8

Hitchhiking and Selective Sweeps

When a mutation B without much selective advantage occurs in the proximity of another mutant gene A with a high selective advantage, the survival chance of gene B is enhanced, and the degree of such enhancement is a function of the recombination fraction between the two loci. Gene B under this situation resembles a hitch–hiker riding along with a host driver — Kojima and Schaeffer (1967)

Draft version 10 April 2014

As first noted by Kojima and Schaeffer (1967) and Maynard Smith and Haigh (1974), the dynamics of a neutral allele are strongly influenced by selection at a linked locus. Over fifty years later, we are still trying to fully understand all of the ramifications of this idea. Chapter 3 provided a brief introduction to two rather different scenarios involving linkage to a selected locus: selective sweeps and background selection. In this chapter we further unpack these concepts, presenting a much richer theoretical treatment and a more detailed account of some of their potential consequences. Results presented here underpin many of the tests for detecting currently ongoing, or very recent, selection developed in Chapter 9.

Our treatment is structured as follows. We start with a review of the basic terminology for different scenarios all loosely referred to as sweeps. Next, we review the population-genetics of hard sweeps, detailing how neutral variation is perturbed by positive selection at linked sites. We then turn to soft sweeps, wherein a preexisting allele is suddenly placed under selection, generating a different pattern of background neutral variation relative to a hard sweep. This naturally leads to a discussion as to whether adaptation to a new challenge occurs by existing variation or by waiting for a new favorable mutation, as well as to the notion of a polygenic sweep (small allele-frequency changes at a number of loci). We conclude with a discussion of the implications of repeated bouts of selection at linked sites (be they recurrent sweeps or background selection) for substitution rates at linked sites, codon usage bias, and whether the current data suggests that a paradigm shift away from Kimura's (1983) classical neutral theory of molecular evolution is needed.

SWEEPS: A BRIEF OVERVIEW

We start with brief overview of the basic terminology and key ideas about sweeps before developing many of these concepts at a more technical level. The casual reader may find this section sufficient from their purposes, while it serves to orient the more diligent reader before proceeding onward.

Hitchhiking, Sweeps, and Partial Sweeps

Although usually attributed to Maynard Smith and Haigh (1974), Kojima and Schaeffer (1967) introduced the term **hitchhiking** to describe the increase in frequency of a neutral allele linked to an allele under directional selection. Plant breeders were also aware of this phenomenon, in the context of **linkage drag** (Brinkman and Frey, 1977), wherein an introgressed favorable region may drag along unfavorable linked genes. The term **selective sweep** (Berry et al. 1991), which is often treated as synonymous with hitchhiking, originally referred to the sweeping away of most variation around a selected site following the fixation of a favorable allele (Figure 8.1A). This cleansing effect occurs because selection reduces the effective population size at linked regions, shortening the coalescence times for surviving neutral alleles relative to pure drift. We return shortly to this important point (Figure 8.3).

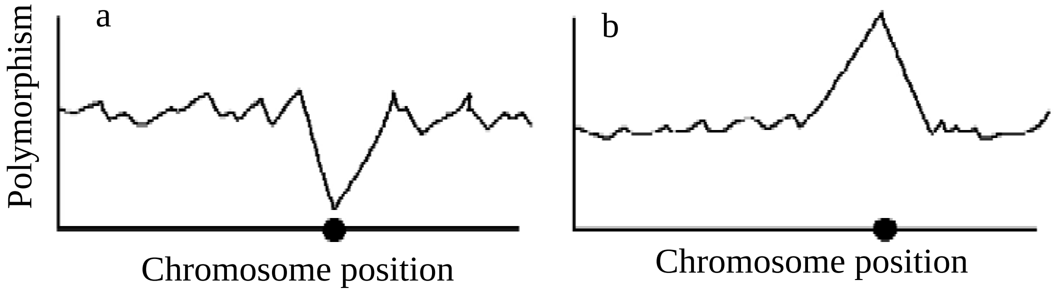

Figure 8.1. **A:** The signature of positive directional selection (a selective sweep) around a selected site (the solid circle). The background levels of linked neutral variation (measured as the average in a sliding window of markers) shows a significant decrease around the selected site, reflecting the decreased effective population size (and hence a shorter time to the most recent common ancestor, TMRCA) for regions linked to this site. **B:** By contrast, stabilizing selection generates an *increase* in the polymorphism level at linked markers, reflecting a longer TMRCA, and hence more opportunities for mutation to generate variation.

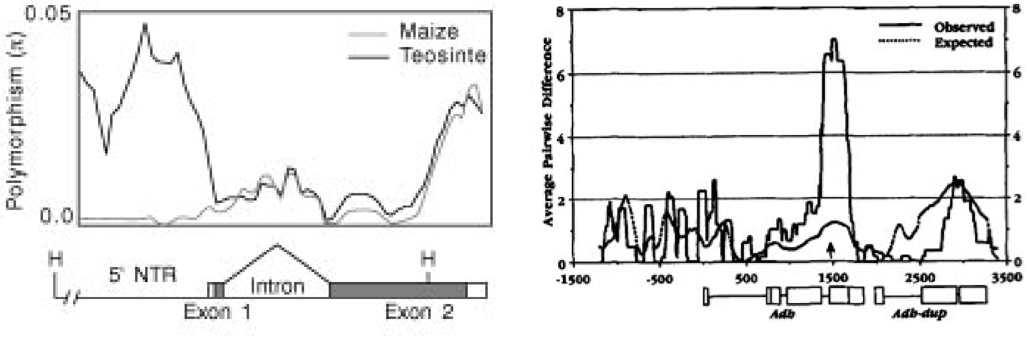

Figure 8.2. Examples of selection influencing levels of polymorphism at linked neutral sites. **Left**: A sliding-window plot of levels of polymorphism around the *tb1* gene in maize (corn) and teosinte, a candidate gene for the domestication of teosinte into corn. Relative to teosinte, maize variation is dramatically reduced in the 5′ UTR region of *tb1*, suggesting a sweep linked to this region. After Wang et al. (1999.) **Right**: Inflated levels of variation are seen around a site that results in a key amino acid change (arrow) in the *Drosophila melangoaster Adh* gene, which has long been suggested to be under balancing selection. This pattern of polymorphism is consistent with this view. After Kreitman and Hudson (1991).

A **partial sweep** refers to the setting where the selected site has not yet reached fixation, either because a sweep is currently underway or because the allele is under **balancing selection**, being driven to some intermediate frequency instead of fixation. As shown in Figures 8.1 and 8.2, a region under long-term balancing selection will show an *increase* in the amount of polymorphism at linked neutral sites (Strobeck 1983; Kaplan et al. 1988; Hudson and Kaplan 1988). This occurs because selection holds alternate alleles at intermediate frequencies for a much longer time than under drift, resulting an older common ancestor relative to the neutral expectation (Figure 8.3), and hence more time for variation to accumulate.

Selection Alters the Coalescent Structure at Linked Neutral Sites

The underpinning for many tests of selection using polymorphism data (Chapter 9) is that ***selection changes the coalescent structure at linked neutral sites***. Describing the structure of genealogical relationships among the alleles in a sample as a **tree**, recent positive selection shortens the **total branch length** (the sum of the lengths of all the branches), decreasing the amount of variation. Conversely, long-term balancing selection generates deeper times to common ancestors (as alleles are retained in the population longer than expected under drift), increasing the amount of variation. This effect is equivalent to a change in the effective population size, with a sweep reducing the effective population size in a linked region (Chapter 2), generating a shorter coalescent times, while balancing selection increases N_e and hence increases coalescent times.

