column has been called a **stratophentic series** (Gingerich 1974). This is not a new enterprise. Almost a century ago, Ronald Brinkmann (1929) published morphological information on close to 3,000 ammonites in the genus *Kosmoceras* based on a centimeter level of resolution while working through a 14-meter section from the Middle Jurassic. The number of available stratophentic series is considerable, and growing.

A recent review by Hunt et al. (2015) examined 709 such series from roughly 200 lineages. The number of samples per sequence (populations measured from different locations within the same stratigraphic column) ranged from 7 to 114 (with a median of 14), covering an approximate time range from 5000 years to more than 50 million years, with most between 10^5 and 10^7 years in duration. Such a temporal series of data makes it possible to look for statistical trends in mean phenotypes or for correlations in rates of change in adjacent intervals, neither of which are expected in a strictly neutral model (at least under the infinite-alleles model).

The motivation for many of these tests was to provide a statistical framework to examine fossil data in the context of the **punctuated equilibrium** debate, wherein the suggestion was made that directional selection is rare in the fossil record, and a pattern of **stasis** (very little change) is the predominate mode (Eldredge and Gould 1972; Gould and Eldredge 1977).

Tests for Departures From Symmetric Random Walks

A number of relatively simple tests of depature from a symmetric random walk have been proposed. All assume uncorrelated changes over time increments and symmetry (both an equal chance of positive and negative increments and that the mean incremental change is zero in the event that it varies in magnitude). Raup (1977) and Raup and Crick (1981) proposed using a simple runs test (null hypothesis of an equal number of positive and negative changes). A sequence showing excessive increments of the same sign is consistent with directional selection, while a sequence with an excessive number of sign reversals is consistent with stabilizing selection (as might be expected for a population mean fluctuating around an optimum).

A more sophisticated approach was taken by Bookstein (1987, 1988), who used results on the maximum excursion of a symmetric random walk. Scale the sequence so that its initial mean is zero, and let S_i denote the mean of the ith sample. The standard error for the expected divergence over the entire sequence (n steps) of a symmetric random walk is $\sigma\sqrt{n}$, where σ^2 is the variance in change per increment. This standard error can be estimated as

$$\widehat{\sigma\sqrt{n}} = S = \sqrt{\sum_i (S_i - S_{i-1})^2} \tag{12.23}$$

For large n, Bookstein obtain an expression for the distribution of the largest scaled excursion $x = \max | S_i | /S$. For $x > 1$, critical values (the upper p of a tail) are very closely given by the correspond $p/4$ critical values for a unit normal. For example, the upper 5% tail correspond to $x = 2.25$, while the value for $p/4 = 0.125$ for a normal is 2.15. The upper 1% and 0.1% upper tail probabilities correspond to x values of 2.8 and 3.5. Series with values exceeding these critical values are said to be **improbably directional**, consistent with directional selection (or an environmental trend). Conversely, some series may not vary enough, and these are said to be **improbably constrained**, consistent with some sort of stabilizing selection or other cause of stasis. The lower 5%, 1%, and 0.1% values correspond to x values of 0.62, 0.49, and 0.41.

A final widely used test against the null of a symmetric random walk is based on **Hurst exponents** (Hurst 1951), also called **scaled range analysis**. The idea behind this approach is that the absolute difference of a symmetric random walk $| x_t - x_o |$ scales as $\sigma\sqrt{t}$. One regresses the standardized range R for a time interval τ on every-increasing values of τ, fitting the log-log regression

$$\ln(R) = H \ln(\tau) + \epsilon \tag{12.24}$$

where the slope H is the Hurst exponent (i.e., $R \propto \tau^H$). Under a symmetric random walk (increments are uncorrelated), absolute trait divergence scales with the square root of time,

giving $H = 0.5$. As increments become more positively correlated, H increases to one (**directional persistence**), consistent with directional selection. As adjacent increments become increasingly negatively correlated, H decreases to zero (**anti-persistence**), consistent with stabilizing selection or some other form of stasis. Roopnarine (2001) discusses permutation tests for significance of $H \neq 0.5$. Gingerich's (1993) **LRI** (log rate versus log interval) method is just a scaled version of this test, where slope G of his LRI regression is simply $G = H - 1$ (Roopnarine et al. 1999).

While straightforward and widely applied in the early literature, the problems with the above methods is that they typically have low power, so that the null hypothesis of a symmetric random walk is hard to reject (Roopnarine et al. 1999; Roopnarine 2001; Sheets and Mitchell 2001). This is especially the case with stratophentic series, with its incompleteness and sporadic coverage due to the vagaries of the fossilization process.

Hunt's Approach for Comparing Different Models

Hunt (2006, 2007, 2008a, 2008b; Hunt and Carrano 2010) notes that low power for tests of departures from symmetric random walks creates a "tyranny of the null hypothesis," potentially over-inflating the role of drift. He suggested that instead of testing against the random null, one should examine a set of candidate models, using Akaike weights (Anderson et al. 2000; Example 12.5) to indicate support for each. The Akaike weights for a set of competing models sum to one, providing a useful indicator of their relative support.

Table 12.1. Summary of the 251 fossil sequences examined by Hunt (2007), each fit using three models of divergence: random walk, directional selection, and stasis. Counts given under the Trait and Fossil group categories are the number of times a model had the highest Akaike weight for a fossil sequence. For example, 13 of the 251 sequences (0.052) had directional selection as the model with the highest support, while 5 of 114 (0.044) size-related traits had directional selection as the most-supported model. Values under the Median column correspond to the median fraction of support over all sequences for a given model. Half of all sequences had a support for directional change model of 0.06 or less, with 95% of all models having the fractional support for directional selection in the 0.02 to 0.08 range. The fossil groups are planktonic and benthic microfossils (Plank and Benth) and macrofossils (Marco).

			Trait				Fossil group		
Model	Median, 95% CI		All	Size	Shape	Other	Plank	Benth	Macro
Directional	0.06	(0.02, 0.08)	13	5	4	4	5	3	5
Random	0.47	(0.39, 0.56)	123	67	43	13	24	57	42
Stasis	0.34	(0.20, 0.50)	115	42	68	5	12	37	66
			251	114	115	22	40	97	113

Hunt initially consider three basic models: a symmetric random walk (incremental mean value zero), a directional (or generalized) random walk (mean increment $\neq 0$), and stasis. For the two random walks, he modeled the incremental δ (the change over an interval) by normal random variables. For the symmetric random walk, $\delta \sim N(0, \sigma_\delta^2)$, while for the general random walk, $\delta \sim N(\mu_\delta, \sigma_\delta^2)$, with parameters fit by maximum likelihood. For stasis, he assumed a simple model initially suggested by Sheets and Mitchell (2001). Instead of constructing likelihood models around increments (so that $z_t = z_0 + \sum \delta_i$), they simply took the trait value at time t to be $z_t \sim N(\mu, \sigma^2)$. While at first glance this appears to be an OU process at stationarity ($t \rightarrow \infty$), this is not the case, as Ornstein-Uhlenbeck has correlated

means (reflecting the shared history on a common evolutionary path). Akaike weights allow for a much more nuanced interpretation of model fit (see Figure 12.5). For example, suppose the weights for a fossil sequence are 0.50, 0.48, and 0.02 for the random, directional, and stasis models. Clearly, stasis and random walk are almost equally likely explanations, but the support for a directional model is very slim.

It should be stressed that the power of this approach is not the initial small set of candidate models chosen by Hunt, but rather that it serves as a much more general framework for examining an ever-richer set of models. Indeed, Hunt et al. (2015) examined more complex models that allow for the sequence to shift between modes (e.g., random vs. stasis) over the time sampled. One could also use this strategy contrast the simple stasis model used here with Ornstein-Uhlenbeck, Estes and Arnold (2006), or Uyeda et al (2011) as competing models of stabilizing selection or with Charlesworth's (1984) model of directional evolution.

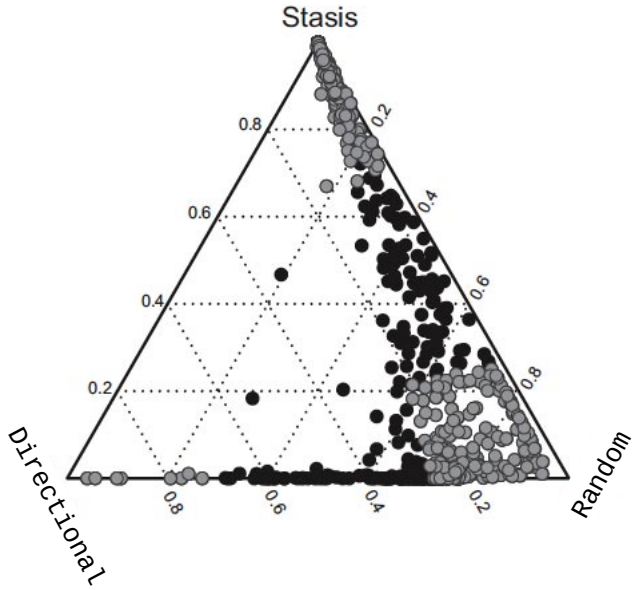

Figure 12.5. A De Finetti diagram of the support for the random walk, directional walk, and stasis models. Each point corresponds to the coordinates of the Akaike weights for these three models for a single stratophentic series. Points near vertices corresponds to almost 100% support. Points along an edge of the triangle indicate very little support in the trait perpendicular to that edge. Points in gray indicate strong support (weight for most supported at least 2.7 the weight of any other model). After Hopkins and Lidgard (2012).

Using this approach, Hunt (2007), Hopkins and Lidgard (2012), and Hunt et al. (2015) examined a growing number of stratophentic series (251, 635, and 709, respectively). The basic conclusions from Hunt et al. (2007) hold for the larger studies as well, and these are given in Table 12.1. For each sequence, the fraction of support for the three models was computed using Akaike weights. In only 13/251 (5.2%) of the sequences was directional selection (generalized random walk with $\mu_\delta \neq 0$) the most strongly-supported model (largest Akaike weight). The random walk model was the most supported (49%), while 46% of the fossil series had stasis as the most supported model. Figure 12.5 presents the relative support for all three models from the analysis of Hopkins and Lidgard (2012), presented as a

De Finetti diagram (or **triangle**). Values in the middle of the triangle have roughly equal support for all three models (rarely seen). Values near the edges of the triangle had very weak support for at least one model, and values near the vertices correspond to very high support for a single model. Note that there is little support along the directional selection axis for any sequence, with most of the support lying along the stasis-random walk axes.

Example 12.5. Anderson et al. (2000) proposed that a series of candidate models be comparied via their **Akaike weights**. Suppose one has a series of models, which are fit using maximum likelihood. If these are not nested, then comparison statistics, such as Akaike information criterion (or AIC; Akaike 1974) can be used to rank them (smaller is better), which reward goodness of fitness (highest log-likelihood L), while penalizing for number of parameters k. With $n < 40k$ observations, Anderson et al. suggest that a corrected version of AIC be used,

$$AIC_c = -2\ln(L) + 2k + \frac{2k(k+1)}{n-k-1} \tag{12.25a}$$

Suppose that one has a set of m candidate models, and compute the support for model i relative to the best-fitting of all m models,

$$\Delta_i = AIC_{c,i} - \min(AIC_c) \tag{12.25b}$$

The resulting Akaike weights for each of the candidate models are given by

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{j=1}^{m}\exp(-\Delta_j/2)} \tag{12.25c}$$

By construction, these sum on one, and give the relative support for model i given the collection of proposed models. Note that Equation 12.25c gives each model equal prior weight. More generally, if one has prior weights π_i for each model,

$$w_i = \frac{\pi_i\exp(-\Delta_i/2)}{\sum_{j=1}^{m}\pi_j\exp(-\Delta_j/2)} \tag{12.25d}$$

An interesting perspective on the rates of macroevolution was offered by Uyeda et al. (2011), who examined a vast data set of divergent traits, whose time span runs from fractions of years to over 350 million years. Their rather striking finding is summarized in Figure 12.6. For periods of a million years or less, rapid evolution could occur, but it was constrained, and did not accumulate over time. This matches the earlier observation by Estes and Arnold (2007) that the expected magnitude of divergence over samples was largely time-independent up to about a million generations. However, around a million years, Uyeda et al. observed rapid divergence, generating that they called a **blunderbuss pattern**, as the spread of values resembles the flared muzzle of the 17th century firearm of the same name.

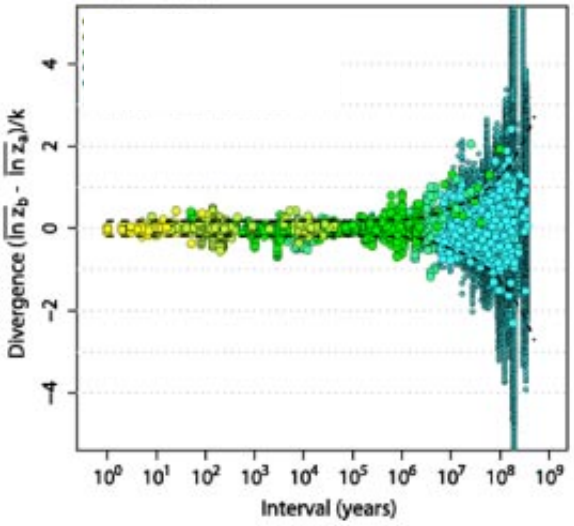

Figure 12.6. The blunderbuss pattern of divergence observed by Uyeda et al. (2011). Bounded variation is seen over the first 10^6 years, after which considerable divergence can occur. Divergence is scored as the log difference between means, scaled by the dimension k of the data ($k = 2$ for area, $k = 3$ for mass, etc.). After Uyeda et al. (2011).

While Estes and Arnold (2007) were able to account for their pattern by stabilizing selection with a fluctuating optimum, this model did not fit the data presented in Figure 12.6. Rather, the best fit was a model of essentially stasis (the Sheets-Mitchell model used by Hunt) coupled with rare random bursts of significant change (average waiting time around ten million years). The model allowing for multiple (as opposed to a single) burst fit the data best. While this pattern is striking and reproducible over the several different taxonomic data sets used by the authors, the underlying mechanism is unclear. Uyeda et al. suggest that it might correlated with the opening of new niches following species turnover (as species live spans are typically in the million-year range).

POPULATION-STRUCTURE BASED TESTS: Q_{ST} VERSUS F_{ST}

Species are often distributed in space as a series of subpopulations isolated by semipermeable migration barriers. In such settings, the variation at a neutral locus represent a balance between input of new variation by mutation and migration, countered by its removal by drift (Chapter 2). One measure of the resulting genetic population structure is Wrights' (1951) F_{ST} statistic, reviewed by Holsinger and Weir (2009). As discussed in Chapter 2, F_{ST} partitions the total genetic variance of the entire metapopulation (measured as heterozygosity under the assumption of panmixia) into the fraction within ($1 - F_{ST}$) and among (F_{ST}) subpopulations (Cockerham 1973; Nei 1987; Weir 1996). As developed in Chapter 9, one signature of selection at a candidate locus is whether it shows too much, or too, little population structure with respect to the F_{ST} value estimated from a set of (presumably) neutral markers. This is the test of Lewontin and Krakauer (1973). An excessive F_{ST} value at the candidate locus suggest too much divergence relative to the drift-migration expectation, and thus directional selection that varies over subpopulations. Too little divergence (F_{ST}

too small) suggests consistent stabilizing selection over subpopulations to retard the divergent effects of drift. Although there are a number of problems with such tests (Chapter 9), they still have a wide appeal, as the required data is reasonably straightforward to collect. The *trait*-based analog of this test is based on $Q_{ST} - F_{ST}$ comparisons. Q_{ST} (developed shortly) is a measure of the genetic population structure underlying our focal trait, which is compared with a neutral-marked based estimate of F_{ST} for the population. While the candidate-gene tests in Chapter 9 compare the candidate locus F_{ST} value with the neutral standard, the *trait*-based test compares Q_{ST} for the candidate trait with this standard. This is a rather active area of research, and reviews can be found Merilä and Crnokrak (2001), McKay and Latta (2002), Whitlock (2008), and Leinonen et al. (2008, 2013). Standard Q_{ST} analysis assumes only a single level of structure, but extensions to more hierarchically structured populations were started by Whitlock and Gilbert (2012). Likewise, multivariate extensions (properly accounting for the genetic correlations among a series of measured traits) are developed by Chenoweth and Blows (2008), Martin et al. (2008), Chapuis et al. (2008), and Ovaskainen et al. (2011).

