Lecture 12:
Divergence-based tests: II.
Rate of adaptive substitutions, Poisson random field models

UNE course:

The search for selection

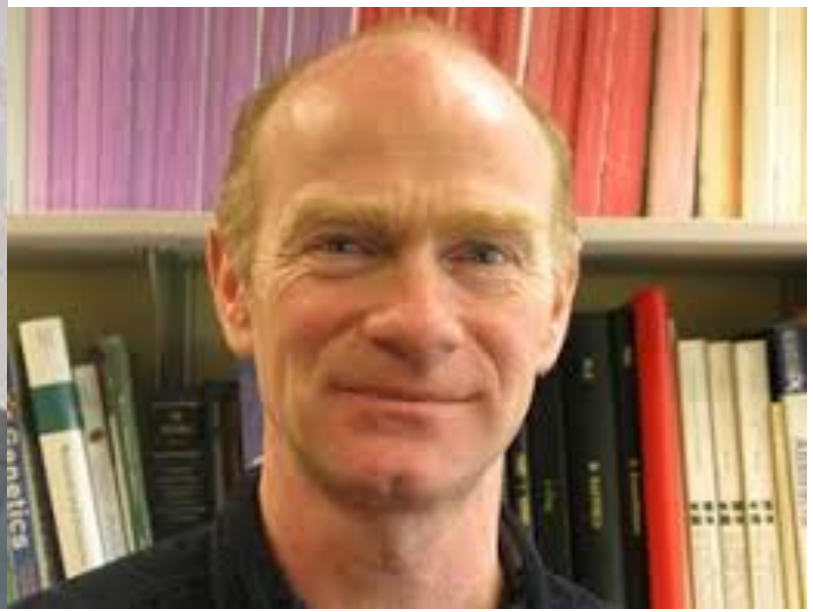
3 -- 7 Feb 2020

Bruce Walsh (University of Arizona)

jbwalsh@email.arizona.edu

Outline

- Fraction, α, of adaptive substitutions
  - Estimation by MK methods
  - Estimation by ML methods
  - Empirical results
- Poisson random field models
  - A ML method using MK-type data
- Connecting the parameters of adaptive evolution



Brian Charlesworth

Adam Eyre-Walker

Peter Keightley

G

E

Jane Charlesworth

Deborah Charlesworth

As above, let μ and $f\mu$ denote the per-site rates at which effectively neutral mutations arise at silent and replacement sites, so that $\mu_a = f\mu n_a$ and $\mu_s = \mu n_s$ are the total rates for the replacement and silent sites in our sample (where n_a and n_s are, respectively, the number of replacement and silent sites). Under neutrality, the expected numbers of effectively neutral substitutions for each class are $D_s = 2\mu_s t$ and $D_{a,n} = 2\mu_a t$. Now suppose there are η_a additional replacement substitutions fixed by positive selection, giving the total number of replacement substitutions as $D_a = D_{a,n} + \eta_a = 2\mu_a t + \eta_a$. Ideally, we would like to estimate both the number, η_a, and the fraction, $\alpha = \eta_a / D_a$, of replacement substitutions that are adaptive. To estimate η_a, note that the expected number of segregating sites for category x is given by $\theta_x a_n$ (Equation 9.21a), yielding $P_s = 4\mu_s N_e a_n$ and $P_a = 4\mu_a N_e a_n$, where the latter assumes that the vast bulk of segregating sites are neutral (adaptive mutations are assumed to be both rare and also fixed quickly, and hence make little contribution to P_a). First note that

$$D_s \frac{P_a}{P_s} = 2\mu_s t \frac{\mu_a}{\mu_s} = 2\mu_a t \tag{10.7a}$$

From above, this last expression is simply the expected number of neutral replacement substitutions, $D_{a,n}$, and because $\eta_a = D_a - D_{a,n}$, our estimate of the number of adaptive replacement substitutions becomes

$$\widehat{\eta}_a = D_a - D_s \frac{P_a}{P_s} \tag{10.7b}$$

as obtained by Charlesworth (1994b), Fay et al. (2001, 2002), and Smith and Eyre-Walker (2002). This immediately suggests an estimator for the fraction, α, of replacement substitutions that are adaptive,

$$\widehat{\alpha} = \frac{\widehat{\eta}_a}{D_a} = 1 - \frac{D_s P_a}{D_a P_s} = 1 - NI \tag{10.7c}$$

Note that a positive estimate of α requires a neutrality index < 1. Using the data from Example 10.6 for noncoding regions on the X chromosome in *D. melanogaster*, $\widehat{\alpha} = 1 - 0.906 = 0.094$ using all polymorphic sites, and $\widehat{\alpha} = 1 - 0.764 = 0.236$ if singletons are ignored. Hence, between roughly 10% and 25% of all substitutions in these noncoding regions might be adaptive. Similarly, Kousathanas et al. (2010) obtained estimates of around 10% adaptive substitutions in the immediate up- and downstream regions around protein-coding genes in the house mouse (*Mus musculus castaneus*).

While Equations 10.7b and 10.7c can be applied to single genes, individual-gene estimates of α are expected to have a large sampling variance and low power. If the actual fraction of adaptive substitutions is small, the modest increase in the number of substitutions will often not be large enough to be significantly different from its neutral expectation, and the resulting estimate of α will not be significantly different from zero. For example, if five substitutions are expected at our focal gene given the ratio of silent to replacement polymorphisms, an observed value of eight substitutions is unlikely to be excessive enough to be declared significantly different from five. However, if three of the eight substitutions were indeed driven to fixation by positive selection, then $\alpha = 0.375$, which is quite substantial.

Despite low power for estimating α at any *single* locus, considerable power can be obtained by estimating the expected value, $E[\alpha] = \overline{\alpha}$, over a *number of loci*. To accomplish this task, Fay et al. (2001, 2002) suggested the estimator

$$\widehat{\overline{\alpha}}_{Fay} = 1 - \frac{\overline{D}_s}{\overline{D}_a}\left(\frac{\overline{P_a}}{\overline{P_s}}\right) \tag{10.8a}$$

where the bar implies the average of that quantity over all sampled genes, e.g., $\overline{D}_s$ is the average number of silent substitutions over all the sampled genes. Note that we use α when referring to a single gene, $\overline{\alpha}$ for its expected value over a set of genes, and $\widehat{\overline{\alpha}}$ as an estimate of $\overline{\alpha}$.

The estimator given by Equation 10.8a has two potential sources of bias, both of which can lead to an overestimation of $\overline{\alpha}$ (Smith and Eyre-Walker 2002; Welch 2006). Let μ and $f\mu$ denote the effectively neutral per-site substitution rates for silent and replacement sites within a gene, where f is allowed to vary over genes. Following Welch (2006), one can show that

$$E\left[\frac{\overline{D}_s}{\overline{D}_a}\right] = \frac{\overline{n}_s}{\overline{n}_a}\frac{1}{E[f]}\left(E\left[\frac{1}{1-\alpha}\right]\right)^{-1} \simeq \frac{\overline{n}_s}{\overline{n}_a}\frac{1}{E[f]}\left[1-\overline{\alpha}-\sigma^2(\alpha)\right] \tag{10.8b}$$

where n_x is the average number of sites of type x over all genes, $E[\cdot]$ is the expectation over all sampled genes, and $\sigma^2(\alpha) = E[\alpha^2] - (E[\alpha])^2$ is the among-gene variance in the fraction of adaptive substitutions (α), with the last approximation following from the delta method (LW Equation A1.3). Equation 10.8b shows that when there is among-locus variation in α (so that $\sigma^2(\alpha) > 0$), $\overline{\alpha}$ is overestimated by Equation 10.8a.