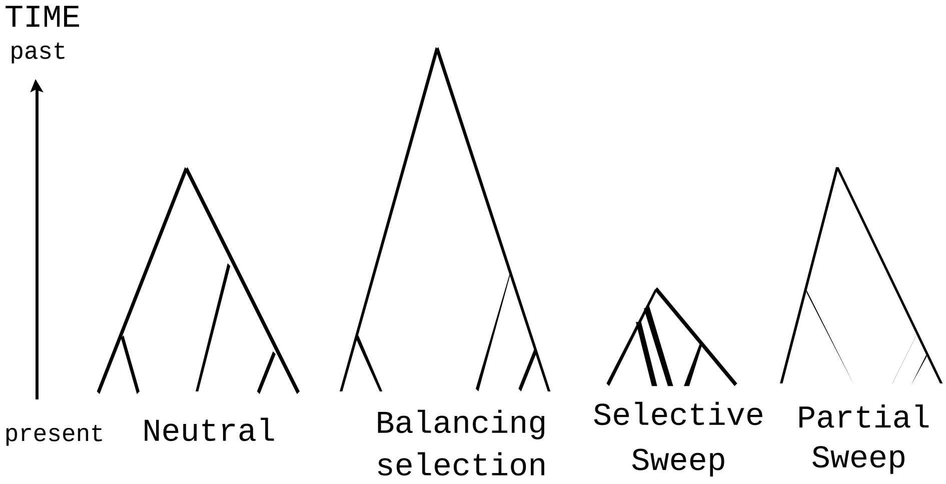

Figure 8.3. Examples coalescent (or genealogical) structures (Chapter 2) for populations under pure drift, balancing selection, a selective sweep, and a partial sweep (ongoing selection where the allele is either not yet fixed or, if under balancing selection, has not yet reached its equilibrium value). The tips of the tree at the bottom of the graph represent five sampled alleles from each population, which eventually coalesce into a single lineage as one goes back in time (the top of the graph). This final coalescent point represents the most recent common ancestor (MRCA) for the sampled alleles. For balancing selection, the time to the MRCA (TMRCA) is greater than for neutral genes, which is turn is greater than a region undergoing a sweep. The shape (topology) of the coalescent is also influenced by selection. Individual coalescent times for a sweep are much more compressed (nodes are closer together) as one moves back in time, while under drift, coalescent times increase as one approaches the MRCA. A partial sweep represents a bit of a mixture, with a sweep-like structure on one part of the genealogy and a drift-like structure in the other.

For a neutral coalescent, trees generated with different N_e have the same expected shape when scaled to the same total length. However, selection at a linked site does more than simply shorten or lengthen the coalescent structure. It ***alters its topology*** as well (Figure 8.3).

Under a selective sweep, the **nodes** of the tree (the coalescent points of the genealogies in the sample) are compressed as one moves back in time, as opposed to being more wide-spread (as is the case with pure drift, Equation 2.40). In the extreme, positive directional selection can generate a **star** (or **palmetto**) **genealogy**, with all lineages coalescing at a single point. In contrast, under pure drift, the expected longest branch lengths are those that coalesce the final two lineages into a single ancestral lineage (Equation 2.40, Figure 2.8). While differences in the total length of the coalescent influence the total *amount* of neutral variation, changes in its *shape* changes the *pattern* of variation from that expected from a simple change in N_e. This is manifested through changes in site/allele frequency spectra (Chapter 2) and patterns of linkage disequilibrium, and these differences underpin a number of tests of selection (Chapter 9). Unfortunately, recovery from a sharp population bottleneck (a crash in population size) generates a very similar, but not quite identical, compression as is seen with directional selection (Barton 1998).

A different coalescent structure is generated during the initial phase of selection when a favorable allele is increasing in frequency (a partial sweep), be it on its way to fixation or increasing to some equilibrium frequency under balancing selection (Figure 8.3). In either case, the resulting tree during the partial sweep phase can be rather unbalanced, with one branch having a sweep-like pattern (reflecting those lineages influenced by the selected allele) and the other a more drift-life pattern (those lineages which have yet to be affected). This coalescent structure is *transient*, and with time will either resolve to a sweep or a balancing selection structure.

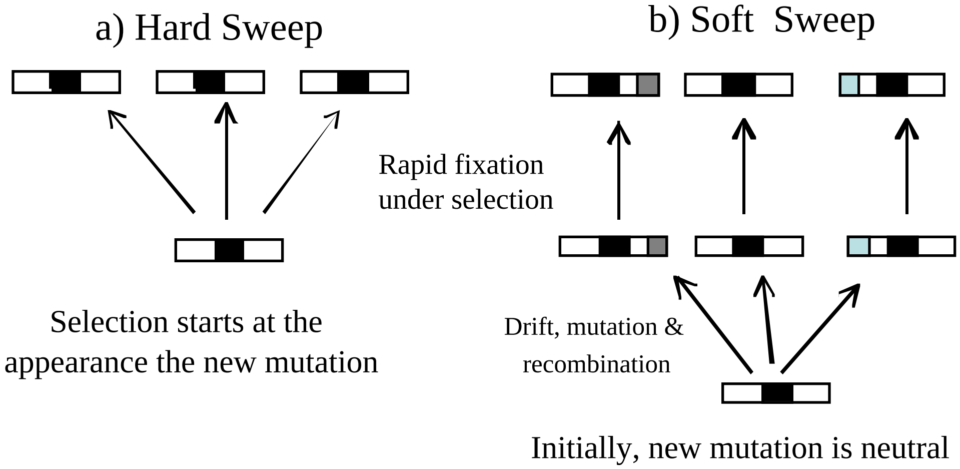
a) Hard Sweep
Selection starts at the appearance the new mutation
b) Soft Sweep
Rapid fixation under selection
Drift, mutation & recombination
Initially, new mutation is neutral

Figure 8.4. A): A hard sweep. A new mutation is immediately favored, resulting in only a single haplotype sweeping to high frequency. **B)**: A single-origin soft sweep. Here a single mutation is initially neutral or even slightly deleterious. It drifts around the population, generating new haplotypes through either mutation or recombination. At some point, an environmental change places this site under strong selection, and it sweeps to fixation carrying along a sample of its existing collection of haplotypes.

Hard versus Soft Sweeps

Not all sweeps, even those involving strong selection, are expected to leave a detectable signal. A **hard sweep** refers to a single favorable new mutation arising and immediately being under selection. The fixation of this mutation drags the haplotype on which it arose to high frequency, leaving a strong signal (Figure 8.4a). In contrast, under a **soft sweep**

(Hermisson and Pennings 2005; Messer and Petrov 2013) multiple haplotypes initially carry the favorable allele. This can occur by two scenarios, which have different consequences for the strength of signal left by the sweep.

Under a **single-origin soft sweep**, the eventually favorable mutation predates the start of selection, being either neutral or perhaps even slightly deleterious when it arose. It drifts around the population, potentially spreading to different haplotype backgrounds, until eventually a change in the environment results in it being favored. This results in selection acting on a more diverse collection of haplotypes, giving a much weaker signal than under a hard sweep. A more formal way to see this difference in the pattern of background variation is that under a **catastrophic sweep** (Perlitz and Stephan 1997), all alleles within a tightly linked region descend from a single founder chromosome τ generations ago, assumed to be at (or near) the start of selection. Conversely, if the frequency of an allele was p at the start of selection, a soft sweep starts as $2pN$ copies. Among these copies (assuming neutrality), the mean coalescent time for a completely linked site in two random individuals is $t = 2pN_e$, where N_e is the effective population size at the start of selection (Innan and Tajima 1997). Thus, there is the potential for substantial divergence ($2t\mu = 4pN_e\mu = p\theta$ per site) among these copies at the start of selection.

Under the second scenario, a **multiple-origin soft sweep**, the fixed favorable allele does not descend from a *single* mutation, but rather a *collection* of multiple independent events (Pennings and Hermisson 2006). Now each recurrent mutation to the favorable allele is associated with an independently chosen-haplotype, potentially creating even more haplotype diversity at fixation than a soft sweep involving a single preexisting allele.

THE BEHAVIOR OF A NEUTRAL LOCUS LINKED TO A SELECTED SITE

We now consider the population-genetics theory of hard sweeps and their effects on linked neutral loci. Parts of this discussion are rather technical, but the main theoretical results are summarized in Table 8.1 and the expected signatures from a hard sweep summarized in Table 8.2. Throughout we assume strong selection ($4N_es \gg 1$) on the favorable allele and (largely) assume no new mutations (either at neutral sites or for the favorable allele) occur during the sweep. This negligible mutation approximation reflects a rapid sweep through the population of a new favorable allele, and hence reduced time for new mutations to arise, and is relaxed in the next section.