$\mathbf{Q_{ST}}$: Partitioning Additive Variance Over Subpopulations

Consider a quantitative trait with a purely additive-genetic basis, and denote genetic variance for the trait in the entire metapopulation under the assumption of panmixia by σ^2_G. From Table 11.3 (taking $f = Q_{ST}$), the within- and among-subpopulation components of variance can be represented as $\sigma^2_{GW} = (1 - Q_{ST})\sigma^2_G$ and $\sigma^2_{GB} = 2Q_{ST}\sigma^2_G$, respectively, for a total variance in a structured population of $(1 + Q_{ST})\sigma^2_G$. Rearranging yields

$$Q_{ST} = \frac{\sigma^2_{GB}}{\sigma^2_{GB} + 2\sigma^2_{GW}} \tag{12.26}$$

While the term Q_{ST} is do to Spitze (1993), this metric was proposed earlier by Prout and Barker (1989, 1993) and Lande (1992), and strongly hinted at by Rogers and Harpending (1983). Although Equation 12.26 is obtained using the results for a collection of partly inbred lines (i.e., Table 11.3), this is a very general result, applying to a wide range of population structures and migration patterns provided the character does indeed have an entirely additive genetic basis (Whitlock 1999). When dealing with haploid populations (Whitlock 2008) or collections of entirely selfed lines (Bonnin et al. 1996; Le Corre 2005; Rhoné et al. 2010), σ^2_{GW} in Equation 12.26 is now weighted by one, rather than two.

Table 12.2. Interpretation of Q_{ST} versus F_{ST} comparisons.

Observation	Interpretation
$Q_{ST} > F_{ST}$	**Divergent selection**: spatial variation in favored trait values
$Q_{ST} = F_{ST}$	Consistent with divergence expected under drift. Does not rule out selection, but does not support it either.
$Q_{ST} < F_{ST}$	**Convergent selection**: similar trait values favored over subpopulations.

Equation 12.26 provides a potential approach for testing the hypothesis of neutral divergence among subpopulation means. If isolates of sufficient numbers of families from multiple subpopulations can be grown in a common environment, then appropriate statistical methods can be used to estimate σ^2_{GW} and σ^2_{GB} (Lynch and Walsh 1998). The resultant

estimate of Q_{ST} can then be compared to a parallel measure of subdivision (F_{ST}) derived from putatively neutral markers. Under the assumption of neutrality, Q_{ST} should not be significantly different from F_{ST}. $Q_{ST} > F_{ST}$ is expected if subpopulation differentiation has been primarily driven by adaptive divergence, whereas the opposite pattern is expected if the mean phenotypes of all or most subpopulations are kept relatively uniform by stabilizing selection for the same optima (Table 12.2). It is important to stress that any $Q_{ST} - F_{ST}$ comparison *must* be performed using the same set of populations to obtain both Q_{ST} and F_{ST}. An analysis using an estimate of F_{ST} from one set of populations and Q_{ST} from another is not trustworthy.

One caveat with respect to this strategy is that, even under neutrality, the expected value of Q_{ST} will not necessarily equal F_{ST} if the trait of interest is influenced by nonadditive genetic effects. As outlined in Chapter 11 and above, with nonadditive gene action, the within- and among-subpopulation components of genetic variation for neutral characters under short-term divergence are no longer equal to $\sigma^2_{GW} = (1 - f)\sigma^2_G$ and $\sigma^2_{GB} = 2f\sigma^2_G$ (where f is the parameter estimated by F_{ST}), but instead are influenced by a number of higher-order terms (see Table 11.3). In general, because the within-population genetic variance declines less rapidly with inbreeding under nonadditivity (and sometimes even increases; Chapter 11), Q_{ST} as defined by Equation 12.26 will tend to be smaller than F_{ST} under neutrality, although exceptions do exist. In particular, Whitlock (1999) showed that additive $\times$ additive variance always results in $Q_{ST} < F_{ST}$ under neutrality. Dominance also causes Q_{ST} and F_{ST} to deviate under neutrality, with the directional of the inequality a function of the population structure. There is disagreement as to the practical importance of these departures, especially given the large variances associated with Q_{ST} estimates (Goudet and Büchi 2006; Goudet and Martin 2007; López-Fanjul et al. 2003, 2006, 2007; Whitlock 2008). Since these often (but not always) result in $Q_{ST} < F_{ST}$, this general behavior makes conclusions regarding adaptive divergence based on elevated Q_{ST} conservative, while rendering observations of $Q_{ST} < F_{ST}$ ambiguous. The requirement of additivity may not be a serious issue for most morphological traits, but given that life-history traits often show considerable nonadditive variance (Chapter 6), these may be more vulnerable to false impressions under a Q_{ST}-F_{ST} comparison.

A second caveat is that the choice of markers used to estimate F_{ST} can introduce bias. The strong assumption is that the markers chosen are neutral, such that any structure within them reflects the neutral population structure. Historically, allozyme markers were commonly used. Since these represent variant protein products, a priori one must assume that some are not neutral, in which case their use introduces bias. The assumption leading to equilibrium values of F_{ST} is that mutation rates are much less than migration rates (Hendry 2002; Kronholm et al. 2010). Microsatellite markers are widely used, but these can have very high mutations rates (on the order of $1/200$ or larger), which can violate the low mutation rate assumption. A second concern with microsatellites is that for F_{ST} to serve as a neutral proxy for the behavior of alleles underlying a focal trait, the mutational structure of the markers and QTLs must be somewhat compatible. Microsatellites alleles (scored as the number of repeat copies) can easily back-mutate, resulting in underestimation of F_{ST}. While microsatellites-specific metrics (such as R_{ST}, Slatkin 1995a; Goodman 1997) have been proposed, these should *not* be used in place of F_{ST} for comparison with Q_{ST}. These adjust for high rates of back-mutations, something not expected at QTL alleles, resulting in different adjusted measures of allelic divergence at the markers versus QTLs. Take together, these issues suggest that the use of microsatellites in F_{ST} vs. Q_{ST} comparisons is problematic (Edelaar and Björklund 2011; Edelaar et al. 2011), which is a serious issue given that the majority of recent studies typically have used these markers. The ever-increasing use of

SNPs avoids these concerns.

P_{ST}: Approximating Q_{ST} with Phenotypic Data

Because of the requirement for assays in a common-garden arena, true joint studies of Q_{ST} and F_{ST} are not common. Pujol et al. (2008) noted that roughly half of the wild population studies they reviewed were not based on estimated additive variances. Instead, a phenotypic-based proxy for Q_{ST} was used, where within- and/or between-population *phenotypic* variances replace the more challenging estimates of additive variation. The former can easily be obtained via a standard ANOVA (e.g., Holand et al. 2011), while the latter requires a series of parent-offspring or sib rearings in a common environment. A modification of this purely phenotypic approach is to use

$$\widehat{\sigma}^2_{GB} = c\,\widehat{\sigma}^2_{PB}, \qquad \widehat{\sigma}^2_{GW} = h^2\,\widehat{\sigma}^2_{PW} \tag{12.27a}$$

where c reflects the fact that only part of an observed phenotypic difference in means may be genetic (Merilä 1997; Leinonen et al. 2006; Sæther et al. 2007; Brommer 2011). Substitution into Equation 12.26 yields the P_{ST} statistic of Leinonen et al. (2006),

$$\widehat{P}_{ST,L} = \frac{\widehat{\sigma}^2_{PB}}{2(h^2/c)\,\widehat{\sigma}^2_{PW} + \widehat{\sigma}^2_{PB}} \tag{12.27b}$$

When $c = h^2$, this reduces to Equation 12.26, with phenotypic variances replacing their genetic counterparts. Holand et al. (2011) suggest doing a sensitivity analysis by varying the value of c, and using simulations to find critical upper and lower c values where Q_{ST} is significantly above and significantly below F_{ST}. While enticing because of their simplicity and relative ease of application (only phenotypic data required), strong caution is advised when replacing Q_{ST} by a phenotypic surrogate (Pujol et al. 2008; Brommer 2011). At a minimum, such estimates should *always* be denoted as P_{ST} whenever *any* variance component is based on a purely phenotypic measure. The biology/ecology of a species might be such that only P_{ST} estimates are possible, but in such cases the investigator needs to seriously consider if such a resulting study can give truly worthwhile results.

Even when genetic data (information from crosses) is used, bias can still be introduced into Q_{ST} estimates. Ideally, additive variances should be estimated from the covariance among paternal half-sibs. When covariance among full-sibs is used, additive variance estimates can be inflated by the presence of dominance and/or maternal effects. Likewise, if between-group differences are not measured in a common garden, shared environmental effects can inflate this estimate. Conversely, a common garden may obscure any evolved plastic response that is part of the adaptive response to specific environments. Following López-Fanjul et al. (2003), write $\sigma^2_{GB} = 2F_{ST}(1+a)\sigma^2_A$ and $\sigma^2_{GW} = (1-F_{ST})(1+b)\sigma^2_A$, where a and b measure the bias. Hence,

$$Q_{ST} = \frac{2F_{ST}(1+a)\sigma^2_A}{2F_{ST}(1+a)\sigma^2_A + 2(1-F_{ST})(1+b)\sigma^2_A} = \frac{(1+a)F_{ST}}{(1+a)F_{ST} + (1-b)(1-F_{ST})} \tag{12.28}$$

which only equals F_{ST} when $a = b$. When $a > b$, $Q_{ST} > F_{ST}$, while for $a < b$, $Q_{ST} < F_{ST}$. See Whitlock (2008) for further discussion of such sources of bias.

A slightly optimistic note was struck by Pujol et al. (2008), who note that the onerous requirement of a common garden may be partly circumvented through the use of genetic-group mixed models from dairy science (e.g., Westell et al. 1988; Quaas 1988), which allow for estimation of both within- and between-group additive variance (a variant of this approach was used by Roberge et al. 2007). A limitation on this approach in natural populations

is that information on the relationships between groups is required to separate of environmental from genetic effects (Chapter 20). While this can be estimated from marker data, this requires a rather dense set of markers (a very large number of SNPs) for the accuracy needed given the expected rather distant connections between groups.

Testing Q_{ST} Versus F_{ST}

The construction of rigorous statistic tests for comparing Q_{ST} with F_{ST} is problematic on several levels. First, both are ratios of variances, so that estimates obtained by directly plugging variance estimates into Equation 12.26 are biased, as the expectation of a ratio is not the same as a ratio of expectations (LW Appendix 1). Second, the sampling distribution of Q_{ST} is complex, as one must use a crossing design to estimate the variance components. Hence, the correct construction of dispersion intervals (such as standard errors in a frequentist setting or credible intervals in a Bayesian setting) is not trivial. Finally, there is the issue of formally comparing a somewhat noisy estimate (F_{ST}) with a very noisy estimate (Q_{ST}), which are obtained using very different designs. Some of these issues are addressed by O'Hara and Merilä (2005), Whitlock (2008), and Whitlock and Guillaume (2009). As noted by O'Hara and Merilä, one significant problem is simply power. The between-group variance is a function of the number of groups, with at least twenty needed for any substantial power. Unfortunately, the typical group was around seven for the studies reviewed by Merilä and Crnokrak (2001).

An importance advance was the observation by Whitlock (2008) that the distribution of Q_{ST} (ignoring, for now, its sampling variance), can often be approximated using the Lewontin-Krakauer distribution for F_{ST} values (Equation 9.8). Provided that F_{ST} is small, simulations by Whitlock confirmed the suggestion by Rogers and Harpending (1983) that the amount of information on population structure from the variance components of a quantitative trait is equivalent to the information from a single-marker F_{ST}. When the average F_{ST} is small, then often to a very good approximation

$$\frac{n_d-1}{\overline{F}_{ST}}\,Q_{ST} \sim \chi^2_{n_d-1}, \quad \text{implying} \quad Q_{ST} \sim \frac{\overline{F}_{ST}}{n_d-1}\,\chi^2_{n_d-1} \tag{12.29a}$$

where $\overline{F}_{ST}$ is the average F_{ST}. This expression assumes that Q_{ST} is estimated without error, a point addressed shortly. The requirement that $\overline{F}_{ST}$ is small arises (in part) from χ^2 being defined over $(0,\infty)$, while Q_{ST} is defined on $(0,1)$, so very little of the χ^2_2 tail must exceed one for this to be a reasonable approximation (Whitlock 2008 recommends an upper limit of $\overline{F}_{ST} < 0.1$). Rearranging Equation 12.29a gives

$$\Pr\left(\frac{Q_{ST}}{\overline{F}_{ST}} > \delta\right) = \Pr\left(\frac{(n_d-1)Q_{ST}}{\overline{F}_{ST}} > \delta(n_d-1)\right) = \Pr\left(\chi^2_{n_d-1} > \delta(n_d-1)\right) \tag{12.29b}$$

Consider $n_d = 2$, such as occurs when comparing two subspecies. How much larger must Q_{ST} be over F_{ST} for this difference to be significant at the $\alpha = 0.05$ level? Since tests involving Q_{ST} are two-sided (being either too large or too small is of interest), and $\Pr(\chi^2_1 > 5.02) = 0.025$, Q_{ST} must be in excess of 5.02 times $\overline{F}_{ST}$ to be significant at the 5% level. For $n = 10$, $\Pr(\chi^2_9 > 19.03) = 0.025$, or $\delta \cdot 9 = 19.03$, or $\delta = 2.1$ and hence only a two-fold difference is required for significance. The same logic can be used for the critical value when $Q_{ST} < F_{ST}$. For example, since $\Pr(\chi^2_9 < 2.7) = 0.025$, a value of Q_{ST} less than one third of $\overline{F}_{ST}$ (2.7/9 = 0.3) is significant at the five percent level when $n_d = 10$.

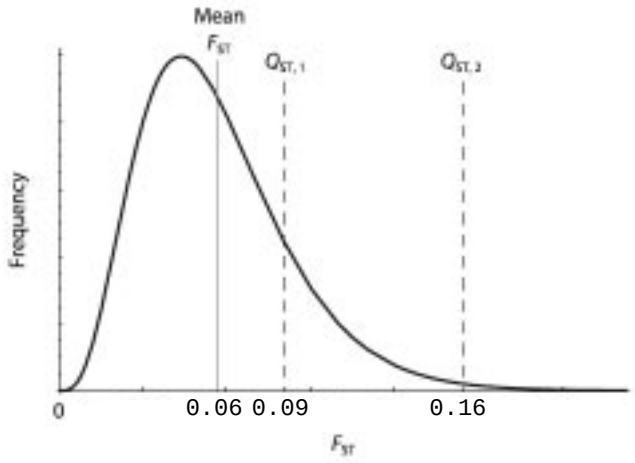

Figure 12.7. For small values of average F_{ST}, the Q_{ST} distribution for a neutral, completely additive trait should approximately following the Lewontin-Krakauer distribution (Equations 9.8, 12.29a). Two traits, one with an estimated $\widehat{Q}_{ST} = 0.125$, and a second with $\widehat{Q}_{ST} = 0.16$ are both larger than $\overline{F}_{ST} = 0.06$, but only trait two is significant. After Whitlock et al. (2008).