A more subtle bias occurs if f and $4N_e\mu$ are *negatively correlated* over genes, as

$$E\left[\frac{\overline{P}_a}{\overline{P}_s}\right] = \frac{\overline{n}_a}{\overline{n}_s}\left(E[f] + \frac{\sigma(4N_e\mu, f)}{4E[N_e\mu]}\right) \tag{10.8c}$$

as obtained by Smith and Eyre-Walker (2002) and Welch (2006). Hence, Equation 10.7d *underestimates* f, and therefore results in an overestimation of $\overline{\alpha}$, if $4N_eu$ and f are negatively correlated (and underestimates $\overline{\alpha}$ if they are positively correlated). Smith and Eyre-Walker (2002) noted that a negative correlation is biologically reasonable, as the effective population size can vary over the genome (Chapters 3 and 8), and regions with smaller N_e are likely have higher f values (Figure 10.1), as more mutations become effectively neutral.

To reduce bias from correlations between f and N_e, Smith and Eyre-Walker (2002) suggested the estimator

$$\widehat{\widehat{\alpha}}_{SEW} = 1 - \frac{\overline{D_s}}{\overline{D_a}} \overline{\left(\frac{P_a}{P_s + 1} \right)} \tag{10.9a}$$

where the second term is the average of the quantity $P_a/(P_s + 1)$ over the sampled genes. Provided that the number of polymorphic silent sites in the sample is modest (five or greater), this adjusted polymorphism ratio is unbiased by correlations between f and N_e, with

$$E\left[\widehat{\widehat{\alpha}}_{SEW}\right] \simeq \overline{\alpha} + \sigma^2(\alpha) \tag{10.9b}$$

A potential concern with Equations 10.8a and 10.9a is bias due to the Yule-Simpson effect. Recalling Equations 10.7c and 10.6c suggests that the estimator

$$\widehat{\widehat{\alpha}}_{TG} = 1 - NI_{TG} = 1 - \frac{\sum_i D_{si} P_{ai} / (P_{si} + D_{si})}{\sum_i P_{si} D_{ai} / (P_{si} + D_{si})} \tag{10.9c}$$

is perhaps the most robust approach to this problem. While Stoletzki and Eyre-Walker (2011) found very close agreement between $\widehat{\widehat{\alpha}}_{TG}$ and $\widehat{\widehat{\alpha}}_{Fay}$ over the data sets they examined, all of the above considerations suggest that the most prudent estimator is $\widehat{\widehat{\alpha}}_{TG}$. We will refer to estimators of α that use departures from the expectation under neutrality in a DPRS table collectively as **MK estimators** (Equations 10.7c, 10.8a, 10.9a, and 10.9c).

While the above sources of bias (among-locus variation in α and correlations between f and $4N_e\mu$; Equations 10.8b and 10.8c) are generally modest and in a predictable direction (overestimation of $\overline{\alpha}$), the presence of mildly deleterious alleles provides a major bias, which can be either positive or negative (Eyre-Walker 2002; Bieren and Eyre-Walker 2004; Welch 2006; Charlesworth and Eyre-Waker 2008; Eyre-Walker and Keightley 2009; Halligan et al. 2010; Schneider et al. 2011; Keightley and Eyre-Waker 2012; Messer and Petrov 2013b). Estimates of α are downwardly biased by the presence of low-frequency deleterious alleles that contribute to P_a but not D_a, thus inflating the polymorphism ratio relative to the divergence ratio (Eyre-Walker 2006; Eyre-Walker and Keightley 2009). As with MK tests, one approach is to count only "common" polymorphisms for P_a and P_s. However, Charlesworth and Eyre-Walker (2008) noted that while this approach is "better than doing nothing," estimates of α still tend to be downwardly biased even after making this correction unless the true α is fairly substantial. Further, the bias is a function of the complex distribution of fitness effects (Charlesworth and Eyre-Walker 2008; Welch et al. 2008; Eyre-Walker and Keightley 2009; Schneider et al. 2011; Keightley and Eyre-Waker 2012).

Messer and Petrov (2013b) suggested that one simple solution is to estimate $\overline{\alpha}$ using different cutoff levels for rare polymorphisms, with $\overline{\alpha}(x)$ denoting the estimate that ignores polymorphisms whose *derived* allele frequency is below x. Note that $\overline{\alpha}(x)$ could be based on any of our previous MK estimators (e.g., Equations 10.8a, 10.9a, and 10.9c) simply by ignoring polymorphisms below this threshold. Recalculating this statistic for increasing values of x, an exponential regression of the form $\alpha(x) = a + b\exp(-cx)$ is fit to the data, and the asymptotic value (the projected value at $x = 1$) is given by the **Messer-Petrov asymptotic estimate** of $\overline{\alpha}$

$$\overline{\alpha}_{MP} = a + b\exp(-c) \tag{10.9d}$$

Maximum-likelihood (ML) estimators of α have been proposed that attempt to account for segregating deleterious mutations (Bierne and Eyre-Walker 2004; Welch 2006; Boyko et al. 2008; Eyre-Walker and Keightley 2009; Schneider et al. 2011; Keightley and Eyre-Waker 2012). This is done by assuming a standard form (such as a gamma) for the distribution of deleterious fitness effects, and then using site-frequency spectrum data to estimate the parameters of this distribution. We sketch the basic outline of this approach in the next section (in the context of Poisson random field models). While it is elegant and powerful when the model assumptions are correct, the concern is that this approach is highly dependent on the assumed functional form (e.g., gamma, normal, or other) of the unknown distribution of fitness effects for the slightly deleterious mutations. Indeed, Kousanthanas and Keightley (2013) found that these models perform poorly when the distribution of fitness effects is multimodal, and they suggested using nonparametric approaches for such cases.

How Common Are Adaptive Substitutions?