Allele-frequency Change

To quantify the impact of a sweep we need to determine how selection influences the frequency q of a neutral allele $\mathbf{m}$ at a linked locus. Let $\mathbf{A}$ denote the favorable allele at the selected site, which has recombination frequency c with the neutral locus. Because $\mathbf{A}$ eventually becomes fixed in the population, we follow the frequency of $\mathbf{m}$ on $\mathbf{A}$-bearing chromosomes to determine the final value of q. Let $q_A(0)$ and $q_a(0)$ denote the frequency of $\mathbf{m}$ on $\mathbf{A}$- and non $\mathbf{A}$- chromosomes the start of selection, with

$$\delta_q = q_A(0) - q_a(0) \tag{8.1a}$$

denoting this initial difference. When $\mathbf{A}$ is introduced as just one or a few copies $q \simeq q_a(0)$. If $\mathbf{A}$ arises as a single copy on an $\mathbf{m}$ chromosome, then $q_A(0) = 1$ (as the only $\mathbf{A}$-bearing chromosome also contains $\mathbf{m}$), giving $\delta_q = 1 - q$. Nonzero values of δ_q imply linkage disequilibrium (nonrandom association) between $\mathbf{A}$ and $\mathbf{m}$, with the frequency of $\mathbf{m}$ on $\mathbf{A}$-bearing chromosomes differing from its unconditional frequency in the general population. Hitchhiking is basically a race between recombination reducing the initial disequilibrium (and hence δ_q) and selection fixing an allele and hence eliminating the chance for further recombination.

Let

$$\Delta_q = q_A(\infty) - q(0) \tag{8.1b}$$

denote the final change in the frequency of linked neutral allele $\mathbf{m}$ after $\mathbf{A}$ has swept through to fixation. Since δ_q and Δ_q represent the initial and final association between $\mathbf{A}$ and $\mathbf{m}$, their ratio

$$f_s = \frac{\Delta_q}{\delta_q} \tag{8.1c}$$

is the fraction of initial associations that persists when $\mathbf{A}$ is fixed, which provides a critical measure the strength of a hitchhiking event. If the sweep is started with a single lineage, f_s is the probability of identity-by-descent at the $\mathbf{m}$ locus among fixed $\mathbf{A}$ chromosomes (Gillepsie 2000; Kim and Nielsen 2004). In the absence of recombination, f_s equals one, resulting in an allele-frequency change of δ_q. With recombination, $f_s < 1$, and our task is to determine how the relative values of selection (s) and recombination (c) determine the values of f_s and Δ_q. We can also express the final allele frequency of $\mathbf{m}$ for the case of $\mathbf{A}$ arising as a single copy ($\delta_q = 1 - q_0$), as

$$q_\infty = q_0 + \Delta_q = q_0 + f_s(1 - q_0) = f_s + q_0(1 - f_s) \tag{8.1d}$$

The derivation of the standard deterministic approximation for Δ_q (Example 8.2) requires a few tricks, and the basic biology can get a bit lost during its development. Hence, we first sketch a rough outline of how selection and recombination compete before presenting more exact results. First, consider the disequilibrium D between $\mathbf{m}$ and $\mathbf{A}$, which (by definition) is just $D =$ freq($\mathbf{Am}$) - freq($\mathbf{A}$)·freq($\mathbf{m}$). We can express this in terms of δ_q and the frequency p of the favorable allele as follows. From the definition of conditional probability,

$$q_A = \text{freq}(\mathbf{m}\,|\,\mathbf{A}) = \frac{\text{freq}(\mathbf{Am})}{\text{freq}(\mathbf{A})} = \frac{\text{freq}(\mathbf{Am})}{p}, \tag{8.2a}$$

with a similar definition for q_a. Conditioning on whether a chromosome contains $\mathbf{A}$, we can express the frequency q of allele $\mathbf{m}$ as

$$\begin{aligned} q &= \text{freq}(\mathbf{m} \mid \mathbf{A})\text{freq}(\mathbf{A}) + \text{freq}(\mathbf{m} \mid \mathbf{a})\text{freq}(\mathbf{a}) \\ &= q_A p + q_a(1-p). \end{aligned} \tag{8.2b}$$

From Equations 8.2a and b, Barton (2000) obtained

$$\begin{aligned} D &= \text{freq}(\mathbf{Am}) - \text{freq}(\mathbf{A}) \cdot \text{freq}(\mathbf{m}) = pq_A - p\left(q_A p + q_a(1-p)\right) \\ &= p(1-p)\left(q_A - q_a\right) = p(1-p)\delta_q, \end{aligned} \tag{8.2c}$$

For a fixed value of p, δ_q declines by $(1-c)$ per generation, so that

$$\delta_q(t) = \delta_q \cdot (1-c)^t \simeq \delta_q\, e^{-ct} \tag{8.2d}$$

Recombination is only effective in changing the frequency of $\mathbf{m}$ on $\mathbf{A}$-bearing chromosomes when there other segregating chromosome types in the population (i.e., $0 < p < 1$). The rapid increase of $\mathbf{A}$ reduces this opportunity, becoming nonexistent when $\mathbf{A}$ is fixed. As shown in Example 8.1, if $\mathbf{A}$ is introduced into the population as a single copy and is destined to become fixed, then its approximate time to fixation is $\tau \simeq 2\ln(4N_e s)/s$. Thus, a crude approximation for the total change in q when $\mathbf{A}$ is fixed is given by the fraction of δ_q that remains after τ generations,

$$\Delta_q \simeq \delta_q e^{-c\tau} \simeq \delta_q \exp\left(-c[2\ln(4N_e s)/s]\right) = \delta_q \left(4N_e s\right)^{-2c/s} \tag{8.2e}$$

Note that it is the ratio c/s that determines the strength of hitchhiking. When $c/s \ll 1$, the total change in the frequency of $\mathbf{m}$ is very close to the value δ_q under complete linkage. As ever-more distant sites are considered (so that c/s increases), Δ_q approaches zero.

Example 8.1. What is the expected time to fixation for an additive allele under strong selection ($4N_e s \gg 1$)? In a strictly deterministic analysis, this is an infinite amount of time, as its frequency gets arbitrarily close to, but never actually reaches, one. However, in a finite population, once the allele frequency is driven sufficiently close to one by selection, it is rapidly fixed by drift. If the scaled strength of selection is large relative to drift ($4N_e s \gg 1$), we can approximate the change in p_t by a deterministic process, provided p is not to very close of zero or one. Near these boundary values, drift determines the dynamics. Hence, a standard approach is to treat p_t as a deterministic process when it is in the range $\epsilon < p < 1 - \epsilon$ for $\epsilon \ll 1$ (Kurtz 1971; Norman 1974; Kaplan et al 1989; Stephan et al. 1992). Once the allele reaches frequency $1 - \epsilon$, it is assumed to be quickly fixed by drift and this additional time is assumed small and ignored.

Let p_t denote the frequency of the favored allele $\mathbf{A}$ at time t. If s is small (but $4N_e s$ large), the deterministic allele-frequency dynamics are well approximated by Equation 5.3a. The solution to this differential equation is given by Equation 5.3b and can alternately be expressed as

$$\frac{p_t}{1-p_t} = \frac{p_0}{1-p_0} e^{st} \tag{8.3a}$$

In particular, the time τ for the frequency of $\mathbf{A}$ to change from $p_0 = \epsilon$ to $p_\tau = 1 - \epsilon$ (where $\epsilon \ll 1$) is obtained by substituting into Equation 8.3a and solving for the time to give

$$\tau = -2\ln(\epsilon)/s \tag{8.3b}$$

Taking $\epsilon = 1/(2N)$, the required time starting from a single copy to reach a frequency very close to one ($1 - 1/[2N]$) is approximately

$$\tau = -2\ln(1/[2N])/s = 2\ln(2N)/s \tag{8.3c}$$

While Equation 8.3c appears often in the literature, it actually *overestimates* the time to fixation in a finite population (and hence *underestimates* the strength of the sweep) and can be improved upon. Again, assume strong selection, $4N_e s \gg 1$. Recall that only a fraction $2sN_e/N$ of single introductions of $\mathbf{A}$ are fixed (Chapter 7). *Conditioned* upon those paths where $\mathbf{A}$ is fixed, its frequency must increase at a faster rate than predicted from the deterministic analysis. Barton (1995, 2000; Otto and Barton 1997) showed that the rate of increase is initially inflated by an amount of $1/(2sN_e/N)$, so that a more accurate estimate of the time for an allele to reach high frequency (essentially become fixed) given it starts as a single copy is given by replacing $\epsilon = 1/(2N)$ by

$$\epsilon = \frac{1}{2N}\,\frac{N}{2sN_e} = \frac{1}{4N_e s},$$

giving

$$\tau = 2\ln(4N_e s)/s. \tag{8.3d}$$

A standard finite population size correction for hitchhiking models starting with $p_0 = 1/(2N)$ is to replace $2N$ by $4N_e s$ to account for this effect.