Figure 12.7 shows the basic structure of tests based on this simple approach: compute $\widehat{Q}_{ST}$ and compare this value with the distribution of values expected for single-locus F_{ST}, where the mean of this later distribution as taken as $\overline{F}_{ST}$, the mean F_{ST} value over all loci in the sample. This approach assumes that a *single* trait is all that is of interest, and that Q_{ST} is measured without error (again, we return to this). In the typical study setting, one has k Q_{ST} values (one for each trait in the study), but uses the same set of markers (and hence the same $\overline{F}_{ST}$ value) for all traits. Here one must compare the Q_{ST} values with the k order statistics drawn from this distribution. The significance of the largest Q_{ST} value is assessed by comparing it with the null distribution of the largest value generated by simulating k draws from the Lewontin-Krakauer distribution (Equation 12.29a). If this is significant, one can then turn to the second largest Q_{ST} value and compare it with the simulated distribution of the second largest order statistic, and so on.

The major flaw with using Equation 12.29a is that it ignores the very important sampling variances of both $\widehat{Q}_{ST}$ and $\overline{F}_{ST}$. Whitlock and Guillaume (2009), building on Equation 12.29a, show how to incorporate these to construct the distribution Q^n_{ST} of values under the assumption of an additive and neutral trait. They start with the clever observation that since $Q_{ST} = F_{ST}$ for a neutral additive trait, we can rearrange Equation 12.26 to give

$$\sigma^2_{GB} = \frac{2\,F_{ST}\,\sigma^2_{GW}}{1 - F_{ST}} \tag{12.30}$$

This trick allows us to generate a random value for σ^2_{GB}, given values of σ^2_{GW} and F_{ST}. Their scheme to generate the ith sample of Q^n_{ST} is as follows:

i) Draw a bootstrap sample of the n markers by sampling with replacement. Compute the mean F_{ST} value for this sample, $\overline{F}_{ST,i}$.

ii) Draw a sample value of $\sigma^2_{GW,i}$ as follows:

a) If σ^2_{GW} computed by a standard ANOVA, recall that $MS/df \sim \chi^2_{df}$, where MS is the observed mean sum of squares and df its degrees of freedom (LW Chapter 18). Generate X_i, a draw from a χ^2_{df} and take the sample i value for MS as $MS/df \cdot X_i$, using these to compute the variance estimates (O'Hara and Merilä 2005; Whitlock and Guillaume 2009). When σ^2_{GW} is estimated as a function of several different sums of squares, compute each in the fashion above, and then use these realizations to compute $\sigma^2_{GW,i}$.

b) If σ^2_{GW} was estimated by Bayesian methods, then $\sigma^2_{GW,i}$ is a sample drawn from the resulting posterior (i.e., Rogell et al. 2012).

iii) Use the sample values of $\overline{F}_{ST,i}$ and $\sigma^2_{GW,i}$ with Equation 12.30 to generate a draw of the between-group variance, $\sigma^2_{GB,i} = 2\overline{F}_{ST,i}\,\sigma^2_{GW,i}/(1-\overline{F}_{ST,i})$

iv) Take $Q^n_{ST,i} = \sigma^2_{GB,i}/(\sigma^2_{GB,i} + 2\sigma^2_{GW,i})$

A plot of the resulting distribution of $\widehat{Q}_{ST} - Q^n_{ST}$ provides a formal statistical test of whether the observed value $\widehat{Q}_{ST}$ is excessively large (95% credible interval is entirely above zero) or small (95% credible interval exclusively below zero). Figure 12.8 shows how the use of **violin plots** provides a useful way to display these results.

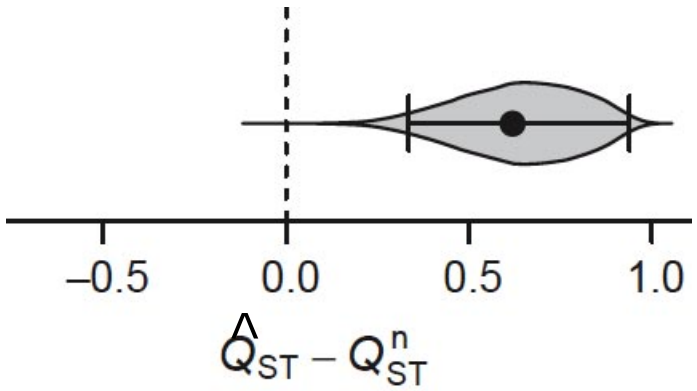

Figure 12.8. Violin plot for the distribution of $\widehat{Q}_{ST} - Q^n_{ST}$ for body length in the sea-run brown trout (*Salmo trutta*), using the resampling scheme suggested by Whitlock and Guillaume (2009). The width of the "violin" indicates the probability mass in that interval, the dot denotes the highest posterior probability, and the error bars the 95% credibility intervals. Here this interval is completely above zero, demonstrating that $\widehat{Q}_{ST}$ is significantly in excess of its predicted neutral value given $\overline{F}_{ST}$. After Rogell et al. (2012).

Empirical Data

Results from the large number of Q_{ST} vs. F_{ST} studies from natural populations are summarized by Merilä and Crnokrak (2001), McKay and Latta (2002), and Leinonen et al. (2008, 2013). Values of Q_{ST} and F_{ST} are positively correlated, with $\rho = 0.24$ (Leinonen et al. 2013). Thus, there is a modest tendency for the structure of quantitative trait variation to parallel the population structure for neutral alleles. The striking finding is that $Q_{ST} > F_{ST}$ for around 70% of all traits, suggesting that heterogeneous variation is directional selection is very widespread (Figure 12.9). Conversely, values of $Q_{ST} < F_{ST}$ are rare, despite that bias in this direction for neutral traits under a variety of conditions (discussed above), suggesting

that stabilizing or uniform selection is far less common. As the former result is qualitatively consistent with adaptive differentiation, they are clearly at variance with the observations on longer-term divergence noted above.

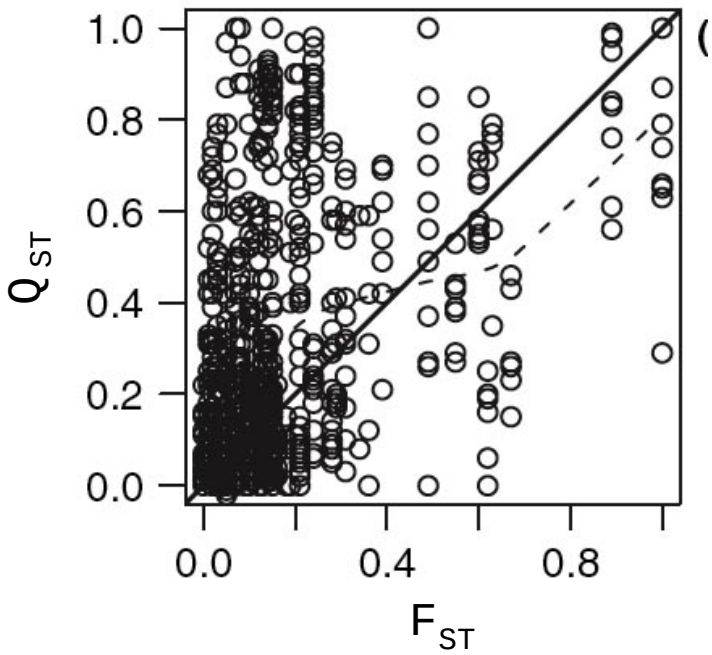

Figure 12.9. The joint distribution of $Q_{ST} - F_{ST}$ seen in the meta-analysis of Leinonen et al. (2008). The solid line represents the neutral expectation $Q_{ST} = F_{ST}$. Instead, there is a very strong tendency for $Q_{ST} > F_{ST}$. While this is consistent with widespread diversifying selection, as discussed in the text, it can also arise from ascertainment bias and using highly polymorphic markers (which underestimate F_{ST}).

One potential explain for this trend of $Q_{ST} > F_{ST}$ is offered by Miller et al. (2008). They found that the evolutionary variance of Q_{ST} is significantly larger than that for F_{ST}, noting a strong positive correlation between Q_{ST} and $Q_{ST} - F_{ST}$. Hence, populations with larger Q_{ST} values tend to also have greater departures from F_{ST}. In particular they note that if traits are **ascertained** (chosen) in a fashion that results in more variable traits being over-represented in the sampling process, this picks up outliers of Q_{ST} given its large evolutionary variance, and this in turn generates excessive $Q_{ST} - F_{ST} > 0$ values even under neutrality. Whitlock (2008) further stresses this concern:

> "it will always be possible to choose a set of traits that have higher than average Q_{ST} values. Traits chosen in this way cannot reliably be used to infer the extent of spatially heterogeneous selection. Examination of the traits chosen for many Q_{ST} studies makes one wonder whether traits are fact always chosen with previous knowledge of the likely results. "

A second source of bias was noted by Edelaar and Björklund (2011) and Edelaar et al. (2011). Markers with high mutation rates underestimate F_{ST}, and the most widely-used markers in $Q_{ST} - F_{ST}$ studies, microsatellites, have high mutation rates. Indeed, these loci are specifically chosen *because* they are so highly polymorphic, so that they contain maximal

information. As shown in Figure 12.10, there is a strong positive relationship between the polymorphism level of a marker (highly polymorphic markers having higher mutation rates) and the excess values of Q_{ST} over F_{ST}. Note that most of this trend is driven by microsatellites, with allozymes showing a consistent excess of Q_{ST} largely independent of their polymorphism level. The use of SNPs to estimate F_{ST} avoids these issues.

Thus, the striking trend of $Q_{ST} > F_{ST}$ is certainly greatly inflated by ascertainment bias, and somewhat inflated by the use of highly polymorphic markers (which is a more recent trend), making it difficult to make any general statement about how common diversity selection structures quantitative traits in subdivided populations. As noted by Whitlock (2008), "While useful, Q_{ST} is a crude measure of the genetic differential of a trait caused by local adaptation."

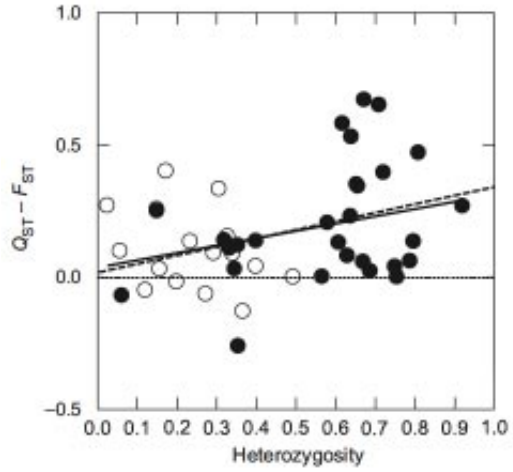

Figure 12.10. Correlation between $Q_{ST} - F_{ST}$ and the polymorphism level at the markers (a proxy for their mutation rate). Filled circles are microsatellites, open are allozymes. From Edelaaret al. (2011).

One check of theory is to compare Q_{ST} and F_{ST} values between control and artificially-selected groups. Morgan et al. (2005) examined the results of a 14-generation replicated selection experiment for increased voluntary wheel-running activity in mice. A base population was split into a control group, both with four replicate lines. The average selection intensity per generation was close to one, and significant response was seen in both the target trait, and (as a correlated response) in body mass. Theory predicts that Q_{ST} for the control versus selected group should exceed F_{ST}, and this was observed for both the directly selected (wheel running) and correlated (body mass) traits. Theory also predicts that Q_{ST} and F_{ST} should be similar *within* the selected lines, and this was indeed seen for wheel-running. Q_{ST} was actually below F_{ST} for body mass, but not significantly so. Support for Q_{ST} as a method for detecting selection was also offered by Rhoné et al. (2010). They examined the response to 12 generations of natural selection on flowering times for a synthetic population of wheat grown in three locations in France, with individuals from remnant seed assessed in a common garden. For generation two seed, Q_{ST} and F_{ST} were not significantly different. However, individuals from generations seven and twelve had Q_{ST} significantly larger than F_{ST}. Finally, agreement with theory was mixed in Porcher et al. (2004, 2006), who examined

eight generations of selection in a series of structured *Arabidopsis* populations. Larger Q_{ST} values were seen under heterogeneous selection, consistent with theory, but F_{ST} increased as well.

QTL-BASED TESTS

The preceding approaches rely on large samples from multiple populations or at two or more intervals. However, there is a situation in which one might test for adaptive divergence with just a single cross between two isolated lineages. By examining a battery of polymorphic markers segregating in the F_2 generation of such a cross, one may search for QTLs associated with phenotypic measures of various traits (Lynch and Walsh 1998). Chapters 9 and 10 discussed a wide variety of test for whether a particular *locus* is under selection, and that locus could be a QTL or GWAS hit from such a mapping experiment.

Here we examine a different class of tests based not on some signature from a *single* QTL, but rather on a signature from an *entire collection* of QTLs for a given trait. The basic idea traces back to three papers, all coincidentally examining crosses for male secondary traits involving *Drosophila simulans* (Coyne 1996; Laurie et al. 1997; True et al. 1997). Under the neutral hypothesis, the relative abundances of "plus" and "minus" marker alleles associated with large vs. small phenotypes are expected to be randomly distributed and thus should not differ significantly from a 1:1 ratio. Naively, this might suggest a simple **sign test**: are there an excessive number of pluses in one line? If so, this is not consistent with neutral drift (which is agnostic with respect to the direction and effect sizes of QTLs being fixed). This general strategy will be biased if the parental lines are intentionally chosen (ascertained) to have extreme phenotypes, as the high line would naturally be expected to be enriched with "plus" alleles. Orr (1998) suggested a way around this problem.

Orr's QTLST and QTLST-EE Sign Tests

Orr noted that by choosing to count the number of plus alleles in the larger line, we have introduced an ascertainment bias, as this line is *expected* to contain an excess of such alleles. We need some appropriate conditioning on this fact to obtain an unbiased statistic of what constitutes excess of plus alleles. The simplest approach is Orr's **equal effects model**, where all n QTL have close to equal effects. Here, the large line must contain at least $[n/2]$ high QTLs, where

$$[n/2] = \begin{cases} n/2+1 & \text{for } n \text{ even} \\ (n+1)/2 & \text{for } n \text{ odd} \end{cases}$$

In other words, the high line must contain *at least* one more high allele than the low line (since all have equal effects). What corresponds to an excessive number of observed plus alleles in the high line (n_{+obs}) now becomes a simple combinatorial problem. The probability of k high alleles in one line (under neutrality) follows from the binomial, where there is an equal chance that a random line gets a plus or a minus allele, giving

$$\Pr(n_+ = k) = \binom{n}{k}(1/2)^k(1/2)^{n-k} = \binom{n}{k}(1/2)^n$$

so that all values of k have $(1/2)^n$ in common, giving

$$\Pr(n_+ \geq n_{+obs} \,|\, n_+ \geq [n/2]) = \sum_{i=n_{+obs}}^{n} \binom{n}{i} \Big/ \sum_{j \geq [n/2]}^{n} \binom{n}{j} \tag{12.31a}$$

This is Orr's **QTL sign test for equal effects**, or **QTLST-EE**. Orr notes that a *minimum* of $n = 6$ detected QTLs is required for this test to be applied. To see this note for $n = 6$ that $[n/2] = 4$, and the most extreme value $n_{+obs} = 6$ gives a p value of $1/22 \sim 0.05$, while for $n = n_{+obs} = 5$, the smallest p is $1/16 \sim 0.0625$. For large n, Orr notes that Equation 12.31a can be approximated by a normal,

$$p \simeq 2\left[1 - \Phi\left(\frac{n_{+}obs - n/2}{\sqrt{n/4}}\right)\right] \tag{12.31b}$$

where $\Phi(x) = \Pr(U \leq x)$ for $U \sim N(0, 1)$.