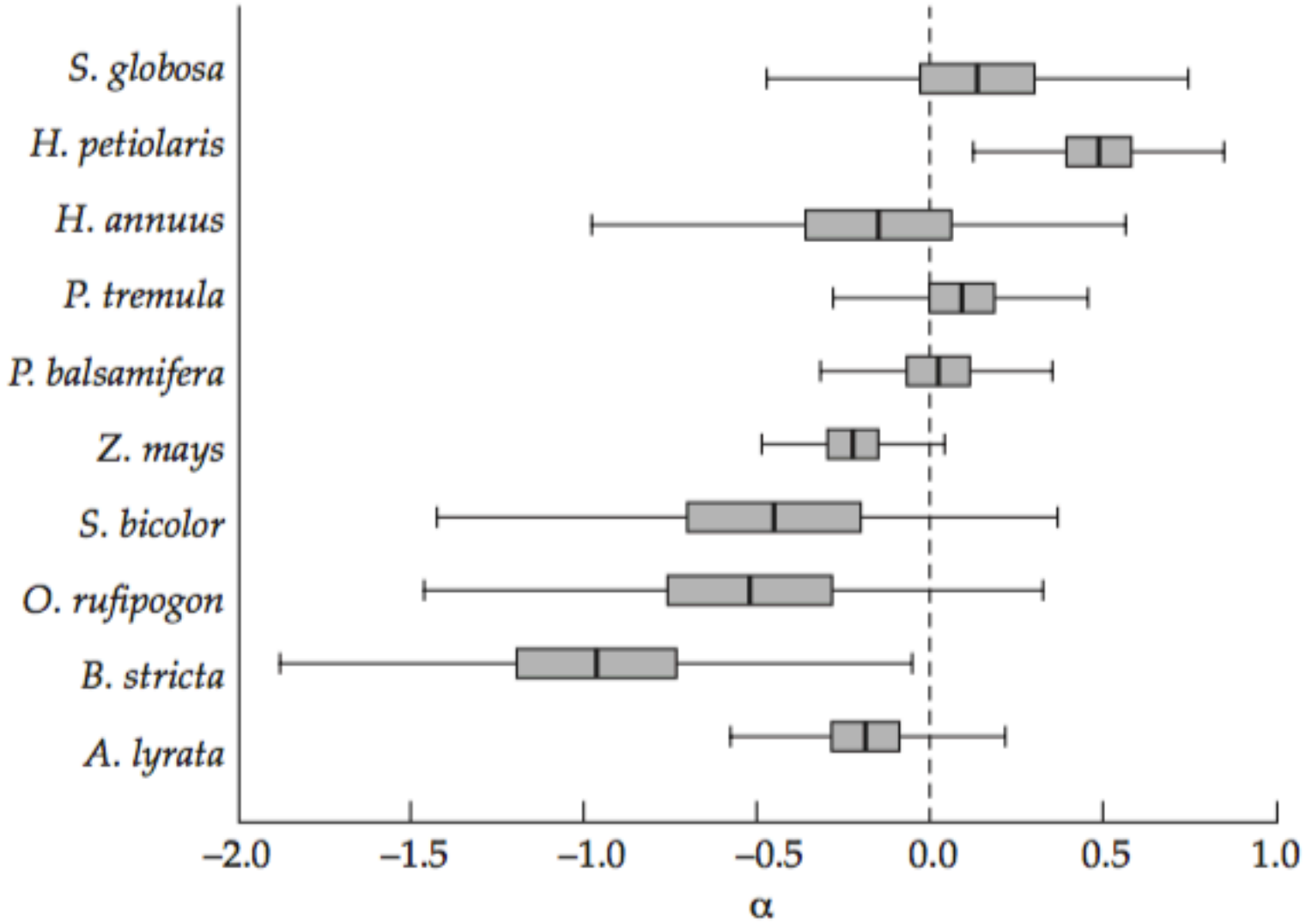
The general observation for *Drosophila* is that estimates of $\overline{\alpha}$ for amino acid substitutions are high, averaging around 50%, with estimates of the fraction of adaptive changes in noncoding regions also approaching 30% in some cases. High $\overline{\alpha}$ values for replacement sites are also observed for the mouse, bacteria, and three plants (*Populus*, *Helianthus*, and *Capsella*), while very low levels are seen in other plants (Table 10.1 and Figure 10.2). Low levels in *Arabidopsis thaliana* were originally attributed to the high levels of selfing in this species (Bustamante et al. 2002), but a close outcrossing relative (*A. lyrata*) similarly shows very low levels of $\overline{\alpha}$ (Foxe et al. 2008). The case receiving the most interest is humans, where an initially rather high estimate of 0.35 by Fay et al. (2001) for a small set of genes was followed by several studies showing much lower values (Table 10.1).

One trend that has been suggested is that $\overline{\alpha}$ increases with effective population size (Eyre-Walker 2006). While intriguing, there are also apparent counterexamples. For example, Bachtrog (2008) found that *D. miranda*, which is thought to have a low effective population size, has a similar value of $\overline{\alpha}$ as *Drosophila* species thought to have a significantly larger values for N_e.

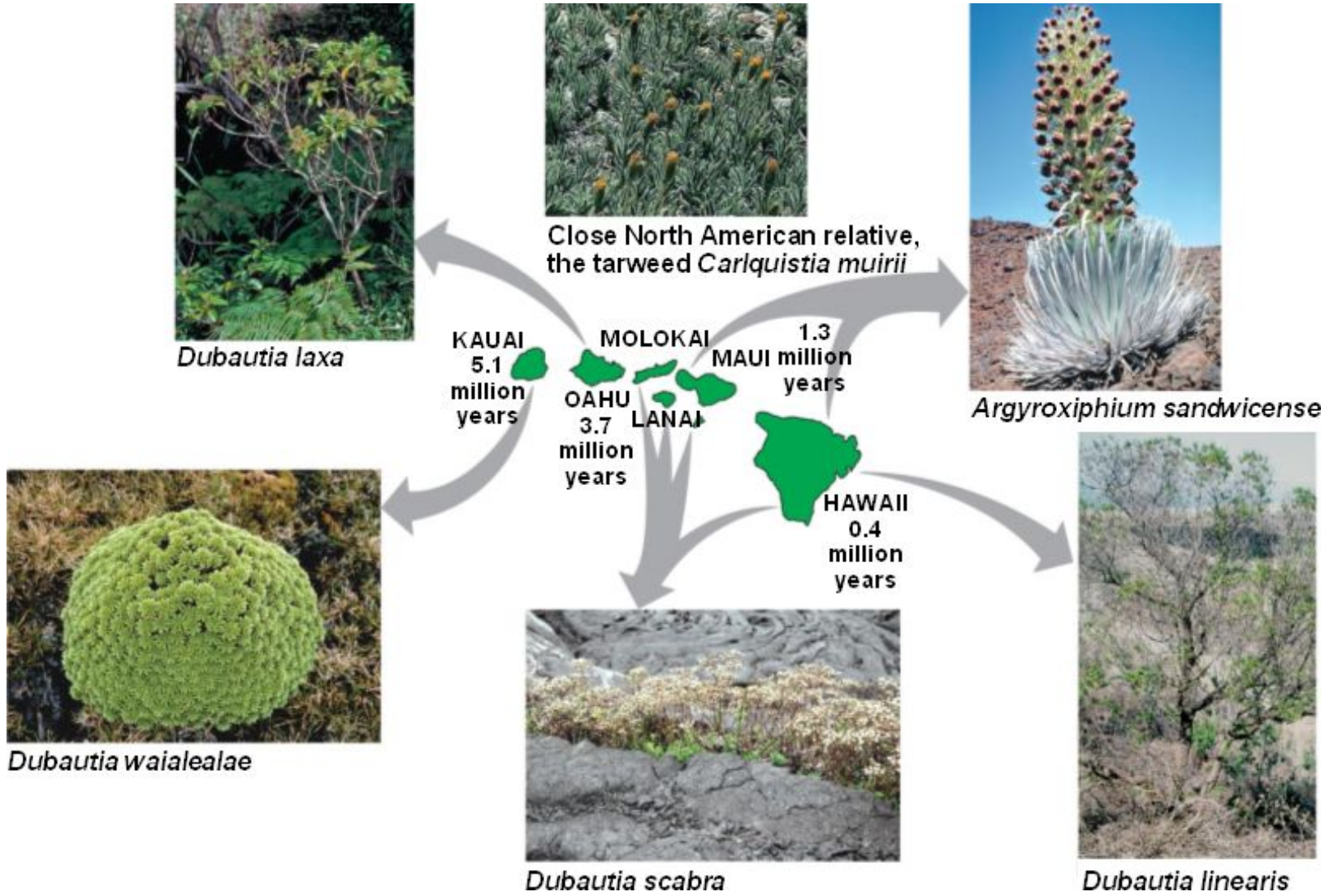
Organism	$\overline{\alpha}$	Method	Reference
Mus musculus castaneus (mouse)	0.57	ML	Halligan et al. 2010
Oryctolagus cuniculus (rabbit)	0.60	MK, ML	Carneiro et al. 2012
Gallus gallus (chicken)	0.20	MK	Axelsson and Ellegren 2009
Drosophila simulans	0.45	MK	Smith and Eyre-Walker 2002
	0.43	ML	Bierne and Eyre-Walker 2004
	0.41	ML	Welch 2006
D. melanogaster	0.44	ML	Bierne and Eyre-Walker 2004
	0.95	PRF	Sawyer et al. 2007
	0.85	ML	Schneider et al. 2011
D. miranda			
Total	0.48	ML	Bachtrog 2008
X chromosome	0.33	MK	Haddrill et. al. 2010
	0.14	ML	
autosomal	0.00	MK	
	0.00	ML	
D. pseudoobscura			
X chromosome	0.44	MK	Haddrill et. al. 2010
	0.70	ML	
autosomal	0.59	MK	
	0.87	ML	