While Equation 8.2e conveys the general notion of competition between recombination and selection, this is a rather crude analysis, only considering the time to fixation for $\mathbf{A}$

(and hence the end of any opportunity for further recombination). An improved analysis would account for how the actual change in the frequency of **A** influences the opportunity for recombination. This problem has received considerable attention, starting with a strictly deterministic analysis by Maynard Smith and Haigh (1974; also see Stephan et al. 2006), followed by analyses allowing for finite population size by Kaplan et al. (1989), Stephan et al. (1992), Otto and Barton (1997), Barton (1995, 1998, 2000), Durrett and Schweinsberg (2004), Etheridge et al. (2006), Pfaffelhuber et al. (2006), Pfaffelhuber and Studeny (2007), and Ewing et al. (2011).

Under a deterministic analysis accounting for the change in **A** (Example 8.2), if p_0 is the starting frequency of **A** at the time of selection, then for $c/s \ll 1$, the change in q at the fixation of **A** is

$$\Delta_q \simeq \delta_q \, p_0^{c/s}, \tag{8.4a}$$

so that $f_s = p_0^{c/s}$. Recalling that

$$x^a = e^{a\ln(x)} \simeq 1 + a\ln(x) \qquad \text{for} \quad |a\ln(x)| \ll 1 \tag{8.4b}$$

and applying this approximation to Equation 8.4a recovers the original result of Maynard Smith and Haigh,

$$\Delta_q \simeq \delta_q \left[1 + \frac{c}{s} \ln(p_0) \right] \tag{8.4c}$$

$$= \delta_q \left[1 - \frac{c}{s} \ln(2N) \right] \quad \text{for} \quad p_0 = \frac{1}{2N} \tag{8.4d}$$

As Equation 8.4d shows, the hitchhiking effect for a favorable mutation introduced as a single copy diminishes with increasing population size, reflecting the longer time to reach fixation in larger populations and hence a greater reduction of any initial association by recombination. This effect, however, is rather modest, scaling as the log of population size.

When dominance is present, so that the fitnesses are $1 : 1 + 2hs : 2s$, c/s in Equation 8.4 is replaced by $c/(2hs)$ for $h \neq 0$. For the case of a completely recessive allele ($h = 0$), Maynard Smith and Haigh (1974) found that

$$\Delta_q \simeq \delta_q \left(1 - \frac{c}{2s} \, p_0 \right) \tag{8.4e}$$

In this case, $\ln(p_0)$ in Equation 8.4c is replaced by p_0, resulting in a much weaker hitchhiking effect, reflecting the much longer fixation time for a recessive and hence greater opportunity for recombination to decay away any initial disequilibrium. Conversely, the decreased fixation time for a favorable dominant allele effectively doubles the strength of selection (with $c/(2s)$ replacing c/s in Equation 8.4a), resulting in a larger region influenced by the sweep (also see Teshima and Przeworski 2006; Ewing et al. 2011).

When an analysis allowing for drift is performed, using the initial frequency $1/(2N)$ for a single copy *underestimates* the effects of hitchhiking, as those alleles that become fixed leave the drift-dominated boundary region faster than predicted by a deterministic analysis (Example 8.1). This can be corrected for by replacing $p_0 = 1/(2N)$ by $p_0 = 1/(4N_e s)$ in all of the above expressions. While this is a reasonable approximation, there is a growing body of very technical literature focusing on the genealogical structure of sample from a hard sweep for those who wish a more refined analysis (Kaplan et al. 1989; Barton 1998; Etheridge et al. 2006; Pfaffelhuber et al. 2006; Pfaffelhuber and Studeny 2007; Ewing et al. 2011).

Example 8.2. To obtain the final change Δ_q in the frequency of a neutral linked marker under a deterministic model of hitchhiking, we follow Barton (2000). Because **m** is neutral, its frequency on either background only changes through recombination, with

$$q_A(t) - q_a(t) = (1-c)^t\,[q_A(0) - q_a(0)] \sim \delta_q\, e^{-ct}$$

Let q_t' denote the frequency of allele **m** in generation t after selection (but before recombination). Recalling Equation 8.2b, we can express the change in q in generation t by selection (but before recombination) as

$$\begin{aligned}\Delta q_t &= q_t' - q_t = p_t' q_A(t) + (1-p_t')q_a(t) - [p_t q_A(t) + (1-p_t)q_a(t)]\\ &= (p_t + \Delta p_t)q_A(t) + (1 - p_t - \Delta p)q_a(t) - [p_t q_A(t) + (1-p_t)q_a(t)]\\ &= \Delta p_t\,[q_A(t) - q_a(t)]\end{aligned}$$

where Δp_t is the change in A. Recalling our previous result for $q_A(t) - q_a(t)$, we have

$$\Delta q_t = \Delta p_t \delta_q\, e^{-ct}$$

The final frequency is just the sum of all these single-generation changes, which we approximate by an integral. Further noting that $\Delta p_t = \Delta p/\Delta t \simeq dp/dt$ gives

$$q = \int_0^\infty \Delta q_t\, dt = \int_0^\infty \Delta p_t \delta_q\, e^{-ct} dt = \int_0^\infty \delta_q\, e^{-ct}\frac{dp}{dt}dt = \delta_q \int_{p_o}^1 e^{-ct} dp$$

where the last integral follows by a change of variables with $p(0) = p_0$ and $p(\infty) = 1$. The trick to evaluating this last integral is to recall Equation 8.3a, and noting that $1 - p_0 \simeq 1$ (since $p_0 \ll 1$), giving

$$\frac{p_t}{1-p_t} = \frac{p_0}{1-p_0}e^{st} \simeq p_0\, e^{st}.$$

Rearranging gives

$$p_0 \frac{1-p_t}{p_t} = e^{-st}$$

Noting that $e^{ab} = (e^a)^b$, we can write $e^{-ct} = e^{-cst/s} = (e^{-st})^{c/s}$. Hence,

$$e^{-ct} = \left(e^{-st}\right)^{c/s} = \left(p_0\,\frac{1-p_t}{p_t}\right)^{c/s} = p_0^{c/s}\left(\frac{1-p_t}{p_t}\right)^{c/s}$$

giving

$$q = \delta_q \int_{p_o}^1 e^{-ct} dp = \delta_q\, p_0^{c/s} \int_{p_o}^1 \left(\frac{1-p_t}{p_t}\right)^{c/s} dp$$

For $c/s < 0.1$, the integral is close to one and we recover Equation 8.4a. For larger c/s, Barton (1998; Otto and Barton 1997) show that a more accurate result is given by

$$\Delta_q \simeq \delta_q\, p_0^{c/s}\ \left[\Gamma\left(1 + [c/s]\right)\right]^2 \Gamma\left(1 - [c/s]\right) \tag{8.5a}$$

where Γ denotes the gamma function (Equation 2.25b). For $c/s \ll 1$, this is approximately

$$\Delta_q \simeq \delta_q\, \left(1 + \frac{c}{s}\left[\ln(p_0) + 0.5772\right]\right) \tag{8.5b}$$

which offers a slight improvement over Equation 8.4c, but only when p_0 is not very small.