Example 12.6. True et al. (1997) found that none of the eight detected QTL for the posterior lobe in the male genitalia in a *Drosophila* cross were **antagonistic**, alleles in the opposite direction of the line value, such as low (minus) alleles in the high line and high (plus) alleles in the low line. Orr suggested that the equal effects model may not be unreasonable for this trait. Assuming this, we have $n = 8$, $[n/2] = 5$, and $n_{+obs} = 8$. Equation 12.31a gives

$$p = \sum_{i=8}^{8}\binom{8}{i}\Big/\sum_{j\geq 5}^{8}\binom{8}{j} = \frac{\binom{8}{8}}{\binom{8}{5}+\binom{8}{6}+\binom{8}{7}+\binom{8}{8}} = \frac{1}{56+28+8+1} = 0.011$$

showing that this is a highly significant excess even for a high line. Orr's normal approximation (Equation 12.31b) gives

$$p \simeq 2\left[1 - \Phi\left(\frac{8-5}{\sqrt{8/4}}\right)\right] = 2\left[1 - \Phi\left(2.1213\right)\right] = 0.034$$

This is rather conservative, but not surprising as this is a large-sample approximation and the number of detected QTL here is very modest.

Under this strong assumption of equal effects, QTLST-EE is a nonparametric test, making no other assumptions and not using any information on the actual difference between the high and low lines. Anderson and Slatkin (2003) note that this test can be highly biased by trait ascertainment (where the investigator, often unconsciously, chooses traits showing excessive divergence, further biasing traits for an excess of plus alleles). They simulated T identical and independently distributed traits, and then chose the most divergent single trait from this set to perform a QTLST-EE. They found that this process of ascertainment introduces a significant bias. For example, their simulations had $n = 10$ QTLs, so that the probability of $n_{+obs} \geq 9$ for a random high line is 0.0285 (Equation 12.31a). However, when the high line was not randomly chosen, but was rather the most divergent in a set of 25 traits, then over 50% of the time it contained at least nine high alleles. This lack of robustness with respect to the trait ascertainment scheme means that significant QTLST-EE results must be interpreted with caution.

Orr's second approach avoids this problem, and indeed simulations show that it is conservative (Anderson and Slatkin 2003; Rice and Townsend 2012). Let R be the difference

between lines, either the actual observed difference, or the difference based on summing the effects over all detected QTLs. With a distribution of QTL effects in hand, one can then conditional on the number of plus alleles *given* that the high line is R in excess of the low line. This is Orr's **QTL sign test**, or **QTLST**. The seemingly problematic issue of the distribution of QTL effects can be easily handled via a bootstrap approach in one of two ways. First, one could use the observed distribution of absolute QTL effects ($|\,a\,|$), and then fit this using some standard distribution. Orr used the gamma (Equation A2.25, Figure A2.2) because of its flexibility and the fact that the exponential, a commonly-assumed distribution of effect sizes (Chapter 26), is a special case. This is best fit as a truncated distribution, as QTL effects below a critical absolute size would not be detected. One can then generate the p value for the observed number of plus alleles through parametric bootstrapping as follows. Generate n draws from this distribution, randomly assign each a sign, and only keep those samples for which the total G of signed QTL effects exceeds R. The resulting distribution of plus alleles in the high line is now conditioned on neutrality (signs drawn at random), the assumed distribution of QTL effects, and the actual divergence R, giving

$$p = \Pr(n_+ \geq n_{+obs} \,|\, G \geq R) = \sum_{i=n_{+obs}}^{n} \Pr(n_+ = i \,|\, G \geq R) \tag{12.32}$$

where $\Pr(n_+ = i \,|\, G \geq R)$ is simply the fraction of times $n_+ = i$ in the bootstrap sample. Alternatively, instead of sampling from the fitted distribution, one could use standard bootstrapping and directly sample (with replacement) from the observed distribution of QTL effects (e.g., Rice and Townsend 2012).

While the QTLST does adjust for ascertainment bias, it does so the expense of power. As noted by Rice and Townsend (2012), the difference R is some indication of the amount of previous selection, and by conditioning on its value, we are removing this evidence. Consider the extreme case where a line fixes all ten plus alleles, and all have equal effect. In order to obtain the observed value of R, we must condition on only those cases where all ten are fixed, giving this test zero power (Griswold and Whitlock 2003 also noted the low power of this test). Rice and Townsend found that both the power, and the false-positive rate, increase with variance of QTL effects.

Orr's tests have a simple appeal, but (as we have shown) there are a number a caveats with their application. Another restriction is that no significant epistasis is allowed, as both tests sum over single-locus effects. There are biological reasons for Orr's tests to be conservative beyond the statistical reasoning discussed above. A trait with a smaller fraction of antagonistic QTLs could imply either stronger selection or more temporally consistent selection. If the sign of directional selection changes on a trait over evolutionary time, then different bouts of selection fix alleles of opposite sign, lowering the observed excess. If the trait is under stabilizing selection and is at its optimal value, random fixations of small-effect alleles are equally likely to be positive or negative. Likewise, if a trait fixes a major allele that causes it to overshoot an optimal value, subsequent fixations of smaller effects in the corrective direction will also reduce any signature from this test. Despite these concerns, this is a useful approach, even if one treats it as nothing more than an exploratory device, rather than a formal test (which was our suggestion for most of the locus-specific tests discussed in Chapters 9 and 10). Further, although framed in QTLs, there is no reason while this could not be used with appropriate GWAS data as well. An example of the latter is Turchin et al. (2012), who showed that frequencies of plus height alleles are systematically elevated in Northern Europeans compared with Southern Europeans.

Applications of QTL Sign Tests

Using QTLST-EE, Rieseberg et al. (2002) performed a meta-analysis of over 2600 QTL effects from 572 traits in 86 studies. Roughly half of the studies involve wild $\times$ domesticated crosses, where strong directional selection is suspected for domestication traits. Restricting analysis to those examples with six or more QTLs/trait (Orr's condition for any power), 35 of the 54 qualifying traits (65%) believed to be involved in domesticated showed significant departures from neutrality (i.e., too few anatagonistic QTLs). By contrast, only 14 of 84 non-domestication traits (15.6%) in such crosses showed significant departures. Using this as a control demonstrates that QTLST-EE behaves in the direction predicted for these crosses.

Table 12.3. Summary of the analysis of Rieseberg et al. (2002) on the signs of QTLs in traits from wild species, lumped by trait categories. Within a category, total and antagonistic (opposite sign) QTLs are given, and their ratio. Under the equal effects assumption (QTL effects are symmetrically distributed about zero) this should be 0.5. All ratios are significant at $p < 0.001$ (using QTLST-EE, with p values adjusted using a sequential Bonferroni correction, Appendix 4), except ** which denotes $p < 0.01$. LS means is the QTL ratio for a category after adjusting using an ANOVA estimating the effects for kingdom, crosses, mating selection, and trait types, where † denotes a mean in excess of two standard deviation from zero.

Trait Category	Antagonistic QTLs	Total QTLs	QTL ratio	LS Means
Animals	73	312	0.234	$0.185 \pm 0.039^{\dagger}$
Plants	128	439	0.292	$0.202 \pm 0.025^{\dagger}$
Interspecific	47	245	0.192	0.137 ± 0.154
Intraspecific	154	506	0.304	0.250 ± 0.243
Outcross	98	425	0.231	0.170 ± 0.174
Self	103	326	0.316	0.217 ± 0.262
Life history	111	540	0.206	0.139 ± 0.175
Morphology	138	508	0.272	0.266 ± 0.255
Physiology	8	40	0.200^{**}	0.176 ± 0.125
Timing	37	124	0.298	0.236 ± 0.219
Total	201	751	0.268	

They then examined 42 qualifying traits in wild populations, finding six (14%) of these had QTL proportions significantly different from neutrality. Given that most studies had an average of slightly more than four QTLs per trait, the above restriction to six or more QTLs discards much potential information. To utilize this, they lumped all traits into a series of categories (Table 12.3). Note that the same study can appear in multiple categories, as a morphological trait involving a selfing plant with an interspecific cross appears in the plant, selfing, interspecific, and morphology categories. To control for this, they performed an ANOVA and then adjusted for common effects (their LS mean column in the table), which more accurately reflects the contribution from the category. As shown in the table, all (unadjusted) categories showed significant departures from neutrality, while just two of adjusted (LS means) did. Rieseberg et al. note that interspecific crosses had a smaller fraction of antagonistic QTLs than did within-species (intraspecific) crosses (LS mean ratios of 0.137 versus 0.250), which they interpret as a stronger role for directional selection in generating between-species difference. They also noted that life-history traits (LS ratio 0.139) were (by this criteria) more strongly selection than morphological traits (LS 0.266). While all of these differences are suggestive, it needs to be stressed than none are significant. Further, this pattern of more strongly-selected life history traits is at odds with the meta-

analyses reviewed in Chapter 29 based on the strength of selection as directly measured in the field, but both approaches have potentially significant biases. Rieseberg notes that the widespread departures of neutrality over all traits could result from an ascertainment bias due to investigators focusing on divergent traits in their crosses. Louthan and Kay (2011) also performed a meta-analysis, focusing on traits expected to be under biotic versus abiotic selection. While larger effects were more common for biotically-selected traits, the fraction of antagonistic QTLs was not different between the two groups. Both trait groups showed significant departures from neutrality using QTLST-EE.

An contrary example to Rieseberg's pattern of life history and physiological traits tending to have fewer antagonistic QTLs ironically comes from Rieseberg's own group (Lexer et al. 2005). A cross of two wild sunflowers (*Helanthus annuus* and *H. petiolaris*) found that morphological traits had a smaller ratio (0.287) than did life history or physiological traits (nonsignificant values of 0.345 and 0.388, respectively). Using QTLST with an exponential distribution of effects showed this difference to be significant. However, all of this signal appears to be due to flower morphology (0.214, significant under QTLST), with other morphological features (root/shoot and leaf traits) having similar, nonsignificant, ratios (0.333, 0.322) to life history and physiological traits.

Two interesting applications of sign tests to particular problems were by Albertson et al. (2003) and Muir et al. (2014). Albertson examined traits in the massive species radiation occurring among cichlid fishes in the African rift lake. One striking feature of this radiation is extensive convergent evolution across lakes in feeding morphology, suggesting directional selection. The authors used QTLST (with effect sizes drawn from a gamma distribution) to examine this, crossing two wild species from Lake Malaw. Since most individual traits had less than six QTLs, they grouped traits, finding for jaw and teeth features that only 4 of the 46 QTLs where antagonistic, a highly significant p value, and supporting directional selection on these feeding traits.

Muir et al. (2014) used QTLST along with an estimated distribution of allelic effects in tomatoes (*Lycopersicon*) to examine leaf-related traits in wild species thought to be associated with adaptation to precipitation. They found no significant departure from neutrality for two leaf and two trichome (leaf hair) traits, but a significant departure from neutrality for two stomatal traits. They computed p values using both QTLST and QLTST-EE, and found (in agreement with Anderson and Slatkin 2003) that QTLST was more conservative, giving p about twice as large than those from QLTST-EE.

DIVERGENCE IN GENE EXPRESSION

The power of quantitative genetics is that its machinery can be applied to *any* character of interest, including -omics traits (e.g., amounts of specific transcripts, proteins, and metabolites). Applications of quantitative genetic-machinery to such omis traits has been coined **genetical genomics** by Jansen and Nap (2001), and trace back to Damerval et al. (1994), who mapped QTLs controlling the spot volume of anonymous maize proteins detected by 2-D gel electrophoresis (LW Figure 15.10).

Level of Gene Expression as a Quantitative Trait

Much of the current work in genetical genomics has focused on the transcriptome, treating the level of expression of a specific gene as a quantitative trait, and then attempting to map **eQTLs** (expression QTLs) or **eSNPs** (in GWAS studies) that influence this trait. Modern genomics tools (initially using **microarray analyses** and more recently by **RNA-Seq**) allow

one to measure the level of expression for essentially the full repertoire of an individual's genes (Schena et al. 1995; Duggan et al. 1999; Brown and Botstein 1999; Wang et al. 2009). The amount of mRNA present (either measured by the intensity of hybridization against probes for a gene or directly from the amount present in massive sequencing of an RNA pool) is a typical quantitative trait, showing both genetic and environmental sources of variation, and further confounded by measurement error. This approach yields thousands of traits, as expression levels at each gene are separate, although potentially highly correlated, characters.

Our treatment of the evolutionary analysis of gene expression glosses over a number of very important concerns in the actual generation and processing of the raw data. Gene expression is both highly environmentally- and tissue-specific, and formally should be viewed as a function-valued trait (Volume 3) — a character whose value is indexed by time and potentially other features (such as tissue type or developmental stage). Below, we simply refer to "the" expression level of a gene, but this is highly context-specific, with many transcripts showing considerable variation within an individual (over both time and tissues). There is also technical variation due to sampling and hybridization/amplification so that an otherwise identical sample might still show considerable variation. Much of the early work used whole organisms (and hence all tissues at the sampled developmental stage), and often pooled multiple individuals. RNA-seq relaxes many of these restrictions as it requires much smaller amounts of material. Whichever method is used to assess expression levels, significant care is required to ensure that traits being compared are indeed the same character (i.e., expression levels at the same development time in the same set of tissues). As highlighted by Lynch and Walsh (1998), accurate quantitative genetic studies, especially with noisy traits, require very significant replication, something lacking in many published studies of expression variation. All of these concerns mean that experimental design is absolutely critical (Kerr and Churchill 2001a,b; Churchill 2002; Yang and Speed 2002; Kerr 2003; Rosa et al. 2005).

While microbes have historically had only a relatively minor role in classical quantitative genetics (largely due to their limited number of easily-scored traits), they have flourished in the genetical genomics era, in part because they allow many of the above design issues to be addressed. Starting with Brem and Kruglyak (Brem et al. 2002, 2005; Brem and Kruglyak 2005), yeast has quickly risen to a model system for the quantitative genetics of gene expression, allowing investigators to examine the heritability of expression and map eQTLs for thousands of transcripts. This work was closely followed by similar analyses in mice, man, and maize (Schadt et al. 2003), and has subsequently been extended to an ever-growing number of species. Such genome-wide transcription studies offer very high phenotypic throughput, allowing thousands of traits to be scored in a single experiment. While there is some modest bias against weakly-expressed genes, the sample of expression levels over thousands of loci offers a largely unbiased view of the quantitative genetics of this class of traits. This is an extremely active area, with the early work reviewed by (among others) Stamatoyannopoulos (2004), de Koning and Haley (2005), Gibson and Weir (2005), Rockman and Kruglyak (2006), Whitehead and Crawford (2006), Ranz and Machado (2006), Fay and Wittkopp (2007), Gilad et al. (2008), Skelly et al. (2009), and Emerson and Li (2010). The conclusion from this early work is that control of gene expression often has considerable heritability, involves both *cis* and *trans* factors, and can be very polygenic.