	[illegible]	[illegible]	
Escherichia coli	0.56	MK	Charlesworth and Eyre-Walker 2006
Arabidopsis thaliana	0.00	PRF	Bustamante et al. 2002
A. lyrata	0.00	PRF	Foxe et al. 2008
Capsella grandiflora (crucifer)	0.40	ML	Slotte et al. 2010
Populus tremula (aspen)	0.30	ML	Ingvarsson 2010
Helianthus annuus (sunflower)	0.75	MK	Strasburg et al. 2009
Humans	0.35	MK	Fay et al. 2001
	0.00	MK	Zhang and Li 2005
	0.06	PRF	Bustamante et al. 2005
	0.12	MK	Gojobori et al. 2007
	0.2	IN	Arbiza et al. 2013

Drawing a clear conclusion from these initial data is problematic for several reasons. First, even in the same species, different genes may be used or different populations may be chosen as the polymorphism benchmark. The effect of the latter is especially prominent in Figure 10.2, with the same divergence data between two sunflower species (*Helianthus annuus* versus *H. petiolaris*) showing a significantly positive estimate of mean α when using *Helianthus petiolaris* as the polymorphism reference population, but a negative (but not significant) estimate when using *H. annuus* as the reference population (reminiscent of Example 10.9). Differences in N_e values between the two species being considered can inflate or deflate estimates of α (Equation 10.10). Second, different studies used different methods, ranging from simple MK-type estimators (Equations 10.8 and 10.9) to much more sophisticated, ML-based estimators that attempt to account for both changes in N_e and the presence of segregating deleterious alleles (Bierne and Eyre-Walker 2004; Welch 2006; Eyre-Walker and Keightley 2009). While they are certainly powerful when the modeling assumptions are correct, the robustness of these ML approaches against model misspecification is unclear.



population sizes in the divergence and polymorphism phases. Only the comparison involving sunflowers (polymorphism data from *Helianthus petiolaris*, divergence between *petiolaris* and *annuus*) had an estimated average α that was significantly positive. Surprisingly, the comparison using polymorphism data from *H. annuus* and the same divergence (*petiolaris* versus *annuus*) gave a negative estimate of average α (but was not significantly different from zero).



http://bio1151.nicerweb.com/Locked/media/ch25/radiation.html

of adaptation? One surprising taxon that shows a very low estimated $\overline{\alpha}$ is the Hawaiian silversword plant genus *Schiedea* (family Caryophyllaceae), a group with rapid (and dramatic) morphological evolution over a very recent time window (Gossmann et al. 2010). One possible resolution to this apparent disconnect is that most current studies have focused on the estimation of α in coding sequences, whereas considerable adaptation (especially over short time scales) may occur at the level of gene regulation. Based upon the estimated α values in noncoding regions, Andolfatto (2005, Wright and Andolfatto 2008) suggested that the number of adaptive substitutions in noncoding regions in *Drosophila* could be far greater than the number of adaptive replacement substitutions. Given that *Drosophila* has a compact genome relative to humans and many other metazoans and land plants, the bulk of adaptive variation may not reside in the coding regions that are the focus of most current estimates of $\overline{\alpha}$. An alternative, and not necessarily exclusive, explanation for the *Schiedea* data is that only a few key genes underlie most of the morphological change, resulting in very little change in the genomewide value of $\overline{\alpha}$.

Estimating the Rate, λ, of Adaptive Substitutions

A quantity that prominently appeared in expressions in Chapter 8 on the effects of recurrent sweeps was λ, the per-generation rate at which adaptive substitutions occur. While it might seem that estimates of λ (the number of adaptive substitutions per site divided by the total time of divergence, $2t$) would be very difficult to obtain, fortunately this is not the case, as they follow almost directly from estimates of α (Smith and Eyre-Walker 2002; Andolfatto 2007). If $d_a = D_a/n_a$ denotes the per-site number of replacement substitutions between two species that separated t generations ago, then an upper bound for λ is simply $d_a/(2t)$. The use of D_a to compute d_a involves the assumption that all substitutions have been observed, so that no corrections for multiple substitutions at the same site are needed, which is not unreasonable when comparing two closely related species. With an estimate of α, the number of adaptive replacement substitutions is just αD_a, yielding **Andolfatto's estimator** (2007),

$$\widehat{\lambda} = \frac{\alpha\, d_a}{2t} \tag{10.11a}$$

for the per-site, per-generation rate of adaptive substitutions.

In order to apply Equation 10.11a, one must have an estimate of the divergence time, t. This can be estimated (scaled as $\tau = t/(2N_e)$ generations) from the ratio of D_s/P_s, as follows. From Equations 10.12a and 10.12b,

$$\frac{E[D_s]}{E[P_s]} = \frac{1}{a_m + a_n}\left(\tau + \frac{1}{m} + \frac{1}{n}\right) \tag{10.11b}$$

where m and n are the sample sizes for the two populations and the sample size feature, a_x, is given by Equation 4.3b. Substituting the observed values of D_s and P_s for their expected values and rearranging provides a simple method-of-moments estimator for the scaled divergence time

$$\widehat{\tau} = (a_m + a_n)\,\frac{D_s}{P_s} - \left(\frac{1}{m} + \frac{1}{n}\right) \tag{10.11c}$$

Using this estimate yields $\widehat{t} = 2N_e\widehat{\tau}$, and substituting into Equation 10.11a yields

$$\widehat{\lambda} = \frac{\alpha\, d_a}{2N_e\widehat{\tau}} \tag{10.11d}$$

Note that the estimate offered by Equations 10.11a and 10.11d for the rate, λ, is typically based on *structural changes*, namely, the adaptive rate of amino acid replacement substitutions in protein-coding genes. A more inclusive estimate would also account for regulatory adaptations, which are expected to be *at least* on par with protein structural adaptations (Chapter 9).