Reduction in Genetic Diversity

How much of a reduction in genetic variation does a sweep induce? As above, we continue to assume (for now) that any effect of mutation occurring *during* the sweep can be ignored (a point we address shortly). The first treatment of this topic, and one of the more widely-cited results on sweeps, is due to Kaplan et al. (1989). They showed that the expected coalescent time for two alleles at a neutral site linked to the site under selection differs significantly from $2N$ (the neutral value) when $c/s < 0.01$, and the sweep has been recent (fixation less than $0.2N$ generations ago, so that the effects of new mutations following fixation are negligible). This leads to their often-quoted approximation that ***neutral sites within 0.01 s/c of a selected site will be significantly influenced by a recent sweep***. The expected total length L of depressed variation associated with a recent sweep becomes

$$L = 0.02\,\frac{s}{c} \tag{8.6a}$$

where the factor of two arises because the influence extends on both sides of the sweep. Assuming c scales as one cM/Mb ($c = 0.01$ for each 10^6 bases), this approximation implies that a recent sweep with a selection coefficient of $s = 0.01$ is expected to influence variation in a region of size $0.02 \cdot (0.01/0.01) = 0.02$ Mb, or roughly 20 kb (Example 8.3 gives a more refined result). Likewise, a selection coefficient of $s = 0.1$ leaves an initial signature over a region of roughly 200 kb. Equation 8.6a can be used to obtain a rough estimate of s. Given the length L of decreased heterozygosity and a value of c for this interval,

$$s \simeq \frac{c \cdot L}{0.02} \tag{8.6b}$$

For example, if a sweep roughly covers 50 kb (or 0.05Mb) in a region where c is roughly 2cM/Mb, then an order of magnitude approximation of s is

$$s \simeq \frac{0.05 \cdot 0.02}{0.02} = 0.05$$

This is a crude approach, requiring a reasonable estimate of the size of the region influenced by the sweep, and a very recently completed sweep. Further, simulation studies have shown that ***sweeps can be asymmetric around the site under selection*** (Kim and Stephan 2002), reflecting the random location of those rare recombination events between $\mathbf{m}$ and the selected site that occur early in the sweep. Simply taking the middle of a region of depressed variation can be a poor approach for localizing the site under selection.

A more accurate expression for the expected fraction of variation remaining after a very recent sweep follows from the expected allele-frequency change (Equation 8.4a). Let q denote the initial frequency of allele $\mathbf{m}$ at a linked neutral marker, with $H_0 = 2q(1-q)$ denoting the initial heterozygosity, typically measured as the nucleotide diversity π, the average per-nucleotide heterozygosity (Chapters 2, 4). Hitchhiking during the fixation of a linked selected allele changes this to $q_h = q + \Delta_q$, and hence the heterozygosity becomes

$$\begin{aligned} H &= 2q_h(1-q_h) = 2(q+\Delta_q)(1-[q-\Delta_q]) \\ &= H_0 - 2(1-2q)\Delta_q - 2\left(\Delta_q\right)^2 \end{aligned} \tag{8.7a}$$

The expected heterozygosity is the average of H over two scenarios. With probability q, the favorable mutation arises on an $\mathbf{m}$ background, giving $q_A(0) = 1$, $\delta_q = 1 - q$, and

$\Delta_q \simeq (1-q)\, p_0^{c/s}$. Conversely, with probability $1-q$, the favorable alleles arises on a non-$\mathbf{m}$ background, giving $q_A(0) = 0$, $\delta_q = 0 - q = -q$, and $\Delta_q \simeq -q\, p_0^{c/s}$. Using these results, the expected allele frequency change is

$$\mathrm{E}(\Delta_q) = q \cdot (1-q)\, p_0^{c/s} + (1-q) \cdot \left(-q\, p_0^{c/s}\right) = 0 \tag{8.7b}$$

Using this result and taking the expectation of Equation 8.7a gives

$$H_h = \mathrm{E}(H) = H_0 - 2\mathrm{E}\,(\Delta_q)^2 \tag{8.7c}$$

where

$$\mathrm{E}\,(\Delta_q)^2 = q\left[(1-q)p_0^{c/s}\right]^2 + (1-q)\left[-q(p_0)^{c/s}\right]^2 \; = q(1-q)p_0^{-2c/s} \tag{8.7d}$$

Combining Equations 8.7c and d gives

$$H_h = H_0 - 2q(1-q)p_0^{-2c/s} = H_0\left(1 - p_0^{-2c/s}\right) \tag{8.8a}$$

Recalling that this results in an approximation (as Equation 8.4a approximates the allele frequency change), our final result is

$$\frac{H_h}{H_0} \simeq 1 - p_0^{2c/s} \simeq -\frac{2c}{s}\,\ln(p_0) \quad \text{for} \quad c/s \ll 1 \tag{8.8b}$$

As a first approximation to account for finite population size, we can improve on Equation 8.8b for a sweep starting from a single mutation by replacing $p_0 = 1/2N$ by $1/(4N_e s)$, giving

$$\frac{H_h}{H_0} \simeq 1 - (4N_e s)^{-2c/s} \tag{8.8c}$$

Stephan et al (1992) and Barton (1998) present more accurate (and complex) expressions for the reduction in heterozygosity in a finite population.

An alternative way to obtain Equation 8.8b is to consider the fraction f_s of the initial associations that persist when $\mathbf{A}$ is fixed (Equation 8.1c), as with probability f_s^2, neutral alleles at our site for two randomly-drawn chromosomes (under a catastrophic sweep) are identical-by-descent and hence (in the absence of mutation) homozygous. The reduction in heterozygosity at the neutral allele immediately following the fixation of $\mathbf{A}$ becomes

$$\frac{H_h}{H_0} = 1 - f_s^2 = 1 - p_0^{2c/s}. \tag{8.8d}$$

Equation 8.8d follows from Equation 8.4a, and hence assumes additive selection. When dominance is present (heterozygote fitness $1 + 2hs$ instead of $1 + s$), Equation 8.8b holds with $2hs$ replacing s (for $h > 0$). For a complete recessive ($h = 0$, fitnesses $1 : 1 : 1 + 2s$), Ewing et al. (2011) showed that

$$\frac{H_h}{H_0} \simeq \frac{\lambda}{1+\lambda}, \quad \text{where} \quad \lambda = \left(c/\sqrt{s}\right)\,\sqrt{4N_e}. \tag{8.8e}$$

As expected, a recessive sweep produces a much weaker signal, reflecting the greater chance for recombination given the much slower time to fixation ($\sim \sqrt{N_e/s}$ generations, Ewing et al. 2011). It is important to stress that Equation 8.8d and 8.8e all refer to the reduction in heterozygosity *immediately* following a sweep. This is the maximal signature, which begins to decay immediately as mutation rebuilds variation, an effect we examine shortly.

Example 8.3. Suppose a recombination rate of 1 cM/Mb (or 0.00001 per kb), and consider the expected reduction in heterozygosity at a site 10 kb away from a sweep ($c = 10 \cdot 0.00001 = 0.00010$). For an additive allele with $s = 0.01$ and $N_e = 10^6$, Equation 8.8b gives $H_h/H_0 \simeq 0.19$, so that (ignoring any new mutation) only 19% of the initial amount of heterozygosity is present immediately following a sweep. For a dominant allele, we replace $s = 0.01$ by $2s = 0.02$ in Equation 8.8b, giving $H_h/H_0 \simeq 0.10$. Finally, suppose the favored allele is recessive. Here,

$$\lambda = \left(c/\sqrt{s}\right)\,\sqrt{4N_e} = \left(0.0001/\sqrt{0.01}\right)\,\sqrt{4\times 10^6} = 2$$

and Equation 8.8c gives $H_h/H_0 \simeq 0.67$. Using the same parameters, the values for H_h/H_0 at different distances away from the selected site are as follows:

	1 kb	5 kb	10 kb	25 kb	50 kb	100 kb
Dominant	0.01	0.05	0.10	0.23	0.41	0.65
Additive	0.02	0.10	0.19	0.41	0.65	0.88
Recessive	0.17	0.50	0.67	0.83	0.91	0.95

The sweep from a dominant allele has the largest effect (roughly twice the reduction for small distances compared to additive selection), while the effect of a recessive allele is fairly weak except at very short distances from the site. For these three modes of gene action and $s = 0.01$, a 50% reduction ($H_h/H_0 = 0.5$) in heterozygosity occurs over a distance of 5 kb on either side of a selected recessive site, 31 kb when additive, and 66 kb when dominant, giving the size of the sweep regions as 10, 62, and 132 kb, respectively.