In order to understand the nature of evolutionary forces that shape the complex webs of gene regulation, appropriate null models are needed. That our initial impression of the origins of a given network structure can be highly misleading was stressed in Kauffman's (1969) classic paper. He showed that randomly-constructed gene networks appear to be

highly coordinated and hence give the appearance of being highly evolved (i.e., highly structured by natural selection). Thus, tests of whether certain features of gene regulation (or any -omics data) are largely neutral are critical to understanding how other forces, such as selection, might shape these features. Much of the machinery developed in this chapter for detecting depatures from neutral drift has been applied to divergence in gene expression.

Finally, before proceeding, a subtle, but important, clarification is in order. Our focus here is detecting selection on particular *traits*, while Chapters 9 and 10 discuss machinery for detecting selection on particular *genes* (or, more correctly, specific *sequences*). Selection on the *expression level* at a specific gene is different from selection acting on the *sequence* of that gene. Suppose only trans-acting factors influence the expression levels at our target gene. Frequencies for different trans-factor alleles at the eQTLs for that transcript would be under selection, *not* the alleles at the target gene itself. A test of a sequence-specific signature of selection would not pick up the target gene, but might detect genes coding for these trans-acting factors (subject to all the caveats discussed in Chapters 9 and 10). Likewise, a genome-wide association study would highlight sequence variation in trans-acting genes, *not* sequence variation in the gene whose expression is under selection. This distinction becomes murky when a gene also has *cis*-acting sites, so while the expression level is the target, sequence variation near the gene (*cis*-acting sites), as well as at more distant (often unlinked) sites, would be the genetic targets of selection. An example of such a *cis*-acting site is in the *tb1* gene involved in the domestication of maize (Chapter 9). This site is roughly 60 kilobases upstream of the *tb1* gene itself, and is due to the insertion of a *Hopscotch* retrotransposon that increases the amount of *tb1* transcripts (Studer et al. 2011).

Rate-Based Tests for Neutrality in Divergence of Gene Expression

Early attempts to detect departures of gene expression evolution from neutral trait predictions used rate-based tests. The earliest studies by Rifkin et al. (2003), Hsieh et al. (2003), and Nuzhdin et al. (2004) all used the same basic approach, based on Lande's F_{MDE} test (Equation 12.20a). While the standard version of this test scales V_B by the between-group variance $2\tau\sigma_m^2$ (where τ is the total divergence time), these authors used a modified version with V_B scaled by an estimate (Var) of the within-group variance. At mutation-drift equilibrium, Equation 11.20b gives the expected within-group variance as $2N_e\sigma_m^2$. Assuming expression values are drawn from a normal, then when L lineages are used to estimate the between-group variance and k individuals/lines to estimate the within-group variance, Equation 12.5a shows that both estimators follow χ^2 distributions,

$$V_B \sim (2\,\tau\,\sigma_m^2)\cdot\chi^2_{L-1}/(L-1) \qquad \text{and} \qquad Var \sim (2N_e\,\sigma_m^2)\cdot\chi^2_{k-1}/(k-1) \tag{12.33a}$$

where here the total divergence time $\tau = 2t$, as we are examing the divergence between lineages with a common ancestor t generations ago. These suggest a modified version of Lande's F_{MDE} test statistic,

$$\mathrm{F}_{MDE'} = \frac{V_B}{Var}\cdot\left(\frac{2N_e\sigma_m^2}{2\tau\sigma_m^2}\right) = \frac{V_B}{Var}\cdot\left(\frac{N_e}{\tau}\right) \sim \frac{\chi^2_{L-1}/(L-1)}{\chi^2_{k-1}/(k-1)} \tag{12.33b}$$

where $\mathrm{F}_{MDE'} \sim F_{L-1,k-1}$ (as it is the ratio of two scaled χ^2, see LW Appendix 5). Hence, the ratio of the between- to within-population variation is compared with an appropriate critical F value scaled by the estimate of t/N_e. A scaled ratio less than a critical value of $F_{\alpha/2}$ is suggestive of too little divergence and hence stabilizing selection, while a scaled ratio in excess of $F_{1-\alpha/2}$ implies too much divergence, and directional selection.

Example 12.7. Rifkin et al. (2003) examined variation in gene expression at the start of metamorphous in six inbred lines of *Drosophila*: four *melanogaster*, one *simulans*, and one *yakuba*. Of the roughly 12,900 transcripts examined, 52% (6,742 genes) showed expression changes in at least one lineage (between species or within the *melanogaster* lines). For 4,549 of these transcripts, the authors could not reject the hypothesis that all six lineage-specific samples came from the same distribution, and these were deemed to be evolutionary stable. Of the remainder, 1,666 transcripts showed no significant variation across the sampled *melanogaster* lines, but a significant difference between *melanogaster* and one of the other species. These were deemed to be under lineage-specific selection. The evolutionary forces acting on the remaining 527 genes were examined using Equation 12.33b. One correction is required, as additive vari ance in expression levels was measured as the between-group variance among a set of fully inbred lines, which is twice ($2f = 2$) the value expected in a random-mating population (Table 11.3). Hence, the critical ratio in Equation 12.33b is with respect to $2N_e/\tau$. Divergence was scored separately between *melanogaster* and each of the other two species ($L = 2$), with Var estimated from the four *melanogaster* lines ($k = 4$). The resulting upper and lower 2.5% values critical values follow first by noting that $\Pr(F_{1,3} \geq 17.4) = 0.025$ and $\Pr(F_{1,3} \leq 0.001) = 0.025$. The effective population size for *melanogaster* was estimated as $N_e = 3 \times 10^6$, while the total divergence times (in generations) were estimated as $\tau = 2t = 4.6 \times 10^7$ (*melanogaster - simulans*) and $\tau = 10.2 \times 10^7$ (*melanogaster - yakuba*). Hence the critical values for excessive divergence are

$$F_{c,mel-sim} = 17.4 \cdot \frac{4.6 \times 10^7}{6 \times 10^6} = 133.4, \quad F_{c,mel-yak} = 17.4 \cdot \frac{10.2 \times 10^7}{6 \times 10^6} = 195.8$$

Transcripts whose ratio of divergence to genetic variation exceeded these values are usually divergent. Using this criteria (as well as the lower threshold for too little divergence), of these remaining 527 genes, 464 where consistent with drift, while 63 were consistent with excessive divergence between at least one species pair.

While straight-forward to apply, a concern with Equation 12.33b is in estimation of t and N_e. One could modify the Hudson-Kreitman-Aguadé (HKA) test developed in Chapter 10 where the shared value of t/N_e (among neutral genes) is estimated using a set of the transcripts. However, unless a large fraction of transcripts follow a neutral pattern, this gives a biased estimate. An alternative is to argument the expression data with data from presumably neutral markers. Under neutrality, the ratio of the expected divergence D_S at synonymous sites divided by their expected polymorphism P_s is $D_s/P_s \simeq t/[2N_e]$ (Equation 10.1b). While this marker-based estimate could be substituted into Equation 12.33b, this is an ad-hoc approach, as the sampling error of this ratio as an estimator of t/N_e is ignored. The idea of combining synonymous marker information with the within- and between-population expression variances was also considered by Warnefors and Eyre-Walker (2012), who proposed a MacDonald-Kreitmen type approach (Chapter 10). They substituted the between population divergence for replacement (nonsynonymous) sites (used σ^2_B instead of D_a) and the within-population variation for the nonsynonymous polymorphisms (σ^2_W instead of P_a). These substitutions allowed for construction of expression analogs of the neutrality index (Equation 10.6a) and direction of selection (Equation 10.6b) statistics discussed in Chapter 10. Again, these are helpful metrics, but not a formal statistical tests.

The rate-based test given by Equation 12.33b was scaled to be independent of σ^2_m, at the cost of assuming/estimating an effective population size. Conversely, Equation 12.20c

can be used to compare the rate divergence against candidate values of σ_m^2 (or h_m^2), which circumvents the potentially problematic issue of estimating N_e, although (as with Equation 12.33b) one must still estimate t. This was the approach used by Lemos et al. (2005). Using a diverse series of lineage comparisons (within mice strains, subpopulations of Drosophila, between species in primates and flies), and assuming h_m^2 is in the range of 10^{-4} to 10^{-2}, they found that the majority of gene expression differences were consistent with stabilizing selection (an average of 85% of all transcripts), with drift the next largest category (average 11.5%), and directional selection the smallest (average of $<$4%).

The above approaches either ignore σ_m^2 or estimate its required value under drift given the observed divergence. A more arduous approach is to directly estimate it (or h_m^2) from a mutation-accumulation experiment (LW Chapter 12). This estimate is then used to predict either within-population variation ($2N_e\sigma_m^2$) and/or between-population divergence ($2t\sigma_m^2$), assessing if these are consistent or too extreme. The general conclusion from such studies is that stabilizing selection plays a prominent role in reducing levels of genetic variance in gene expression below the neutral expectation — both within and among species, levels of variation are much lower than expected based on the estimated mutational variance. For example, using lines of the nematode *C. elegans* from a long-term (280 generations) mutation-accumulation experiment, Denver et al. (2005) estimated σ_m^2 for several thousand genes. By comparing levels of variation among a global collection of natural isolates, they found that ratios of standing levels of genetic variance to estimated σ_m^2 were generally no greater than a few hundred. Given that this ratio provides an estimate of $4N_e$ under the assumption of neutrality in a selfing organism (as opposed to $2N_e$ in an outcrosser, as σ_A^2 is inflated by $2f = 2$ with complete inbreeding), these observations provide a firm rejection of the hypothesis that gene expression levels evolve in a neutral fashion. Rifkin et al. (2005) were able to estimate mutational heritabilities for mutation-accumulation lines of *D. melanogaster* by factoring out the variance at the individual fly level to obtain an estimate of σ_e^2. They found a median $h_m^2 \simeq 2.4 \times 10^{-5}$ across all genes, and showed that although interspecific variance in the expression of a gene was correlated with its mutational variance (in qualitative accordance with the neutral theory), the absolute level of divergence was too low to be compatible with neutrality, consistent with the results from Denver et al..

Largely Neutral Evolution of Expression Levels in Primates?

While the above results strongly suggest a leading role for stabilizing selection, there has been considerable discussion regarding evolution of gene expression in the great apes. One key argument has been the expected pattern of divergence under pure drift. Under a Brownian-motion model, the expected divergence (measured by the between-group variance) scales linearly with divergence time t (Equation 12.10, under the infinite-alleles assumption). In contrast, under an Ornstein-Uhlenbeck process (drift countered by stabilizing selection), the divergence approaches an asymptotic value (Equation 12.22c). Bedford and Hartl (2009) used such a process to fit the pattern of expression divergence within a clade of seven species of *Drosophila*. In accordance with the OU model (and consistent with stabilizing selection), they found that the divergence variance does not linearly increase with time, but rather quickly approaches an asymptotic value.

In contrast, Khaitovich et al. (2004, 2005) have argued that gene expression can evolve in a mostly neutral fashion, based in large part on an observed linear increase in among-species expression divergence with time with the clade of great apes. They also noted the observation of Riftkin et al. (2005), namely a positive correlation between levels of within- and among-species variation for the expression of different genes. Such a pattern is expected under neutrality as both divergence and standing variation are functions of

σ_m^2. However, this is not strong support for neutrality, as a number of other features can create such a correlation. For example, genes whose expression is strongly influenced by the environment may naturally exhibit higher levels of variation both within and among samples. Likewise, linearity by itself is suggestive, but not sufficient. Unless the actual rate of divergence is consistent with the rate of polygenic mutation, linear patterns of evolutionary diversification need not imply neutrality, and may instead be a consequence of random fluctuating selection.

An important complication is that Khaitovich et al. use human probes to measure differences in expression among their species, but sequence divergence between the probes and target sites generate reduced levels of hybridization, resulting in an increase in expression divergence over time. Broadley et al. (2008) report a similar linear divergence of expression variance with time in a series of 14 taxa in the Brassicaceae, but again the probe was based on a single species (*Arabidopsis*). Evaluating primates more broadly (human, chimpanzee, orangutan, and rhesus macaque), Gilad et al. (2006) found that the among-species variance in expression of most genes did not increase with divergence time, contrary to the neutral expectation. This study was well designed in that it employed only probes for which the sequences were identical across all four species, removing any species-specific bias in hybridization. Thus, the conclusion that primate gene expression is evolving in a neutral fashion appears to be questionable, and has in fact been essentially retracted by in a more recent analysis (Chaix et al. 2008), which suggests a rate elevation specific to the human lineage.

Transcriptional Q_{ST}

Gibson and Weir (2005) suggested that Q_{ST}-F_{ST} comparisons can be applied to gene expression data, proposing the term ***t*** $\boldsymbol{Q}_{ST}$ for such a transcriptome scan. By scoring a very large number of traits at once, the vexing issue of ascertainment bias (wherein researchers are naturally drawn to the most variable traits to score) that plagues Q_{ST} tests can be avoided. However, the other problems with this approach persist. One is the requirement for estimated genetic variances of within- and between-population components, as opposed to their phenotypic proxies (P_{ST}). The second is low power, especially when only two populations are compared. Perhaps because of these concerns, this approach has not been not widely applied to expression data. It was hinted at by Whitehead and Crawford (2006), but these authors ultimately resorted to rate-based tests for comparing transcripts. It was applied by Roberge et al. (2007) to a very recent population divergence, created by the installation of a fish latter in 1981 on the Sainte-Marguerite River in Quebec. This created up-stream and down-stream populations of Atlantic salmon (*Salmo salar*), which showed an F_{ST} divergence of just over three percent after roughly six generations of presumed differential natural selection. The authors used an animal model to estimate the between-group genetic variance for transcripts from these two populations, but then simply looked for transcripts showing up as Q_{ST} outliers. The found 16 such transcripts, whose average Q_{ST} was roughly three times the F_{ST} value between these two subpopulations, leading the authors to suggest that they have been under extensive directional selection since the population subdivision. However, Equation 12.29b shows that Q_{ST} must be roughly five times as large as F_{ST} for significance when $L = 2$. Thus, while are certainly hints of selection, low power prevents a definitive conclusion.

Cis vs. *Trans*, Local vs. Distant, and Allele-Specific Expression (ASE)

Finally, before discussing applications of sign-based tests to gene expression data, we need to review a few additional features of regulation. Historically, the term *cis* refers to a control

element that only acts on a gene residing on the same DNA molecule. *Cis*-acting control elements are either binding sites for diffusible factors (proteins, small RNAs), target sequences for gene-processing features (intron splice sites, poly-A sites, etc.), or exert some local control over chromatin structure. By contrast, *trans*-acting factors are diffusible, exerting their influence throughout the genome, presumably by coding for proteins or RNAs that interact with specific *cis* sites to control regulation. Formally, the terms *cis* and *trans* refer to this difference in functionality. However, the early eQTL mappers co-opted them to refer to sites that closely mapped to the gene coding for the transcript and those that mapped further away (often on different chromosomes). Rockman and Kruglyak (2006) suggested that the terms **local** and **distant** are more appropriate for describing eQTL location. Given that the uncertainty region for a typical QTL spans tens of centimorgans (and hence tens of megabases), what was called a *cis* eQTL could actually several genes away and act in *trans*.