Example 10.12. The estimated amino acid divergence between human and chimpanzee proteins is $d_a = 0.008$ (Chimpanzee Sequencing and Analysis Consortium 2005), with a divergence time of roughly 7 million years. If we take $\alpha = 0.10$ (10% of replacement substitutions are adaptive, the rough average for human studies in Table 10.1), then from Equation 10.11a, our estimate of the rate of adaptive replacement substitutions per site, per generation is

$$\lambda = \frac{0.10 \cdot 0.008}{14 \cdot 10^6} = 5.7 \cdot 10^{-11} \text{ per site, per year}$$

Assuming a generation time of 25 years, this corresponds to a rate of $2.3 \cdot 10^{-12}$ per site, per generation.

As a point of comparison, Andolfatto (2007) contrasted X chromosome genes in *Drosophila melanogaster* (for polymorphism data) and *D. simulans* (as the outgroup for divergence). The estimated α was 0.5, while $d_a = 0.028$, and $t = 10^7$ generations, yielding

$$\lambda = \frac{0.50 \cdot 0.028}{2 \cdot 10^7} = 7.0 \cdot 10^{-10} \text{ per site, per generation}$$

Hence (for these data), *Drosophila* have a 12-fold higher per-site adapation rate than humans.

Poisson Random field (PRF) models

- Sawyer and Hartl suggested using an ML method applied to MK data to estimate the
  - Scaled mutation rates
  - Scaled strength of selection
  - Fraction of deleterious, neutral, and advantageous mutations

Stan Sawyer

Dan Hartl

THE SAWYER-HARTL POISSON RANDOM FIELD MODEL

Another approach for extracting information from DPRS tables on the nature and amount of selection is the **Poisson random field (PRF) model** of Sawyer and Hartl (1992). Their initial version assumed that all sites within a region evolve independently and that the strength of selection on all replacement sites was the same. Strongly deleterious mutations were allowed to occur, but the assumption is that these do not contribute to either polymorphism (observed segregating sites) or divergence, and they are accounted for by simply reducing the mutation rate to exclude such mutations. Under this model, the observed counts (P_s, D_s, P_a, and D_a) in a DPRS table follow independent Poisson distributions, whose expected values are functions of four parameters (θ_a, θ_s, τ, and γ). With four observations (the DPRS entries) and four unknowns, we can estimate these parameters, but we cannot assess how well the model fits the data. Two of the parameters are the scaled total mutation rates, $\theta_a = 4N_e\mu_a$ and $\theta_s = 4N_e\mu_s$, while the third parameter is the scaled divergence time, $\tau = t/(2N_e)$. Of most interest is the fourth parameter, the scaled strength of selection, $\gamma = 2N_e s$. Sawyer and Hartl assumed there was additive fitness, so that a new mutation has a fitness of $1 + s$ as a heterozygote and $1 + 2s$ as a homozygote. In contrast to MK approaches, the PRF model does not estimate the fraction, α, of adaptive substitutions directly, but knowledge of γ can allow one to do so indirectly (Example 10.13).

$$E[D_s] = \theta_s \left(\tau + \frac{1}{m} + \frac{1}{n} \right) \tag{10.12a}$$

$$E[P_s] = \theta_s \left(\sum_{j=1}^{m-1} \frac{1}{j} + \sum_{j=1}^{n-1} \frac{1}{j} \right) = \theta_s \left(a_m + a_n \right) \tag{10.12b}$$

$$E[D_a] = \theta_a \left(\frac{2\gamma}{1 - e^{-2\gamma}} \right) \left(\tau + G(m, \gamma) + G(n, \gamma) \right) \tag{10.12c}$$

$$E[P_a] = \theta_a \Big(F(m, \gamma) + F(n, \gamma) \Big) \tag{10.12d}$$

where

$$F(n, \gamma) = \int_0^1 \left(\frac{1 - x^n - (1 - x)^n}{x(1 - x)} \right) \left(\frac{1 - e^{-2\gamma(1-x)}}{1 - e^{-2\gamma}} \right) dx \tag{10.13a}$$

$$G(n, \gamma) = \int_0^1 x^{n-1} \left(\frac{1 - e^{-2\gamma(1-x)}}{2\gamma(1 - x)} \right) dx \tag{10.13b}$$

The basic similarities, and fundamental differences, between MK estimators (e.g., Equations 10.7–10.9) and the PRF approach can be easily obscured by the imposing nature of the PRF equations. The similarity is that both approaches use the same data, the four values in a DPRS table. However, the two approaches estimate different quantities and have different underlying model assumptions. MK estimators make no assumption about the nature or strength of selection on replacement sites, but instead estimate f, the reduction in the effectively neutral substitution rate at replacement sites, and α, the fraction of replacement substitutions at a gene that are adaptive. The effect of purifying selection enters only through f, while the effects of positive selection enter only through α.

In contrast, the PRF equations estimate θ_a and θ_s, the scaled total mutation rates over all sites of the two categories within the gene. The ratio of θ_a/θ_s (suitably corrected for the number of sites within each category; see Equation 10.7d) is *not* an estimate of f, as the PRF model *does* allow for slightly deleterious alleles to be segregating (i.e., the estimate of $\gamma = 2N_e s$ might be negative). The original Sawyer-Hartl model was very restrictive, with only a single fitness class for replacement sites (which is approximately treated as an average selection coefficient over mutations). Extensions (discussed shortly) remove this restriction, allowing for neutral, deleterious, and advantageous classes, with either constant values of γ within each class, or (more generally) class-specific *distributions* of γ values. Thus, the PRF model does *not* estimate α directly, but given estimates of γ, we can compute the expected fraction of substitutions that are fixed by positive selection (Example 10.13 and Equation 10.16c).

The original Sawyer-Hartl analysis equated the observed entries in a DPRS table with their corresponding expected values (Equations 10.12a–10.12d), and then solved for the unknowns of interest (the ratio $\theta_a/\theta_s = \mu_a/\mu_s$, the scaled average strength of selection $\gamma = 2N_e s$, and the scaled time of divergence $\tau = t/[2N_e]$). A value of γ significantly different from zero implies selection on replacement sites, with $\gamma > 0$ implying positive selection and $\gamma < 0$ implying negative selection (the latter applies only to mildly deleterious alleles, as the PRF model treats very deleterious alleles by lowering the mutation rate: these are assumed to be not seen as either polymorphisms or divergences). This original model, which only assumes a single selective class with silent sites being neutral, can be placed in a likelihood framework by recalling that each observed entry in a DPRS table is an independent Poisson random variable. The probability that the count X in a specific category is x, given its expected value ζ, follows from the Poisson distribution,

$$\text{Prob}(X = x \,|\, \zeta) = \zeta^x \exp(-\zeta)/x!, \quad \text{where} \quad \zeta = E[X]$$