Finally, we can examine the accuracy of Kaplan and Hudson's approximation (Equation 8.6), which states that a sweep roughly influences a region of length $L/2 = 0.01s/c$ on either side of the selective site. We do so by using Equation 8.8b to find the value of c/s that results in a reduction in heterozygosity of at least 50% ($H_h/H_0 = 0.5$). Assuming a single copy at the start of selection,

$$\frac{2c}{s}\,\ln(2N) = 0.5, \quad \text{or} \quad \frac{c}{s} = \frac{0.25}{\ln(2N)} \tag{8.9}$$

The dependence on N is very weak. For example, for $N = 10^4$, the critical c/s value (which Kaplan and Hudson approximate as 0.01) is actually 0.025, while for $N = 10^9$, it is 0.012.

The Messer-Neher Estimator of s

As Equation 8.6b illustrates, one can manipulate any of the above expressions for reduction in H to obtain an estimate of s (given H_0 and c). In Chapter 9, this approach is placed into a more sophisticated likelihood framework where the entire spatial pattern of genetic variation (as a function of the distance c from a putative site) is used to estimate s. These approaches all require knowledge of the recombination rate c. How can we estimate s in situations with little, or no, recombination, such as on an organelle, or a very small chromosome, or in a highly-inbred species? A very creative solution to this problem, which equally applies when c is known, was presented by Messer and Neher (2012). As a favored allele is exponentially increasing in frequency during its initial sojourn to fixation, new mutations can appear on its initial haplotype. While rare, these mutations will still be more common that any appearing after fixation, as each hitchhikes up to some modest frequency. By considering a region around a putative size and counting the resulting haplotypes, Messer and Neher obtained a very simple approximate relationship between the frequency of the ordered haplotype classes. Letting n_0 denote the number of the most frequent haplotype, n_1 the next frequent and so forth, they found

$$\frac{n_i}{n_0} \simeq \left(\frac{\mu}{i\,s}\right)^{(1-\mu/s)} \simeq \frac{\mu}{i\,s} \tag{8.10}$$

Namely, a power-law relationship as a function of the region-wide mutation rate μ and the strength s of the favored site. For example, for $\mu/s = 0.02$, n_0 is 50 times more frequent than n_1 and 100 times more than n_2. Based on the relationship given by Equation 8.10, Messer and Neher develop regression-based estimator of s. While this approach works on genomic regions showing little to no recombination (the authors applied it to HIV), it also requires deep sequencing. If $\mu/s = 0.1$, using the frequencies of the five most common haplotypes $(n_0, \cdots, n_4)$ requires accurate estimates of frequencies with excepted value around 2%, implying a sample size of around 10^3 sequences. While c does not appear in Equation 8.10, obviously recombination can also generate novel haplotypes. Messer and Neher found that simply replacing μ by $\mu + c$ performs well if the ancestral diversity was high, but overestimates the rate of formation of new haplotype when this diversity was low. They found that simply pruning the collection of haplotype to ignore those clearly generated by recombination slightly underestimates s but otherwise performs rather consistently over c.

Recovery of Variation Following a Sweep

The signal left by even a strong sweep is a transient one, as new mutation will eventually restore heterozygosity at the neutral site back to its equilibrium value ($H_0 = 4N_e\mu$) before the sweep. Kim and Stephan (2000) find that the expected heterozygosity t generations after a sweep is approximately

$$E[\,H(t)\,] \simeq H_0 \left(1 - (4N_e s)^{-2c/s} \cdot e^{-t/(2N_e)}\right) \tag{8.11}$$

where $-H_0(4N_e s)^{-2c/s} = -H_0\, f_s$ is the reduction immediately following the conclusion of the sweep. Mutation following the cleansing sweep recovers variation, and this can be envisioned as a decay in the initial reduction $-H_0\, f_s$, eventually driving this value back to zero (and hence full variation). Note from Equation 8.11 that the initial reduction decays by an amount $1/(2N_e)$ each generation, as $(1 - 1/2N_e)^t \simeq \exp(-t/2N_e)$. The expected time to recover half the variation lost during the sweep (its half-life) is $\exp(-t_{0.5}/2N_e) = 0.5$ or $t_{0.5} = -2\ln(0.5)N_e \simeq 1.4N_e$. Note the important result that the ratio $E[\,H(t)\,]/H_0$ is *independent* of the actual mutation rate μ. The reason is that a low (or high) mutation rate means both a slow (or fast) accumulation of new mutations following the sweep, but a low (or high) target heterozygosity to reach.

Effects of Sweeps on the Variance in Microsatellite Copy Number

The above results for the behavior of nucleotide diversity (heterozygosity) during and after a sweep apply to SNP data. Since per-nucleotide mutation rates are very low (Chapter 4), the infinite-sites model offers a good approximation for such data, as back mutations are unlikely and mutations are rare in general, so that the role of recurrent neutral mutation during the sweep can largely be ignored. Both of these assumptions are violated when microsatellite (STR, simple tandem repeat) markers are considered. These have high mutation rates (on the order of 10^{-2} to 10^{-4}) and recurrent mutation can regenerate the same allele (scored in STRs by the number of repeats at a site). Further, when dealing with STR data, a common measure of variability is not heterozygosity but rather the variance V in copy number among alleles at the microsatellite marker.

The behavior of V during a sweep was examined by Wiehe (1998), using a simple stepwise mutation model (an STR allele of length k has equal probability of changing to length $k+1$ or $k-1$). If V_0 denotes the initial variance in copy number, its expected value V_h immediately following the sweep has a very similar form to Equation 8.8b,

$$\frac{V_h}{V_0} = 1 - \beta \cdot p_0^{2c/s} \tag{8.12a}$$

The difference being a scaling factor $\beta < 1$, which discounts the removal of variation by the sweep by the continual input from new mutation. Wiehe showed that when the total mutation rate scales with allele length ($k\mu$ is the rate of an allele of length k), β has a closed solution,

$$\beta = p_0^{4\mu/s} \tag{8.12b}$$

which reflects the relative strengths of mutation and selection (akin to recombination versus selection) during the sweep, giving

$$\frac{V_h}{V_0} = 1 - p_0^{(4\mu+2c)/s} \tag{8.12c}$$

When $4\mu + 2c > s$, little depression in the copy-number variance following a sweep is expected, as mutation rates are sufficiently high that new STR alleles are generated at a high rate even as the sweep is occurring, so that even the fixation of a single original haplotype ($c = 0$) will still show significant variation.

Using Slatkin (1995b), the rate of recover in V following the sweep is a modification of Equation 8.11,

$$V(t) = V_0 \left(1 - p_0^{(4\mu+2c)/s} \cdot e^{-t/(2N_e)}\right) \tag{8.12d}$$

As with Equation 8.11, $t_{0.5} \simeq 1.4N_e$ generations is the time to recover half of the decrease in V immediately following the bottleneck. It is often stated that microsatellites recover faster from a sweep because of their high mutation rates. This is due to mutations arising *during* the sweep, as the time to recover following the sweep (the time to decay the reduction present immediately following the sweep) is independent of the mutation rate.

The Site-Frequency Spectrum

Recall that Chapter 2 introduced the concept of a frequency spectrum, the expected distribution of the frequencies of different alleles or sites in a sample. In particular, the site-frequency spectrum $\phi(x)$ gives the expected frequency of sites having frequency x for the derived (most recent) allele. Under the equilibrium neutral model, this is given by the Watterson distribution (Equation 2.34a), with most sites having very low frequencies of the derived allele. As shown in Figure 8.5, a sweep transforms the (unfolded) site-frequency spectrum of derived alleles from the L-shaped Watterson distribution to a more U-shaped one (Fay and Wu 2000; Kim and Stephan 2002), resulting in an ***excess of sites with high-frequency derived alleles*** and also ***an excess of sites with rare alleles***. If considering the folded frequency spectrum (the spectrum over (0, 0.5) associated with the minor allele frequency), these result in an increase in the fraction of sites with rare minor allele frequencies. Przeworski (2002) showed that both features in the unfolded spectrum are present immediately following a sweep, but that the excess of sites with high-frequency derived alleles rapidly dissipates (within $0.2N_e$ generations) as they become fixed. The excess of rare alleles persists a bit longer (roughly

$0.5N_e$ generations), as it is sensitive to new mutations generating rare alleles immediately after the sweep.