There are formal genetic procedures for determining whether a region truly does act in *cis*, and these are exploited by a few of the sign-based tests. These employ a more nuanced view of the expression from a single gene in an individual, as not simply its total amount of mRNA, but rather the levels of expression from different *alleles* at that gene, **allele-specific expression**, or **ASE** (Wright and Moyer 1966; Knight 2004). When the gene products from different alleles can be distinguished (either by hybridization or sequencing), then the expression of each can be followed (Yan et al. 2002). Cowles et al. (2002) and Wittkopp et al. (2004) both proposed that the use of hybrids (crosses of different, often inbred, lines) can distinguish *cis* from *trans*. If allele-specific differences seen in the parental lines persists in an F_1 hybrid, these are (at least in part) due to *cis*-acting factors, as the two alternative alleles in the hybrid both experience the same environment and the same set of *trans*-acting factors. Wittkopp et al. (2008) define the amount C of *cis*-acting differences by the ratio of expressions of the two alleles in the hybrid (C = allele 1/allele 2). The amount of *trans*-acting differences is estimated by $T = P - C$, where P = stain 1/strain 2 is the difference in expression in the parental homozygous lines. Since parent-of-origin (i.e., imprinting) effects can also create ASE, reciprocal crosses are used to rule out such an effect. Emerson et al. (2010) offer an alternative approach for determing the amount of *cis* and *trans* effects.

Applications of Sign-Based Tests to Expression Data

While Orr's tests were framed in the increasingly-dated technology of QTL mapping, their central underlying idea (changes under neutrality are unbiased with respect to their direction, or more generally, completely reflect the directionality inherit in new mutations), fits very nicely with genomics era data. We already mentioned a GWAS application by Turchin et al. (2012). There is also an increasing use of sign-based methodology to explore the nature of selection on gene expression. Standard QTL-based tests discussed above are not directly applicable, as most genes have very few detected eQTLs, and thus don't qualify for testing based on Orr's rule of at least six QTLs/trait. As reviewed by Fraser (2011), two rather different approaches have been used to circumvent this limitation.

The first is simply to shift focus from the expression levels at single genes to the pattern of expression over a *set* of genes as the trait, pooling these to create a setting with more than six eQTLs. Bullard et al. (2010) used this approach in a cross of two closely related yeast species, *Saccharomyces cerevisiae* and *S. bayanus*. One key requirement is that each eQTL is independent, as a single eQTL that simultaneously influences k genes should be weighted as one change, not k changes in the same direction. The use of *cis*-regulatory alleles ensures independence over a set of loosely-linked genes. Bullard et al. accomplished this by only considering alleles showing ASE. An excessive number of up-regulated ASEs from one species indicates lineage-specific selection, and a number of pathways were detected show-

ing this feature. Fraser et al. (2011) used a similar approach in a cross of two subspecies of the mouse (*Mus musculus*). They chose gene sets defined by shared GO (Gene Ontology Consortium) membership, finding over 100 genes with evidence of lineage-specific selection. These studies are important, as (at least for these two crosses) they suggest that adaptation via gene expression changes may be widespread, is highly polygenic, and involves *cis*-regulatory sites of many functionally related genes.

As noted by Fraser et al. (2011) and Fraser (2011), a significant result (an excess in one direction) is not necessarily indicative of directional selection, as regulatory mutations are biased towards down-regulation. A pattern seen could simply be a relaxation of purifying selection in one lineage, resulting in a series of neutral, but down-regulated, substitutions given this inherent mutational bias. In essence, this is the same limitation of the McDonald-Kreitman test discussed in Chapter 10. Fraser et al. (2011) note that a simple solution is to examine expression levels in an outgroup, assessing whether a directional change was due to up-regulation (relative to the outgroup) in one lineage, which is likely due to selection, or down-regulation, which could simply be due to relaxed selection. However, selection for reduced expression in a pathway cannot formally be ruled out in the latter case.

A second modification of a sign-based test for expression data was offered by Fraser et al. (2010), and applies to genes whose expression levels are influenced by both *cis* and *trans* eQTLs. The central premise of sign-based tests is that directionality is random, so that in a cross of lines A $\times$ B, if an eQTL from A is a *cis* up-regulator, this should provide no information as to whether a *trans* acting factor from A (acting on the expression level at the same target gene) is an up- or down-regulatory allele. *Cis* and *trans*-acting alleles whose influence is in the same direction (up, up or down, down) are called **reinforcing**, those acting in opposite directions are called **opposing**. With a collection of genes whose expression is influenced by both *cis* and *trans* eQTLs, a simple 2×2 contingency table can be constructed and tested for departures from randomness. If a significant departure is seen, it is straightforward to estimate the amount of excess in a particular class (e.g., Example 10.1). Fraser et al. applied this approach in a cross of two yeast (*Saccharomyces cerevisiae*) strains that diverged roughly 10^7 generations ago, and found an excess of roughly 242 genes showing reinforcing levels of *cis* and *trans*. While this approach suggests significant regulatory evolution over the genome, it does not indicate *which* specific transcripts are involved. This result is reminiscent to those from some of the approaches for detecting genome-wide signatures of selection examined in Chapter 10. Evidence of a genome-wide pattern is seen, but no particular gene can be singled out with confidence as being the target of selection.

Artieri and Fraser (2014) used the above machinery to examine the nature of selection on the *transitional* profiles of mRNA, using **ribosome profiling**. Extracting mRNA bound to ribosomes creates a sample of the mRNA pool actually undergoing translation. Using this enriched pool, Artieri and Fraser examined the translation rates of specific transcripts within and between two species of yeast, finding both *cis*- and *trans*-acting regulatory divergence. They report that the majority of translation divergence appears to buffer the amounts of mRNA, consistent with stabilizing selection on expression levels acting at both the transcriptional and translational stages of gene regulation.

Example 12.8. A related approach to test for selection on specific transcripts using *cis* and *trans* data was suggested by Emerson et al. (2010), who combined their polymorphism data with divergence data from Tirosh et al. (2009) to examine the within- and between-species expression

control in the yeasts *Saccharomyces cerevisiae* and *S. paradoxus*. They used a MacDonald-Kreitman approach (Chapter 10) by examining the fraction of *cis* and *trans* controlled transcripts measured within and between species. Their resulting contingency table

	Polymorphism	Divergence
Cis	396	1270
Trans	412	541

was highly significant, with *trans* polymorphisms being slightly more common than *cis*, but over twice as many *cis* sites fixed. Such a pattern could arise from either an excessive number of *cis* fixations between species, an excessive amount of *trans* polymorphism within a species, or a combination of both. Analogous to arguments with interpretation of MacDonald-Kreitman data discussed in Chapter 10, an excessive amount of polymorphism could arise from a high mutation rate for slightly deleterious alleles. This generates an excess amount of within-species polymorphism that does not transfer to between–species differences (as they are not fixed). Wittkopp et al. (2008), working with *Drosophila*, also noted an excess of *cis* sites being fixed over *trans* sites, and suggested that both fixation of some *cis* mutations by directional selection coupled with a larger number of slightly deleterious *trans* alleles likely underlies this pattern.

Example 12.10. Fraser (2013) suggested yet another approach for detecting selection on the expression level of specific transcripts. First, the genes making up a specific functional set are chosen (i.e., UV protection), and then an expression score for a population is calculated from the mean frequency of all eSNPS that up-regulate members of this set. This is done over a series of roughly sixty human populations that have lived and evolved under different environment conditions (such as summer UV flux), computing the correlation between the environment variable and expression score. The significance of this value is tested using a randomization approach, wherein the correlation between expression score and environmental variable is computed using a random set of genes. This is repeated 10^6 times to generate a distribution for this correlation under the null. This is very similar to methods examined in Chapter 8 to look for SNP frequency-environment factor associations, which was done on a SNP by SNP basis (Hancock et al. 2010a, b, 2011; Fumagalli et al. 2011), focusing on particular genes, as opposed the entire set of regulatory actors in some gene set. Using Fraser's approach, significant signals were detected for transcript sets involved in UV response, immune cell proliferation, and diabetes. Further, comparison to a catalogue of putative locally-adaptive human SNPs (Hancock et al. 2011) found a roughly ten-fold enrichment for eSNPS and SNPs in *cis*-regulatory regions over nonsynonymous SNPs (after correcting for the expected size of each category). This strongly suggesting a more important role for regulation on local adaptation over structural changes.

Evolution of Expression Levels: Drift, Directional or Stabilizing Selection?

The study of gene regulatory evolution is still in a rather embryonic state. Historically, the field has moved from an early era of speculation that regulatory changes may be at least (if not more) important than structural changes (Britten and Davidson 1969, 1971; King and Wilson 1975), to a broader acceptance favoring regulation, at least in some groups/traits (Carroll 2005, 2008; Wray 2007), but see Hoekstra and Coyne (2007) for a countering view. The initial view provided from transcriptome-wide studies is that stabilizing selection is important, but than some drift in expression level can occur, which is more constrained as

phylogenetic distance increases. There is also clear evidence that directional selection on some *cis*-acting sites has occurred between closely-related species.

Taken together, these results (along with those above for more general traits) suggest, perhaps not surprisingly, that at both the phenotypic and gene-regulatory levels, evolution is primarily characterized by periods of stabilizing selection, although relatively brief episodes of directional selection cannot be ruled out. However, it must also be emphasized that the interpretation of conservative rates of evolution is far from clear-cut. In principle, evolutionary divergence rates that are below the expectation of the infinite-elleles model of mutation may be a consequence of the general opposition of selection to all allelic changes associated with the trait, but there might also simply be a fraction of mutations that is truly neutral and another that has strong negative pleiotropic effects on fitness. In that case, an observed level of divergence could actually be entirely based on neutral mutations, but with the appropriate measure of mutational variance being lower than the actual value observed in mutation-accumulation experiments (where even highly deleterious mutations can accumulate). Alternatively, if the House-of-cards is a more appropriate model for mutation, then one would expect cumulative levels of divergence to plateau in time rather than to increase indefinitely, not because of direct selective constraints but because of limited availability of alternative allelic states.

Literature Cited

Ackermann, R. R., and J. M. Cheverud. 2004. Detecting genetic drift versus selection in human evolution. *Proc. Natl. Acad. Sci. USA* 101: 17946–17051. [12]

Aieri, C. G., and H. B. Fraser. 2014. Evolution at two levels of gene expression in yeast. *Genome Res.* 24: 411-421. [12]

Akaike, H. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions* 19: 716–723. [12]

Albertson, R. C., J. T. Streelman, and T. D. Kocher. 2003. Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes. *Proc. Natl. Acad. Sci USA* 100: 5252–5257. [12]

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildlife Manage.* 64: 912–923. [12]

Anderson, E. C., and M. Slatkin. 2003. Orr's quantitative trait loci sign test under conditions of trait ascertainment. *Genetics* 165: 445–446. [12]

Anderson, W. W. 1973. Genetic divergence in body size among experimental populations of *Drosophila pseudoobscura* kept at different temperatures. *Evolution* 27: 278–284. [12]

Baer, C. F., N. Phillips, D. Ostrow, A. Avalos, D. Blanton, A. Boggs, T. Keller, L. Levy, and E. Mezerhane. 2006. Cumulative effects of spontaneous mutations for fitness in *Caenorhabditis:* role of genotype, environment and stress. *Genetics* 174: 1387–1395. [12]

Bailey, D. W. 1959. Rates of subline divergence in highly inbred lines of mice. *J. Heredity* 50: 26-30. [12]

Beaulieu, J. M., D.-C. Jhwueng, C. Boettiger, and B. C. OMeara. 2012. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evol.* 66: 2369–2383. [12]

Bedford, T., and D. L. Hartl. 2009. Optimization of gene expression by natural selection. *Prc. Natl. Acad. Sci. USA* 106: 1133–1138. [12]

Bjöklund, M. 1991. Sexual dimorphisms and mating system in the grackles (*Quiscalus* spp.: Icterinae). *Evolution* 45: 608–621. [12]

Bonnin, I., J.-M. Prosperi, and I. Olivieri. 1996. Genetic markers and quantitative genetic variation in *Medicago truncatula* (Leguminosae): a comparative analysis of population structure. *Genetics* 143: 1795–1805. [12]

Bookstein, F. L. 1987. Random walk and the existence of evolutionary rates. *Paleobiology* 13: 446464. [12]

Bookstein, F. L. 1988. Random walk and the biometrics of morphological characters. *Evol. Biol.* 23: 369–398. [12]

Brem, R. B., and L. Kruglyak. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA* 102: 1572–1577. [12]

Brem, R. B., J. D. Storey, J. Whittle, and L. Kruglyak. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436: 701–703. [12]

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296: 752–755. [12]

Brinkmann, R. 1929. Statistisch-biostratigraphische Untersuch- ungen an mitteljurassischen Ammoniten über Artbegriff und Stammesenwicklung. *Abh., Ges. Wiss. Göttingen, Math-Phys Kl., N.F.* 13: 1–249. [12]

Britten, R. J., and E. H. Davidson. 1969. Gene regulation for higher cells: a theory. *Science* 165: 349–357. [12]

Britten, R. J., and E. H. Davidson. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Quarterly Review Biol.* 46: 111–138. [12]

Broadley, M. R., P. J. White, J. P. Hammond, N. S. Graham, H. C. Bowen, Z. F. Emmerson, R. G. Fray, P. P. M. Iannetta, J. W. McNicol, and S. T. May. 2008. Evidence of neutral transcriptome evolution in plants. *New Phytologist* 180: 587–593. [12]

Brommer, J. E. 2011. Whither P_{ST}? The approximation of Q_{ST} by P_{ST} in evolutionary and conservation biology. *J. Evol. Biol.* 24: 1160–1168. [12]

Brown, P. O., and D. Botstein. 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21: 33–37. [12]

Bryant, E. H., L. M. Combs, and S. A. McCommas. 1986. Morphometric differentiation among experimental lines of the housefly in relation to a bottleneck. *Genetics* 114: 1213–1223. [12]

Bullard, J. H., Y. Mostovoy, S. Dudoit, and R. B. Brem. 2010. Polygenic and directional regulatory evolution across pathways in *Saccharomyces*. *Proc. Natl. Acad. Sci. USA* 107: 5058–5063. [12]

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Amer. Natl.* 164: 683–695. [12]

Carroll, S. B. 2005. Evolution at two levels: on genes and form. *PLoS Biol.* 3: e245. [12]

Carroll, S. B. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134: 25–36. [12]

Casellas, J., and J. F. Medrano. 2008. Within-generation mutation variance for litter size in inbred mice. *Genetics* 179: 2147–2155. [12]

Chaix, R., M. Somel, D. P. Kreil, P. Khaitovich, and G. Lunter. 2008. Evolution of primate gene expression: drift and corrective sweeps? *Genetics* 180: 1379–1389. [12]

Chakraborty, R., and M. Nei. 1982. Differentiation of quantitative characters between populations or species. I. Mutation and random genetic drift. *Genet. Res.* 39: 303–314. [12]

Chang, S. M., and R. G. Shaw. 2003. The contribution of spontaneous mutation to variation in environmental response in *Arabidopsis thaliana:* responses to nutrients. *Evolution* 57: 984–994. [12]

Chapuis, E., G. Martin, and J. Goudet. 2008. Effects of selection and drift on G matrix evolution in a heterogeneous environment: a multivariate $Q_{st} - F_{st}$ test with the freshwater snail *Galba truncatula*. *Genetics* 180: 2151–2161. [12]

Charlesworth, B. 1984. Some quantitative methods for studying evolutionary patterns in single characters. *Paleobiology* 10: 308–318. [12]