The likelihood of the data in the DPRS table for gene i is thus given by

$$L_i = \prod_{j=1}^{4} \left(\frac{\zeta_{i,j}^{x_{i,j}} \exp(-\zeta_{i,j})}{(x_{i,j})!} \right) \tag{10.15}$$

where $x_{i,j}$ denotes the observed DPRS table values for category j in gene i, with

$$x_{i,1} = P_{s,i}, \quad x_{i,2} = P_{a,i}, \quad x_{i,3} = D_{s,i}, \quad x_{i,4} = D_{s,i}$$

and $\zeta_{i,j}$ are the corresponding gene-specific expected values,

$$\zeta_{i,1} = E[P_{s,i}], \quad \zeta_{i,2} = E[P_{a,i}], \quad \zeta_{i,3} = E[D_{s,i}], \quad \zeta_{i,4} = E[D_{a,i}]$$

Note from Equations 10.12a–10.12d that $\zeta_{i,1}$ through $\zeta_{i,4}$ are functions of the unknown parameters ($\theta_{a,i}, \theta_{S,i}, \gamma_i, \tau$) that we wish to estimate by ML. A numerical search over all possible values of these parameters for the combination that maximizes Equation 10.15 given the data (treating the $x_{i,j}$ as fixed constants) yields the ML solutions (LW Appendix 4). Under the assumption of independence across genes, the combined likelihood over k genes becomes

$$L = \prod_{i=1}^{k} L_i$$

where θ_a, θ_S, and γ can potentially vary over the genes, while the divergence time, τ, is shared by all. Hence, for M genes, there are $3M + 1$ unknown parameters.

As noted, this basic model can be expanded by considering more realistic fitness models. For example, Nielsen et al. (2005a) allowed three fitness classes for replacement sites: neutral, deleterious, and beneficial (advantageous). While fitness is assumed to be the same within each class, this is a significant improvement over the original Sawyer-Hartl model. The resulting likelihood now has four parameters for selection (as opposed to one, γ). These are p_b, p_0, and p_d, the frequencies of beneficial, neutral, and deleterious mutations (where $p_b = 1 - p_0 - p_d$), and γ_b and γ_d, the scaled selection coefficients for the beneficial and deleterious alleles (which are assumed to be the same over all genes). Nielsen et al. applied their method to a set of 50 human genes with prior evidence for possible positive selection. The resulting ML estimates were $p_d = 0.748, p_0 = 0.172$, and $p_b = 0.080$ as the fraction of deleterious, neutral, and advantageous mutations, and $\gamma_d = -34.96$ and $\gamma_b = 267.11$ as the scaled strengths of selection of deleterious and advantageous mutations. Note that even in this case where genes were ascertained as likely to be under positive selection, most mutations were still deleterious. A similar analysis of two *Drosophila melanogaster* data sets by Schneider et al. (2011) found that ~1.5% of all replacement mutations were adaptive (i.e., $p_b \sim 0.015$), but with a much smaller scaled strength of selection, $\gamma_b \sim 10$.

While the PRF model does not directly estimate the fraction of adaptive replacements (α), this can be obtained from the estimates of γ and the fraction, p_b, of advantageous mutations as follows. The expected rate of effectively neutral substitutions at replacement sites is μp_0 (the neutral mutation rate), whereas the expected number of favorable mutations arising in each generation is $2N\mu p_b$, where μp_b is the favorable mutation rate. For large γ, each favorable mutation has a fixation probability of $2sN_e/N$ (Chapter 7), for an expected per-generation substitution rate of favorable alleles of

$$\lambda \simeq (2N\mu p_b)(2sN_e/N) = \mu p_b(2\gamma) \tag{10.16a}$$

The fraction of adaptive substitutions is the rate of adaptive substitutions divided by the total rate of substitutions (adaptive plus neutral),

$$\alpha = \frac{\lambda}{\lambda + \mu p_0} \tag{10.16b}$$

Substituting Equation 10.16a yields

$$\alpha = \frac{2\gamma\mu p_b}{2\gamma\mu p_b + \mu p_0} = \frac{2\gamma}{2\gamma + (p_0/p_b)} \tag{10.16c}$$

Example 10.13. What is the estimate of α for the subset of human genes considered by Nielsen et al. (2005a) that was previously discussed (immediately proceeding Equation 10.16a)? Here $p_b = 0.08, p_0 = 0.172$, and $\gamma_b = 267.11$. While only 8% of all new replacement mutations were deemed to be advantageous, α is considerably larger than 0.08, as Equation 10.16c yields

$$\alpha = \frac{2 \cdot 267.11 \cdot 0.08}{(2 \cdot 267.11 \cdot 0.08) + 0.172} = 0.996$$

The reason for this high value is that the estimated advantageous mutation rate (0.08μ) is just slightly below half of the estimated neutral rate (0.172μ), while the fixation probabilities for advantageous mutations are over 500 times greater. If we lumped the neutral and deleterious mutations rates together and assumed these were all effectively neutral (i.e., replacing 0.172 by $1 - p_b = 0.920$), our estimate of α would still be very high (0.980). It is also important to recall that Nielsen et al. focused on a highly biased set of genes, which were chosen to be enriched for positive selection. It is thus likely that the p_b, γ, and α estimates based on this set of loci are larger than those for typical human genes.

Now consider the Schneider et al. (2011) values for *Drosophila melanogaster* ($p_b \sim 0.015$, $\gamma_b \sim 10$). If we assume that all of the remaining mutations are neutral ($p_0 = 1 - p_b = 0.985$), Equation 10.16c yields

$$\alpha = \frac{2\gamma}{2\gamma + (p_0/p_b)} = \frac{20}{20 + 0.985/0.015} = 0.23$$

If we assume that 50% of all new mutation are deleterious ($p_0 = 1 - p_b - 0.5 = 0.485$), then $\alpha = 0.38$. A key point of this example is that α can be quite substantial even when p_b is very small.

Equation 10.16c relates the selection estimates p_b and γ from a PRF model with the selection estimate α from an MK approach. Inspection shows that small p_b (or more precisely a small value of p_b/p_0) does not mean that α is small, as $\alpha > 0.5$ when $2\gamma > p_0/p_b$. One final result emerges from Equation 10.16a. Because μp_b is the rate of beneficial mutation, which (in keeping with our notation from Chapter 8) we denote by μ_b, Equation 10.16a becomes

$$\lambda = 2\gamma\mu_b \tag{10.16d}$$

which immediately suggests the **Bachtrog estimator** (2008),

$$\mu_b = \frac{\lambda}{2\gamma} \tag{10.16e}$$

Doris Bachtrog

One critical difference between PRF and MK analyses is the contribution of information from silent sites (e.g., P_s, D_s), a point stressed by Li et al. (2008). Estimates of selection under an MK analysis are in the form of estimates of α, which are critically dependent upon P_s and D_s (e.g., Equations 10.8a and 10.9a), in addition to D_a and P_a. Conversely, under the PRF model, positive selection is estimated only through γ. An examination of Equations 10.12c and 10.12d shows that estimates of γ depend *only* on D_a and P_a, and that information from silent sites (P_s and D_s) does not enter into them. As a consequence, the control for demographic effects on P_a provided by P_s does not enter, and over- or under-inflated estimates of P_a from population structure can significantly bias estimates of γ. Further, Equation 10.14a (from which the PRF equations follow) is an *equilibrium* model, which assumes that the population size has been stable for sufficient time to reach the mutation-selection-drift equilibrium. Chapter 9 was littered with the bodies of tests that critically depend on this same assumption.