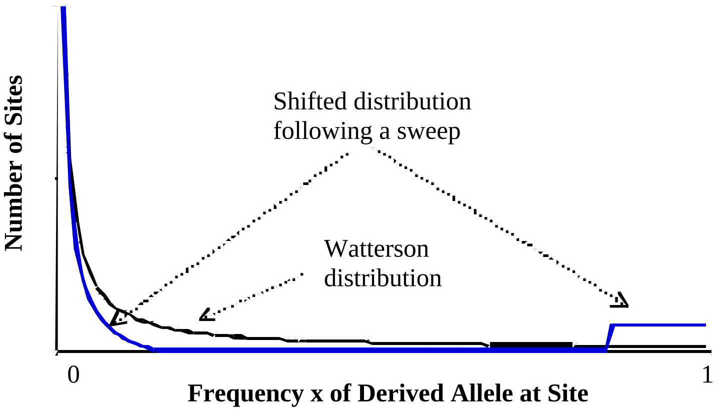

Figure 8.5. The effect of a hard sweep on the unfolded site-frequency spectrum of derived alleles. Under the equilibrium neutral model, this distribution is hyperbolic (Equation 2.24a), an L-shaped curve that is monotonically declining, with most derived alleles being at low frequencies. The effect of a sweep is to shift some derived alleles to very high frequencies, while shifting the others to frequencies near zero, resulting in a more U-shaped distribution.

To see how this transformation occurs, consider a particular site where the derived allele has frequency x before a sweep. Assume that the site-frequency spectrum before the sweep follows the Watterson distribution $\phi(x) = (\theta/x)dx$ (Equation 2.34a), which requires that the equilibrium-neutral conditions hold (Chapter 2), and $\theta = 4N_e\mu$ refers to per-nucleotides values. Assuming the sweep initiates from a single favorable mutation, then with probability x it is initially associated with the derived allele at our linked site, increasing its frequency from x to $f_s + x(1 - f_s)$ (Equation 8.1d). Else, with probability $1 - x$, the favorable allele is associated with the ancestral allele, decreasing the derived-allele frequency from x to $x(1-f_s)$. To visualize the transformation of the frequency spectrum from these two different events, decompose the site-frequency spectrum as $x\phi(x)dx + (1-x)\phi(x)dx = \theta dx + \theta(x^{-1} - 1)dx$. The first piece ($\theta dx$) corresponds to a uniform distribution (a constant for all values of x) over the range $f_s \leq x \leq 1-1/(2N)$. This range follows as f_s is the resulting frequency of a derived allele near zero at the start of the sweep, while the upper limit for a segregating site is $1-1/(2N)$. Conversely, when the favorable mutation is associated with the ancestral copy, the distribution of sites originally with frequency x is down-shifted to $\theta(x^{-1}-1)dx$, which is now associated with a frequency of $x(1-f_s)$, and has resulting range of $1/(2N) \leq x \leq 1-f_s$. The middle range of the transformed frequency spectrum ($1 - f_s < x < f_s$) essentially is zero. Putting all of these together, Fay and Wu (2000) approximate the resulting sweep-transformed site-frequency spectrum as

$$\phi(x) = \begin{cases} \theta\left(\dfrac{1}{x} - \dfrac{1}{1-f_s}\right), & \dfrac{1}{2N} \leq x \leq 1 - f_s \\ 0, & 1 - f_s < x < f_s \\ \dfrac{\theta}{1-f_s}, & f_s \leq x \leq 1 - \dfrac{1}{2N} \end{cases} \tag{8.13}$$

If two *concurrent* sweeps are influencing the same region, the resulting site-frequency spectrum is rather different from the pattern for a single hard sweep. Simulations by Chevin et al. (2008) found an *excess of immediate frequency alleles* in such cases, mimicking the signature of balancing selection. However, they also generate both an excess of high-frequency derived alleles and a deficiency of low-frequency alleles. The combination of these three features seems unique to concurrent sweeps.

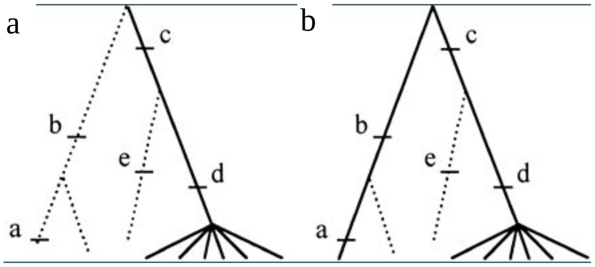

Figure 8.6. The genealogy of a sample of alleles following a selective sweep. Solid branches represent sampled alleles, while dotted lines indicate lineages lost due to the fixation of the favorable allele. **A:** In the absence of recombination, lineages not initially associated with the favorable mutation are lost. Here all sequence contain the derived c and b alleles, and there is a star phylogeny for the surviving sequences. **B:** When recombination occurs, other lineages may become associated with the favorable allele, resulting in the MRCA for some sequences being much deeper (earlier) than the start of the sweep. Here a single recombinant is present in the sample, so that c and d are high-frequency derived alleles, while b and a are at low-frequencies. After Fay and Wu (2000).

Recombination and the Genealogical Structure

As shown in Figure 8.3, a sweep changes both the size and shape of the genealogy of linked neutral alleles. In particular, many of the alleles are sampled from an approximately star genealogy, with the nodes of the coalescent being very compressed, so that the pattern resembles a radiation from a single point, namely the start of selection (Figure 8.6A). One consequence of a star-like phylogeny is that mutations following the start of selection generate an excess of rare alleles, as they are confined to one or a few external branches of the genealogy of the sampled alleles. As a consequence, even after a sweep is finished, mutation will still generate an excess of rare alleles during the recovery of the background variation around the selected site.

Recombination also has an important impact on the genealogy, especially when the favorable haplotype is still rather rare. In this setting, most recombination events involving this haplotype will be with other lineages not carrying the favorable allele. This results in the favorable allele being transferred across lineages, generating sites near the sweep with alleles whose coalescent times predate the start of the sweep (e.g., Figure 8.6B).

The Pattern of Linkage Disequilibrium

The pattern of linkage disequilibrium (LD) generated by a sweep has been extensively studied (Thomson 1977; Gillespie 1997; Przeworski 2002; Kim and Nielsen 2004; Stephan et al. 2006; McVean 2007; Jensen et al. 2007; Pfaffelhuber et al. 2008), and turns out to be both complicated and surprising (Figure 8.7). The conventional wisdom has been that a selective sweep increases LD around the site of selection (Thomson 1977; Przeworski 2002), with the increase in LD *during* a sweep offering a signal for selection (Chapter 9). Starting with Kim and Nielsen (2004), it was realized that the spatial and temporal patterns in LD associated with a sweep are far more subtle.