Charlesworth, J., and A. Eyre-Walker. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proc. Natl. Acad. Sci. USA* 104: 16992–16997. [12]

Chenoweth, S. F., and M. Blows. 2008. Q_{ST} meets the G matrix: the dimensionality of adaptive divergence in multiple correlated quantitative traits. *Evol.* 62: 1437–1449. [12]

Cherry, L. M., S. M. Case, J. G. Kunkel, J. S. Wyles, and A. C. Wilson. 1982. Body shape metrics and organismal evolution. *Evolution* 36: 914–933. [12]

Churchill, G. A. 2002. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32: 490–495. [12]

Cockerham, C. C. 1973. Analyses of gene frequencies. *Genetics* 74: 679-700. [12]

Cockerham, C. C., and H. Tachida. 1987. Evolution and maintenance of quantitative genetic variation by mutations. *Proc. Natl. Acad. Sci. USA* 84: 6205–6209. [12]

Cockerham, C. C., and B. S. Weir. 1983. Variance of actual inbreeding. *Theor. Pop. Biol.* 23: 85–109. [12]

Cowles, C. R., J. N. Hirschhorn, D. Altshuler, and E. S. Lander. 2002. Detection of regulatory variation in mouse genes. *Nature Genet.* 32: 432–437. [12]

Coyne, J. A. 1996. Genetics of a difference in male cuticular hydrocarbons between two sibling species, *Drosophila simulans* and *D. sechellia*. *Genetics* 143: 1689–1698. [12]

Damerval, C., A. Maurice, J. M. Josse, and D. de Vienne. 1994. Quantitative trait loci underlying gene product variation: a novel perspective by analyzing regulation of genome expression. *Genetics* 137: 289–301. [12]

de Koning, D.-J., and C. S. Haley. 2005. Genetical genomics in humans and model organisms. *Trends Genet.* 21: 377–381. [12]

Denver, D. R., K. Morris, J. T. Streelman, S. K. Kim, M. Lynch, and W. K. Thomas. 2005. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans. Nature Genetics* 37: 544–548. [12]

Dickinson, W. J. 2008. Synergistic fitness interactions and a high frequency of beneficial changes among mutations accumulated under relaxed selection in *Saccharomyces cerevisiae. Genetics* 178: 1571–1578. [12]

Downie, D. A. 2003. Effects of short-term spontaneous mutation accumulation for life history traits in grape phylloxera, *Daktulosphaira vitifoliae. Genetica* 119: 237–251. [12]

Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent. 1999. Expression profiling using cDNA microarrays. *Nat. Genet.* 21: 10–14. [12]

Edelaar, P., and M. Björklund. 2011. If F_{ST} does not measure neutral genetic differentiation, then comparing it with Q_{ST} is misleading. Or is it? *Mol. Ecol.* 20: 1805–1812. [12]

Edelaar, P., P. Burraco, and I. Gomez-Mestre. 2011. Comparisons between Q_{ST} and F_{ST}- how wrong have we been? *Mol. Ecol.* 20: 4830–4839. [12]

Eisen, E. J., and J. P. Hanrahan. 1974. Genetic drift and inbreeding depression measured from control populations of mice. *Can. J. Genet. Cytol.* 16: 91–104. [12]

Eldredge, N., and S. J. Gould. 1972. Punctuated equilibria: an alternative to phyletic gradualism. *In* T. J. M. Schopf (ed.), *Models in paleobiology*, pp. 82–115. Freeman, San Franscion. [12]

Emerson, J. J., L.-C. Hsieh, H.-M. Sung, T.-Y. Wang, C.-J. Huang, . H.-S. Lu, M.-Y. J. Lu, S.-H. Wu, and W.-H. Li. 2010. Natural selection on *cis* and *trans* regulation in yeasts. *Genome Res.* 20: 826–836. [12]

Emerson, J. J., and W.-H. Li. 2010. The genetic basis of evolutionary change in gene expression levels. *Phil. Trans. Roy. Soc. B* 365: 2581–2590. [12]

Estes, S., and S. J. Arnold. 2007. Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. *Amer. Natur.* 169: 227–244. [12]

Eyre-Walker, A., and P. D. Keightley. 2007. The distribution of fitness effects of new mutations. *Nature Rev. Genet.* 8: 610–618. [12]

Fay, J. C., and P. J. Wittkopp. 2007. Evaluating the role of natural selection in the evolution of gene regulation. *Heredity* 100: 191–199. [12]

Felsenstein, J. 1985. Phylogenies and the comparative method. *Amer. Natl.* 125: 1–15. [12]

Felsenstein, J. 2004. *Inferring phylogenies.* Sinauer, Sunderland, MA [12]

Felsenstein, J. 2008. Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *Amer. Nat.* 171: 713–725. [12]

Fraser, H. B. 2011. Genomewide approaches to the study of adaptive gene expression evolution. *Bioessays* 33: 469–477. [12]

Fraser, H. B. 2013. Gene expression drives local adaptation in humans. *Genome Res.* 23: 1089–1096. [12]

Fraser, H. B., T. Babak, J. Tsang, Y. Zhou, B. Zhang, M. Mehrabian, and E. E. Schadt. 2011. Systematic detection of polygenic *cis*-regulatory evolution. *PLoS Genet.* 7: e1002023. [12]

Fraser, H. B., A. M. Moses, and E. E. Schadt. 2010. Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc. Natl. Acad. Sci. USA* 107: 2977–2982. [12]

Fumagalli, M., M. Sironi, U. Pozzoli, A. Ferrer-Admettla, L. Pattini, and R. Nielsen. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics* 7: e1002355. [12]

Garland, T., A. W. Dickerman, C. M. Janis, and J. A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. *System. Biol.* 42: 265–292. [12]

Gibson, G., and B. Weir. 2005. The quantitative genetics of transcription. *Trends Genet.* 21: 616–623. [12]

Gilad, Y., A. Oshlack, G. K. Smyth, T. P. Speed, and K. P. White. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 440: 242–245. [12]

Gilad, Y., S. A. Rifkin, and J. K. Pritchard. 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* 24: 408–415. [12]

Gingerich, P. D. 1974. Stratigraphic record of early Eocene *Hyopsodus* and the geometry of mammalian phylogeny. *Nature* 248: 107–109. [12]

Gingerich, P. D. 1993. Quantification and comparison of evolutionary rates. *Amer. J. Sci.* 293: 453–478. [12]

Goodman, S. J. 1997. R_{st} Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Mol. Ecol.* 6: 881–885. [12]

Goudet, J., and L. Büchi. 2006. The effects of dominance, regular inbreeding and sampling design on Q_{ST}, an estimator of population differentiation for quantitative traits. *Genetics* 172: 1337-1347. [12]

Goudet, J., and G. Martin. 2007. Under neutrality, $Q_{ST} \leq F_{ST}$ when there is dominance in an island model. *Genetics* 176: 1371–1374. [12]

Gould, S. Jay, and N. Eldredge. 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology* 3: 115–151. [12]

Griswold, C. K., and M. C. Whitlock. 2003. The genetics of adaptation: the roles of pleiotropy, stabilizing selection and drift in shaping the distribution of bidirectional fixed mutational effects. *Genetics* 165: 2181–2192. [12]

Gu, X. 2004. Statistical framework for phylogenomic analysis of gene family expression profiles. *Genet.* 167: 531–542. [12]

Haldane, J. B. S. 1949. Suggestions as to quantitative measurement of rates of evolution. *Evolution* 3: 51–56. [12]

Hall, D. W., R. Mahmoudizad, A. W. Hurd, and S. B. Joseph. 2008. Spontaneous mutations in diploid *Saccharomyces cerevisiae:* another thousand cell generations. *Genet. Res.* 90: 229–241. [12]

Hancock, A. M., D. B. Witonsky, E. Ehler, G. Alkorta-Aranburu, C. Beall, A. Gebremedhin, Re. Sukernik, G. Utermann, J. Pritchard, G. Coop, and A. Di Rienzo. 2010a. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc. Natl. Acad Sci. USA* 107: 8924–8930. [12]

Hancock, A. M., G. Alkorta-Aranburu, D. B. Witonsky, and A. Di Rienzo. 2010b. Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Phil. Trans. R. Soc. B* 365: 2459–2468. [12]

Hancock, A. M., D. B. Witonsky, G. Alkorta-Aranburu, C. M. Beall, A. Gebremedhin, R.S ukernik, G. Utermann, J. K. Pritchard, G. Coop, and A. Di Rienzo. 2011. Adaptations to climate-mediated selective pressures in humans. *PLoS Genetics* 7: e1001375. [12]

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51: 1341–1351. [12]

Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution* 50: 1404–1417. [12]

Hendry, A. P. 2002. $Q_{ST} >=\neq< F_{ST}$? *Trends. Ecol. Evol.* 17: 502. [12]

Hill, W. G. 1972. Estimation of genetic change. I. General theory and design of control populations. *Anim. Breed. Abstr.* 40: 1–15. [12]

Hoekstra, H. E., and J. A. Coyne. 2007. The locus of evolution: evo devo and the genetics of adaptation. *Evol.* 61: 995–1016. [12]

Holand, A. M., H. Jensen, J. Tufto, and R. Moe. 2011. Does selection or genetic drift explain geographic differentitation of morphological characters in house sparrows? *Genet. Res. Camb.* 93: 367–379. [12]

Holsinger, K. E., and B. S. Weir. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet.* 10: 639–650. [12]

Hopkins, M. J., and S. Lidgard. 2012. Evolutionary mode routinely varies among morphological traits within fossil species lineages. *Proc. Natl. Acad. Sci. USA* 109: 20520–20525. [12]

Hsieh, W.-P., T.-M. Chu, R. D. Wolfinger, and G. Gibson. 2003. Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genet.* 165: 747–757. [12]

Hunt, G. 2006. Fitting and comparing models of phyletic evolution: random walks and beyond. *Paleobiology* 32: 578–601. [12]

Hunt, G. 2007. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. *Proc. Natl. Acad. Sci. USA* 104: 18404–18408. [12]

Hunt, G. 2008a Gradual or pulsed evolution: when should punctuational explanations be preferred? *Paleobiology* 34: 360–377. [12]

Hunt, G. 2008b. Evolutionary patterns within fossil lineages: model-based assessment of modes, rates, punctuations and process. *In* P. H. Kelley and R. K. Bambach (eds.), *From evolution to geobiology: research questions driving paleontology at the start of a new century. Paleontological Society Papers* 14: 117–131. [12]

Hunt, G., and M. T. Carrano. 2010. Models and methods for analyzing phenotypic evolution in lineages and clades. *In* J. Alroy and G Hunt (eds.), *Quantitative methods in paleobiology*, pp. 245–269. [12]

Hunt, G., M. J. Hopkins, and S. Lidgard. 2015. Simple versus complex models of trait evolution and stasis as a response to environmental change. *Proc. Natl. Acad. Sci. USA* 112: 4885–4890. [12]

Hurst, H. E. 1951. Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.* 116: 770–808. [12]

Jansen, R. C., and J.-P. Nap. 2001. Genetical genomics: the added value from segregation. *Trends Genet.* 17: 388–391. [12]

Kalinka, A. T., K. M. Varga, D. T. Gerrard, S. Preibisch, D. L. Corcoran, J. Jarrells, U. Ohler, C. M. Bergman, and P. Tomancak. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468: 811–814. [12]

Kauffman, S. A. 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* 22: 437–467. [12]

Kavanaugh, C. M., and R. G. Shaw. 2005. The contribution of spontaneous mutation to variation in environmental responses of *Arabidopsis thaliana:* responses to light. *Evolution* 59: 266–275. [12]

Keightley, P. D., and M. Lynch. 2003. Toward a realistic model of mutations affecting fitness. *Evolution* 57: 683–685. [12]

Kerr, M. K. 2003. Design considerations for efficient and effective microarray studies. *Biometrics* 59: 822–828. [12]

Kerr, M. K., and G. A. Churchill. 2001a. Experimental design for gene expression microarrays. *Biostatistic*s 2: 183–201. [12]

Kerr, M. K., and G. A. Churchill. 2001b. Statistical design and the analysis of gene expression microarray data. *Genet. Res.* 77: 123–128. [12]

Khaitovich, P., S. Pääbo, and G. Weiss. 2005. Toward a neutral evolutionary model of gene expression. *Genetics* 170: 929–939. [12]

Khaitovich, P., G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Pääbo. 2004. A neutral model of transcriptome evolution. *PLoS Biol.* 2: 682–689. [12]

Kimura, M. 1983. *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, UK. [12]

Kimura, M., and J. F. Crow. 1963a. On the maximum avoidance of inbreeding. *Genet. Res.* 4: 399–415. [12]

King, M.-C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188: 107–116. [12]

Knight, J. C. 2004. Allele-specific gene expression uncovered. *Trend Genet.* 20: 113–116. [12]

Kronholm, I., O. Loudet, and J. de Meaux. 2010. Influence of mutation rate on estimators of genetic differentiation-lessons from *Arabidopsis thaliana*. *BMC Genetics* 11: 33. [12]

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30: 314–334. [12]

Lande, R. 1977. Statistical tests for natural selection on quantitative characters. *Evolution* 31: 442–446. [12]

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. *Evolution* 33: 402–416. [12]

Lande, R. 1992. Neutral theory of quantitative genetic variance in an island model with local extinction and colonization. *Evolution* 46: 381–389. [12]

Laurie, C. C., J. R. True, J. Liu, and J. M. Mercer. 1997. An introgression analysis of quantitative trait loci that contribute to a morphological difference between *Drosophila simulans* and *D. mauritiana*. *Genetics* 145: 339–348. [12]

Le Corre, V. 2005. Variation at two flowering time genes within and among populations of *Arabidopsis thaliana*: comparison with markers and traits. *Mol. Ecol.* 14: 4181–4192. [12]

Leinonen, T., J. M. Cano, H. Makinen, and J. Merilä. 2006. Contrasting patterns of body shape and neutral genetic divergence in marine and lake populations of threespine sticklebacks. *J. Evol. Biol.* 19: 1803–1812. [12]

Leinonen, T., R. J. S. McCairns, R. B. O'Hara, and J. Merilä. 2013. $Q_{ST} - F_{ST}$ comparisons: evolutionary and ecological insights from genomic heterogeneity. *Nat. Reviews Genet.* 14: 179–190. [12]

Leinonen, T., R. B. O'Hara, J. M. Cano, and J. Merilä. 2008. Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. *J. Evol. Biol.* 21: 1-17. [12]

Lemos, B., G. Marroig, and R. Cerqueira. 2001. Evolutionary rates and stabilizing selection in large-bodied opossum skulls (Didelphimorphia: Didelphidae). *J. Zoology* 255: 181–189. [12]

Lemos, B., C. D. Meiklejohn, M. Cáceres, and D. L. Hartl. 2005. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution* 59: 126-137. [12]

Lewontin, R. C., and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. *Genetics* 74: 175–195. [12]

Lexer, C., D. M. Rosenthal, O. Raymond, L. A. Donovan, and L. H. Rieseberg. 2005. Genetics of species differences in the wild annual sunflowers, *Helianthus annuus* and *H. petiolaris*. *Genetics* 169: 2225–2239. [12]

López-Fanjul, A. Fernández, and M. A. Toro. 2003. The effect of neutral nonadditive gene action on the quantitative index of population divergence. *Genetics* 164: 1627-1633. [12]