In contrast, because MK estimates involve the ratio of P_a/P_s, recent demographic effects influencing polymorphism levels are accounted for, and there is no assumption about the population being at an equilibrium value for the current amount of genetic variation (see the discussion following Equation 10.5d). Thus, while both MK and PRF approaches face bias from differences in population size between the divergence and polymorphism phases, PRF approaches have additional bias introduced by any nonequilibrium patterns in the polymorphism data. As noted by Li et al. (2008), tests of selection using PRF theory (i.e., γ significantly greater than zero) are closer to an HKA than an MK test, as the former compares the P/D ratio over different genes and lacks the internal control of comparing polymorphism levels from two different classes *within* the same gene.

Bayesian extensions

- $P(\theta|d) = \text{Constant}^* L(d|\theta)\, p(\theta)$
- Essentially an extension of the ideas of ML
  - See WL Appendix 2 for a full introduction
- MCMC methods allow straightforward calculation of draws from the posterior
    - See WL Appendix 3
- An extremely flexible framework for dealing with complex models

Bayesian Extensions

More fined-grain variation in the fitness of replacement mutations was allowed by Bustamante et al. (2002) and Sawyer et al. (2003) in the form of Bayesian models (an approach discussed more fully in Chapter 19 and in great detail in Appendices 2 and 3). Instead of returning a point estimate, $\widehat{\theta}$, for an unknown parameter, θ (or vector of parameters, $\boldsymbol{\Theta}$), a Bayesian analysis returns the full distribution (the **posterior**), $\varphi(\theta \mid \mathbf{x})$, for that parameter, given any previous information (the **prior** for $\boldsymbol{\Theta}$) and the likelihood given the data, $\mathbf{x}$.

Bayesian analysis of PRF data typically uses a **hierarchical model**, the motivation for which comes from random-effects models (Chapter 19). Suppose we have p parameters of interest. Treating the parameters as fixed effects requires p degrees of freedom, but often there are more parameters than observations ($p \gg n$). In some settings, we can treat these p quantities as random effects: draws from some unknown distribution, such as a normal, with unknown mean and variance. Because all draws (realizations) are assumed to come from this common distribution, we can borrow information across observations to estimate the distribution parameters, using (for the case of a normal) only two degrees of freedom (estimation of the unknown mean and variance).

Bayesian hierarchical models take this idea a step further. Consider data structured into a number of categories (say, genes), with multiple observations (draws) from each category (say, new mutations in a particular gene). Assuming that the draws from a given category are all from the same distribution (say, a normal with a category-specific mean and variance), then when the number of categories is large, so too is the parameter set (all of the category-specific means and variances). A hierarchical model reduces the number of parameters to estimate by assuming that the mean (and/or variance) for each category-specific distribution is *itself* a draw from a second distribution. Once each draw is made, these parameter values are fixed for that category. This reduces the estimation problem to one of simply estimating the parameters in the second distribution.

An example of this approach was presented by Bustamante et al. (2002), who assumed that all new replacement mutations at gene i have the same selection value, γ_i, but allowed these gene-specific values to vary among loci. This was done by assuming each γ_i to be a random variable drawn from a normal distribution with a mean of μ_γ and a variance of σ_γ^2, both estimated from the data. In other words, this model allows selection to vary over loci (but not between replacement mutations in the same gene) as a function of just two parameters $(\mu_\gamma, \sigma_\gamma^2)$. Formally, the selection coefficient associated with the jth new replacement mutation at locus i is

$$\gamma_{i,j} = \gamma_i, \quad \text{where } \gamma_i \text{ is a single draw from a } N(\mu_\gamma, \sigma_\gamma^2) \tag{10.17a}$$

Because the divergence time, τ, is a common factor over all genes, this allows information to be borrowed across loci (i.e., all loci contribute to the estimation of τ), improving power, while only loci with sufficient polymorphism and divergence information (a rough rule of thumb is $P_a + D_a \geq 4$) are likely to be informative about γ. Figure 10.3 shows an example of

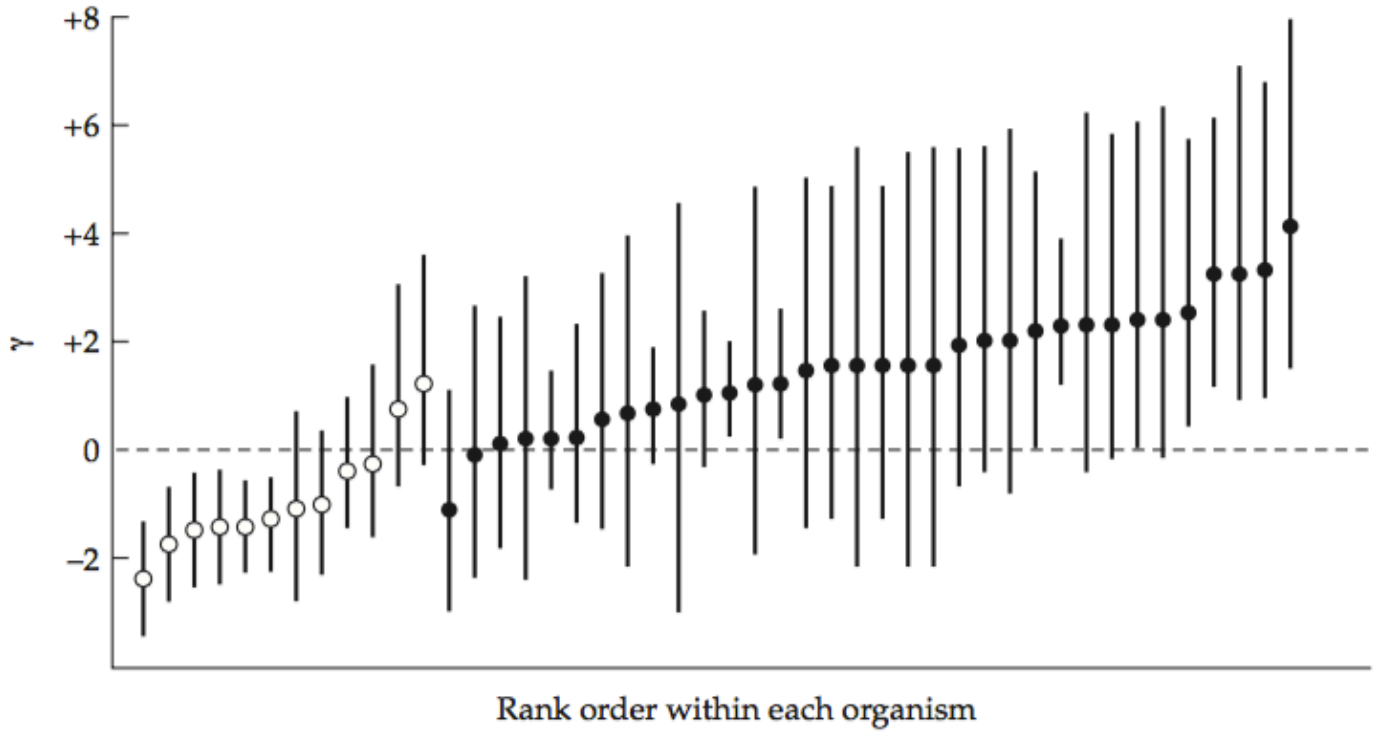


Figure 10.3 Bustamante et al. (2002) examined 12 genes from *Arabidopsis thaliana* (using a single allele from *A. lyrata* to compute divergence) and 34 genes from *D. melanogaster* (with a single allele for *D. simulans*). This figure plots the resulting posterior distribution for γ for each gene (i.e., the locus-specific value, γ_i from Equation 10.17a). The circle represents the mean, and the vertical lines denote the 95% credible intervals (the shortest span of the posterior distribution containing 95% of the probability; Appendix 2). These are plotted by rank order within the two species, with *Arabidopsis* plotted first (open circles) and *D. melanogaster* second (filled circles). If the vertical line is entirely below zero, selection on mutations at this locus is significantly negative (i.e., purifying selection). For lines entirely above zero, selection on new variants is significantly positive. Half (6 of 12) of the *Arabidopsis* genes are significantly negative, while none are significantly positive. Conversely, no *Drosophila* genes are significantly negative, while 9/34 are significantly positive.

Sawyer et al. (2003) extended the Bustamante et al. (2002) approach by allowing each new mutation (j) at locus i to potentially have a different fitness value, $\gamma_{i,j}$, with $\gamma_{i,j} \sim N(\mu_{\gamma,i}, \sigma_w^2)$. Hence, each new mutation has a fitness value drawn from a distribution with a locus-specific mean, $\mu_{\gamma,i}$, and a variance, σ_w^2, that is common over all loci (allowing us to share information over genes). This is a two-stage hierarchical model, where (as in Equation 10.17a) the mean fitness effect, $\mu_{\gamma,i}$, for locus i is drawn from a normal distribution with a mean of μ_γ and a variance of σ_γ^2. Once the locus-specific mean fitness effects are assigned, the fitness of a new replacement mutation at gene i is drawn from a *second* normal, with this locus-specific mean, $\mu_{\gamma,i}$, and a variance, σ_w^2, assumed to be common over all loci (again allowing us to share information over genes). This model can be more compactly written as

$$\gamma_{i,j} \sim N(\mu_{\gamma,i}, \sigma_w^2), \quad \text{where} \quad \mu_{\gamma,i} \sim N(\mu_\gamma, \sigma_\gamma^2) \tag{10.17b}$$

which has three distribution parameters to estimate: $\mu_\gamma, \sigma_\gamma^2$, and σ_w^2. Comparison with Equation 10.17a shows that *each* replacement mutation at a given locus is now a random draw (as opposed to all having the same value), and that (as before) the locus-specific mean also varies. This increased flexibility comes at the cost of only a single additional parameter, σ_w^2, the variance in gene-specific γ values about their mean (under the assumption of homoscedasticity).

Example 10.14. Sawyer et al. (2007) applied their 2003 model (Equation 10.17b) to a sample of 91 genes from an African population of *D. melanogaster*, using a *D. simulans* sequence to assess divergence. After ignoring very strong deleterious mutations that are unlikely to contribute to polymorphisms, they found that approximately 95% of all new replacement mutations are deleterious (estimates of $\gamma_{i,j} < 0$), with an estimated 70% of all replacement polymorphisms observed in the sample being deleterious. Conversely, they estimated that over 95% of the fixed differences at replacement sites are due to positive selection ($\gamma_{i,j} > 0$), albeit it was fairly weak. Within this class of replacement substitutions with estimated positive values, 46% were estimated to have $\gamma_{i,j} < 4$, 85% have $\gamma_{i,j} < 8$, and 99% have $\gamma_{i,j} < 14$.

The parameters of adaptive evolution

- How are the various parameters discussed above connected?
  - Advantageous mutation rate, u_b
  - Fraction of advantageous substitutions, α
  - K_a to K_s ratio, ω
  - Rate, λ, of adaptive evolution
  - Strength, γ, of selection

Table 10.2 Summary of the key parameters of adaptive evolution and their connections. Chapter 8 first introduced several of these (α, γ, and μ_b), while ω and f were introduced in this chapter.

α	The fraction of substitutions that are adaptive
γ	The scaled strength of selection, $2N_e s$
μ	The total per-site mutation rate
μ_s	The effectively neutral per-site mutation rate at silent sites (usually assume $\mu_s \simeq \mu$)
μ_b	The adaptive (benefical) mutation rate
p_b	The fraction of new mutations at a site that are advantageous, $\mu_b = p_b \mu$
λ	The rate of adaptive fixations, $\lambda = 2\gamma\mu_b$
$f = p_0$	The fraction of neutral mutations
$1 - f$	The amount of constraint on a site (relative to some standard, typically silent sites)
ω	The ratio of the replacement- to silent-site substitution rates

$$\omega = f + 2\gamma p_b = \frac{2\gamma p_b}{\alpha} \quad \text{(Equations 10.25a and 10.25c)}$$

$$\gamma = \frac{\omega - f}{2p_b} = \frac{\omega - p_0}{2p_b} \quad \text{(Equation 10.25b)}$$

$$\alpha = \frac{\lambda}{\lambda + \mu p_0} = \frac{2\gamma}{2\gamma + p_0/p_b} = \frac{2\gamma p_b}{\omega} \quad \text{(Equations 10.16b, 10.16c, and 10.25c)}$$

We can connect these parameters as follows. Assume that silent sites are taken as the neutral benchmark, so that (as a first approximation) their per-site mutation rate, μ_s is also the actual mutation rate, μ. Two types of mutations contribute to the rate of replacement substitutions: a fraction f (notationally interchangeable with p_0, as $f = p_0$) that is effectively neutral and a much smaller (perhaps zero) fraction p_b that are favored. Effectively neutral substitutions accrue at a rate of $f\mu_s$, while (Equation 8.24a) beneficial substitutions accrue at rate $\lambda = (2N\mu_b)(2sN_e/N) = 2(2N_e s)\mu_b = 2\gamma\mu_b = 2\gamma p_b \mu_s$. Hence

$$\omega = \frac{K_a}{K_s} = \frac{f\mu_s + 2\gamma p_b \mu_s}{\mu_s} = f + 2\gamma p_b = p_0 + 2\gamma p_b \tag{10.25a}$$

so that very strong, or frequent, selection ($\gamma\, p_b > 1$) is required for $\omega > 1$. Similarly, we can rearrange this equation to solve for γ,

$$\gamma = \frac{\omega - f}{2p_b} = \frac{\omega - p_0}{2p_b} \tag{10.25b}$$

If $f = 0.5$ and $p_b = 0.01$, so that half of the mutations are effectively neutral and 1% are favored, $\gamma = 25$ is required for $\omega = 1$, while $\omega = 3$ requires $\gamma = 125$. If p_b is 0.001, a value of $\gamma = 400$ only gives $\omega = 1.3$, which is a sufficiently small deviation to avoid detection in many cases. Finally, to connect α and ω, from Equations 10.16b and 10.25a, we have

$$\alpha = \frac{2\gamma p_b}{2\gamma p_b + p_0} = \frac{2\gamma p_b}{\omega} \tag{10.25c}$$