López-Fanjul, A. Fernández, and M. A. Toro. 2006. The effect of genetic drift on the variance/covariance components generated by multilocus additive $\times$ additive epistatic systems. *J. Theor. Biol.* 239: 161-171. [12]

López-Fanjul, A. Fernández, and M. A. Toro. 2007. The effect of dominance on the use of the $Q_{ST} - F_{ST}$ contrast to detect natural selection on quantitative traits. *Genetics* 176: 725-727. [12]

Louthan, A. M., and K. M. Kay. 2011. Comparing the adaptive landscape across trait types: larger QTL effect size in traits under biotic selection. *BMC Evol. Biol.* 11: 60. [12]

Lynch, M. 1988. Design and analysis of experiments on random drift and inbreeding depression. *Genetics* 120: 791–807. [12]

Lynch, M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. *Amer. Natur.* 136: 727–741. [12]

Lynch, M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution* 45: 1065–1080. [12]

Lynch, M., and W. G. Hill. 1986. Phenotypic evolution by neutral mutation. *Evolution* 40: 915–935. [12]

Lynch, M., and B. Walsh. 1998. *Genetics and analysis of quantitative traits.* Sinauer Assocs., Inc., Sunderland, MA. [12]

Martin, G., E. Chapuis, and J. Goudet. 2008. Multivariate $Q_{ST} - F_{ST}$ comparisons: a neutrality test for the evolution of the G matrix in structured populations. *Genetics* 180: 2135–2149. [12]

Martins, E. P. 1994. Estimating the rate of phenotypic evolution from comparative data. *Amer. Natl.* 144: 193–209. [12]

Martins, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Amer. Natl.* 149: 646–667. [12]

McKay, J. K., and R. G. Latta. 2002. Adaptive population divergence: markers, QTL and traits. *Trends Ecol. Evol.* 17: 285–291. [12]

Merilä, J. 1997. Quantitative trait and allozyme divergence in the greenfinch (*Carduelis chloris*, Aves: Fringillidae). *Biol. J. Linn. Soc.* 61: 243–266. [12]

Merilä, J., and P. Crnokrak. 2001. Comparison of genetic differentiation at marker loci and quantitative traits. *J. Evol. Biol.* 14: 892–903. [12]

Miller, J. R., B. P. Wood, and M. B. Hamilton. 2008. F_{st} and Q_{st} under neutrality. *Genet.* 180: 1023–1037. [12]

Morgan, T. J., M. A. Evans, T. Garland, J. G. Swallow, and P. A. Carter. 2005. Molecular and quantitative genetic divergence among populations of house mice with known evolutionary histories. *Heredity* 94: 518–525. [12]

Muir, C. D., J. B. Pease, and L. C. Moyle. 2014. Quantitative genetic analysis indicates natural selection on leaf phenotypes across wild tomato species (*Solanum* sect. *Lycopersicon*; Solanaceae). *Genetics* 198: 1629–1643. [12]

Nei, M. 1987. *Molecular evolutionary genetics.* Columbia Univ. Press, New York, NY. [12]

Nuzhdin, S. V., M. L. Wayne, K. L. Harmon, and L. M. McIntyre. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol. Biol. Evol.* 21: 1308–1317. [12]

O'Hara, R. B., and J. Merilä. 2005. Bias and precision in Q_{ST} estimates: Problems and some solutions. *Genetics* 171: 1331–1339. [12]

Orr, H. A. 1998. Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. *Genetics* 149: 2099-2104. [12]

Ostrow, D., N. Phillips, A. Avalos, D. Blanton, A. Boggs, T. Keller, L. Levy, J. Rosenbloom, and C. F. Baer. 2007. Mutational bias for body size in rhabditid nematodes. *Genetics* 176: 1653–1661. [12]

Ovaskainen, O., M. Karhunen, C. Zheng, J. M. Cano Arias, and J. Merilä. 2011. A new method to uncover signatures of divergent and stabilizing selection in quantitative traits. *Genetics* 189: 621–632. [12]

Perry, G. H., P. Melsted, J. C. Marioni, Y. Wang, R. Bainer, J. K. Pickrell, K. Michelini, S. Zehr, A. D. Yoder, M. Stephens, J. K. Pritchard, and Y. Gilad. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* 22: 602–610. [12]

Porcher, E., T. Giraud, I. Goldringer, and C. Lavigne. 2004. Experimental demonstration of a causal relationship between heterogeneity of selection and genetic differentiation in quantitative traits. *Evolution* 58: 1434–1445. [12]

Porcher, E., T. Giraud, and C. Lavigne. 2006. Genetic differentiation of neutral markers and quantitative traits in predominantly selfing metapopulations: confronting theory and experiments with *Arabidopsis thaliana. Genet. Res.* 87: 1–12. [12]

Prout, T., and J. S. F. Barker. 1989. Ecological apects of the heritability of body size in *Drosophila buzzatii*. *Genetics* 123: 803-813. [12]

Prout, T., and J. S. F. Barker. 1993. *F* statistics in *Drosophila buzzatii*: selection, population size and inbreeding. *Genetics* 134: 369–375. [12]

Pujol, B., A. J. Wilson, R. I. C. Ross, and J. R. Pannell. 2008. Are $Q_{ST} - F_{ST}$ comparisons for natural populations meaningful? *Mol. Ecol.* 17: 4782–4785. [12]

Quaas, R. L. 1988. Additive genetic model with groups and relationships. *J. Dairy Sci.* 71: 1338-1345. [12]

Raup, D. M. 1977. Stochastic models in evolutionary palaeontology, *In* A. Hallam (ed.), *Patterns of evolution as illustrated by the fossil record*, pp. 59-78. Elsevier, Amsterdam. [12]

Raup, D. M., and R. E. Crick. 1981. Evolution of single characters in the Jurassic ammonite *Kosmoceras*. *Paleobiology* 7: 200–215. [12]

Ranz, J. M., and C. A. Machado. 2006. Uncovering evolutionary patterns of gene expression using microarrays. *Trends Ecol. Evol.* 21: 29–37. [12]

Reyment, R. A. 1982. Phenotypic evolution in a Cretaceous foraminifer. *Evolution* 36: 1182–1199. [12]

Rhoné, B., R. Vitalis, I. Goldringer, and I. Bonnin. 2010. Evolution of flowering time in experimental wheat populations: a comprehensive approach to detect genetic signatures of natural selection. *Evolution* 64: 2110–2125. [12]

Rice, D. P., and J. P. Townsend. 2012. Resampling QTL effects in the QTL sign test leads to incongruous sensitivity to variance in effect size. *Genes, Genomes, Genetics* 2: 905–911. [12]

Rich, S. S., A. E. Bell, D. A. Miles, and S. P. Wilson. 1984. An experimental study of genetic drift for two quantitative traits in *Tribolium. J. Heredity* 75: 191-195. [12]

Rifkin, S. A., D. Houle, J. Kim, and K. P. White. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* 438: 220–223. [12]

Rifkin, S. A., J. Kim, and K. P. White. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat. Genet.* 33: 138–144. [12]

Rieseberg, L. H., A. Widmer, A. M. Arntz, and J. M. Burke. 2002. Directional selection is the primary cause of phenotypic diversification. *Proc. Natl. Acad. Sci USA* 99: 12242–12245. [12]

Roberge, C., H. Guderley, and L. Bernatchez. 2007. Genomewide identification of genes under directional selection: gene transcription Q_{ST} scan in diverging Atlantic salmon subpopulations. *Genetics* 177: 1011–1022. [12]

Rockman, M. V., and L. Kruglyak. 2006. Genetics of global gene expression. *Nat. Rev. Genet.* 7: 862–872. [12]

Rogell, B., J. Dannewitz, S. Palm, E. Petersson, J. Dahl, T. Prestegaard, T. Järvi, and A. Laurila. 2012. Strong divergence in trait means but not in plasticity across hatchery and wild populations of sea-run brown trout *Salmo trutta*. *Mol. Ecol.* 21: 2963–2976. [12]

Rogers, A. R., and H. C. Harpending. 1983. Population structure and quantitative characters. *Genetics* 105: 985–1002. [12]

Rohlfs, R. V., P. Harrigan, and R. Nielsen. 2013. Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Mol. Biol. Evol.* mst190. [12]

Roopnarine, P. D. 2001. The description and classification of evolutionary mode: a computational approach. *Paleobiology* 27: 446465. [12]

Roopnarine, P. D., G. Byars, and P. Fitzgerald. 1999. Anagenetic evolution, stratophenetic patterns, and random walk models. *Paleobiology* 25: 41–57. [12]

Rosa, G. J. M., J. P. Steibel, and R. J. Tempelman. 2005. Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comp. Funct. Genom.* 6: 123–131. [12]

Roseman, C. C. 2004. Detecting interregionally diversifying natural selection on modern human cranial form by using matched molecular and morphometric data. *Proc. Natl. Acad. Sci. USA* 101: 12824–12829. [12]

Russell, W. A,, G. F. Sprague, and H. L. Penny. 1963. Mutations affecting quantitative characters in long-time inbred lines of maize. *Crop Sci.* 3: 175–178. [12]

Sæther, S. A., P. Fiske, J. A. Kålås, A. Kuresoo, L. Luigujõe, S. B. Piertney, T. Sahlman, and J. Höglund. 2007. Inferring local adaptation from $Q_{ST} - F_{ST}$ comparisons: neutral genetic and quantitative trait variation in European populations of great snipe. *J. Evol. Biol.* 20: 1563–1576. [12]

Savalli, U. M. 1993. An application of the neutral model to the evolution of tail length in the genus *Euplectes* (Aves, Ploceidae). *Evolution* 47: 696–699. [12]

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P, S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422: 297–302. [12]

Schena, M., Dari S., R. W. Davis, and P. O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467–470. [12]

Schultz, S. T., M. Lynch, and J. H. Willis. 1999. Spontaneous deleterious mutation in *Arabidopsis thaliana. Proc. Natl. Acad. Sci. USA* 96: 11393–11398. [12]

Shaw, F. H., C. J. Geyer, and R. G. Shaw. 2002. A comprehensive model of mutations affecting fitness and inferences for *Arabidopsis thaliana. Evolution* 56: 453–463. [12]

Shaw, R. G., D. L. Byers, and E. Darmo. 2000. Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana. Genetics* 155: 369–738. [12]

Sheets, H. D., and C. E. Mitchell. 2001. Why the null matters: statistical tests, random walks and evolution. *Genetica* 112-113: 105–125. [12].

Skelly, D. A., J. Ronald, and J. M. Akey. 2009. Inherited variation in gene expression. *Ann. Rev. Genomics Hum. Genet.* 10: 313–332. [12]

Slatkin, M. 1995a. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457–462. [12]

Spicer, G. S. 1993. Morphological evolution of the *Drosophila virilis* species group as assesed by rate tests for natural selection on quantitative characters. *Evolution* 47: 1240–1254. [12]

Spitze, K. 1993. Population structure in *Daphnia obtusa:* quantitative genetic and allozymic variation. *Genetics* 135: 367-374. [12]

Stamatoyannopoulos, J. A. 2004. The genomics of gene expression. *Genomics* 84: 449–457. [12]

Studer, A., Q. Zhao, J. Ross-Ibarra, and J. Doebley. 2011. Identification of a functional transposon insertion in the maize domesticaion gene *tb1. Nature Gen.* 43: 1160–1165. [12]

Tirosh, I., S. Reikhav, A. A. Levy, and N. Barkai. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* 324: 659–662. [12]

True, J. R., J. Liu, L. F. Stam, Z.-B. Zeng, and C. C. Laurie. 1997. Quantitative genetic analysis of divergence in male secondary sexual traits between *Drosophila simulans* and *Drosophila mauritiana. Evolution* 51: 816–832. [12]

Turchin, M. C., C. W. K .Chiang, C. D. Palmer, S. Sankararaman, D. Reich, J. N. Hirschhorn, and Genetic Investigation of ANthropometric Traits (GIANT) Consortium. 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* 44,: 1015–1019. [12]

Turelli, M., J. H. Gillespie, and R. Lande. 1988. Rate tests for selection on quantitative characters during macroevolution and microevolution. *Evolution* 42: 1085–1089. [12]

Uyeda, J. C., T. F. Hansen, S. J. Arnold, and J. Pienaar. 2011. The million-year wait for macroevolutionary bursts. *Proc. Natl. Acad. Sci. USA* 108: 15908–15913. [12]

Van Valen, L. 1985. Why and how do mammals evolve unusually rapidly? *Evol. Theory* 7: 127–132. [12]

Vassilieva, L. L., A. M. Hook, and M. Lynch. 2000. The fitness effects of spontaneous mutations in *Caenorhabditis elegans. Evolution* 54: 1234–1246. [12]

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57–63. [12]

Warnefors, M., and A. Eyre-Walker. 2012. A selection index for gene expression evolution and its application to the divergence between humans and chimpanzees. *PLoS One* 7: e34935. [12]

Weir, B. S. 1996. *Genetic data analysis II.* Sinauer Assocs., Inc., Sunderland, MA. [12]

Weir, B. S., P. J. Avery, and W. G. Hill. 1980. Effect of mating structure on variation in inbreeding. *Theor. Popul. Biol.* 18: 396–429. [12]

Westell, R. A., R. L. Quaas, and L. Dale Van Vleck. 1988. Genetic groups in an animal model. *J. Dairy Sci.* 71: 1310–1318. [12]

Whitehead, A., and D. L. Crawford. 2006. Variation within and among species in gene expression: raw material for evolution. *Mol. Ecol.* 15: 1197–1211. [12]

Whitlock, M. C. 1999. Neutral additive genetic variance in a metapopulation. *Genet. Res.* 74: 215-221. [12]

Whitlock, M. C. 2008. Evolutionary inference from Q_{ST}. *Mol. Ecol.* 17: 1885-1896. [12]

Whitlock, M. C., and K. J. Gilbert. 2012. Q_{ST} in a hierarchically structured population. *Mol. Ecol. Res.* 12: 481–483. [12]

Whitlock, M. C., and F. Guillaume. 2009. Testing for spatially divergence seelction: Comparing Q_{ST} to F_{ST}. *Genetics* 183: 1055-1063. [12]

Wittkopp, P. J., B. K. Haerum, and A. G. Clark. 2004. Evolutionary changes in *cis* and *trans* gene regulation. *Nature* 430: 85–88. [12]

Wittkopp, P. J., B. K. Haerum, and A. G. Clark. 2008. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat. Genet.* 40: 346–350. [12]

Wray, G. A. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat. Rev. Genet.* 8: 206–216. [12]

Wright, D. A., and F. H. Moyer. 1996. Parental influences on lactate dehydrogenase in the early development of hybrid frogs in the genus *Rana*. *J. Exper. Zoo.* 163: 215–229. [12]

Wright, S. 1921. Systems of mating. II. The effects of inbreeding on the genetic composition of a population. *Genetics* 6: 124–143. [12]

Wright, S. 1951. The genetical structure of populations. *Ann. Eugenics* 15: 323–354. [12]

Wyles, J. S., J. G. Kunkel, and A. C. Wilson. 1983. Birds, behavior, and anatomical evolution. *Proc. Natl. Acad. Sci. USA* 80: 4394-4397. [12]

Yan, H., W. Yuan, V. E. Velculescu, B. Vogelstein, and K. W. Kinzler. 2002. Allelic variation in human gene expression. *Science* 297: 1143–1143. [12]

Yang, Y. H., and T. Speed. 2002. Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* 3: 579–588. [12]

Zeng, Z.-B., and C. C. Cockerham. 1993. Mutation models and quantitative genetic variation. *Genet.* 133: 729–736. [12]