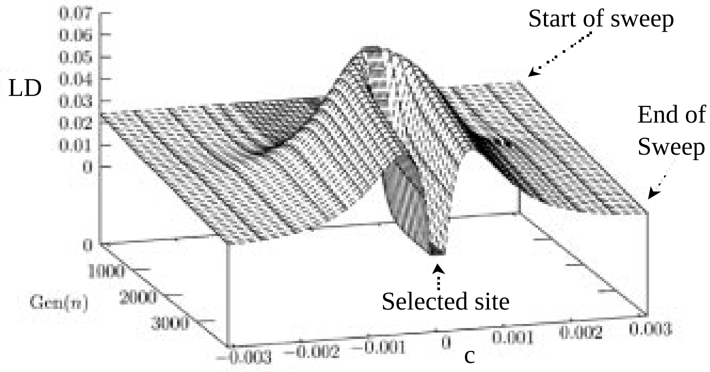

Figure 8.7. The dynamics of linkage disequilibrium around a selected site during the time course of a sweep (which starts at generation 0). One sees a strong signal of LD *across* the site during the early phase of the sweep (the partial sweep stage), but *little to no* LD across the site upon fixation. This 3D figure plots the spatial pattern of expected LD under a deterministic model of selection whose position corresponds to $c = 0$, with the more distant slices (those towards the back of the graph) representing older patterns. Along any one slice, the curve plots the expected LD between the target of selection and a linked site at distance c. Initially, a sweep results in a sharp increase in LD in a region through the selected site. However, as the favorable allele reaches intermediate frequency, the LD immediately adjacent to the site starts to decay, while LD on either side largely remains intact. Upon fixation (the forward-most slice), the result is very little LD at the site (often below the starting background) which is flanked by strong regions of LD on either side. As a deterministic analysis, this graph represents the average behavior over a large number of identical sweeps. Any particular realization will be far noisier. After Stephan et al. (2006).

While LD does indeed increase during the early phase of the sweep of a favorable allele to fixation, it actually starts to *decrease* around the site once the frequency of the favorable allele reaches roughly 0.5 (Stephen et al. 2006). Upon fixation, the result is a region tightly linked around the sweep that has an LD level *lower* than the background level at unlinked neutral loci, and hence potentially reduced from its initial starting value. Conversely, on either side of the selective site, LD significantly increases, so that strong LD can be found on the left and/or right sides of a selected site, with no association *across* the site – LD between sites to the left and to the right of a sweep is close to zero. Thus, a recently-completed sweep potentially leaves a very unusual spatial pattern in LD, with a plot of LD showing peaks on either side of the selected site, surrounded by a valley of little LD at the actual site itself (Figure 8.7). Further, while LD is inflated around the sides of a selected site, it can actually be slightly *decreased* at sites of intermediate distance (McVean 2007).

The plot in Figure 8.7 is based on a deterministic analysis of a three locus model (one selected, two neutral) by Stephan et al. (2006). As such, it depicts a very smooth and symmetric view of the LD on either site of the selected site, representing the *average* behavior over a large number of identical sweeps. In reality, there is considerable variance in the amount of LD due to finite population size, the stochastic location of rare recombination events, and differences in allele frequencies across markers at the start of the sweep. Simulation studies (e.g., Kim and Nielsen 2004) often find a very asymmetric pattern of LD across a selected site, with a strong signal on one side and little to no signal on the other.

This unusual pattern of LD around the sweep has a genealogical explanation (McVean 2007). Early on in a sweep, strong LD is expected because of the rapid increase of the favorable haplotype. During this phase, there is some chance that the favorable allele will recombine into other haplotypes, with these rare recombination events transfering the favorable allele to other backgrounds (e.g., Figure 8.6b), generating a few new haplotypes (containing alleles that are segregating prior to the start of the sweep) also associated with the favorable allele. As these new haplotypes are also swept along, they result in blocks of LD as $\mathbf{A}$ approaches fixation. Recombination events on either side of the sweep are independent, and hence do not create LD *across* the region. However, either following (or even during) the sweep, new mutations can arise. Because these are at low frequency, they generate only small amounts of LD, but as neutral alleles present before the sweep become fixed (the fixation of high-frequency derived alleles), these new segregating loci contribute the bulk of the low levels of LD seen. The role of new mutations appearing after the start of the sweep on LD is especially important in areas adjacent to the selected site where little to no recombination has occurred during the fixation of the favorable allele.

Age of a Sweep

A number of workers have considered various estimates of the time since the start of a sweep, typically under the assumption of a catastrophic sweep (a single copy of a new mutation is swept to fixation), no recombination, and a negligible amount of mutation at neutral sites during the sweep (Wiehe and Stephan 1993; Perlitz and Stephan 1997; Jensen et al. 2002; Enard et al. 2002; Przeworski 2003; Li and Stephan 2005, 2006). The simplest estimate follows from the infinite-sites model (Chapter 2). Assume S segregating sites are observed in a sample of n sequences for a nonrecombining region around the site of a sweep. Under the infinite-sites model, the expected number of segregating sites in a sample is $E(S) = \mu T_n$, where μ is the total mutation rate over the entire region of interest and T_n is the total branch length of the entire genealogy of the sample. Under a catastrophic sweep that started τ generations ago, the coalescent tree has its nodes sharply compressed, and can be approximated by a star phylogeny. In this case, the total branch length is $n\tau$ (as the length along each of the n branches is roughly τ), giving $\mu n\tau$ as the expected number of segregating sites, and leading to a simple method-of-moments estimator of the time τ,

$$\widehat{\tau} = \frac{S}{\mu\, n} \tag{8.14}$$

More sophisticated approaches for estimating τ are discussed in Chapter 9.

Example 8.4. Akey et al (2004) found a 115-kb region on human chromosome 7 showing signatures of a sweep: excess rare alleles, excess high-frequency derived alleles, and a reduction in nucleotide diversity. Eleven segregating sites were found in a sample of 45 African- and European-Americans chromosomes. Assuming a mutation rate of 10^{-9} per site per generation, the total mutation rate over the entire region is $115,000 \cdot 10^{-9} = 0.000115$ per generation, giving

$$\widehat{\tau} = \frac{11}{0.000115 \cdot 45} = 2126 \text{ generations}$$

Assuming a generation time of 25 years for humans, this translates into 53,140 years. Example 9.14 shows how confidence intervals are obtained under this model.

Geographic Structure

All our analyses thus far have assumed a panmictic population. One simple consequence of population structure is that hard sweeps from independent mutations in different subpopulations generate a soft-sweep signature when the data is combined (Ralph and Coop 2010; Messer and Petrov 2013).

While there has only been preliminary analysis of the effect of geographic structure (Slatkin and Wiehe 1998; Santiago and Caballero 2005), it is clear that it can be dramatic. For example, Santiago and Caballero consider a simple two-subpopulation model, with weak migration. As expected, a sweep fixing a favored allele in one subpopulation results in a decrease in variation around the selected site in that subpopulation. However, it can also result in an *increase* in the variation around that site in the second subpopulation following the spread and fixation of the favorable allele. In effect, the sweep and subsequent migration has the effect of transforming some between-population variation into within-population variation. The net result is that diversity in one subpopulation *increases* for a short distance as one moves away from the site, and also shows an excess of sites with immediate allele frequencies, mimicking signatures for balancing selection. Finally, while a sweep restricted to one subpopulation can result in increased between-population divergence in allele frequencies (increasing F_{st}), Santiago and Caballero also found that a sweep can often *reduce* F_{st}. Clearly, models incorporating sweeps in structured populations are an important future research area (Stephan 2010a).

Table 8.1. Summary of various features associated with a selective sweep of a favorable allele $\mathbf{A}$ with fitnesses $1 : 1 + 2hs : 1 + 2s$ (for $h \neq 0$). Let q denote the frequency of a neutral marker at the start of selection at distance (recombination fraction) c from a strongly selected site ($4N_es \gg 1$). Assume the frequency of the favorable allele is p_0 at the start of selection, and let q_h and H_h denote the final frequency for a neutral allele initially associated with $\mathbf{A}$ and the heterozygosity at a neutral site immediately following the sweep. V refers to copy-number variation at an STR, and $\beta < 1$ is a function of the STR mutation rate (e.g., Equation 8.12b).

Fraction f_s of initial associations remaining at fixation:

$$f_s \simeq \begin{cases} p_0^{-c/(2hs)} \simeq 1 - \dfrac{c}{2hs}\,\ln(p_0) & \text{for } p_0 \gg 1/(2N_es) \\[2ex] (4N_es)^{-c/(2hs)} \simeq 1 - \dfrac{c}{2hs}\,\ln(4N_es) & \text{for } p_0 = 1/(2N) \end{cases}$$

Total change in the frequency of a linked neutral allele: $\Delta_q \simeq (1-q)f_s$

Final frequency of a linked marker: $q_h = q + \Delta_q = f_s + q(1-f_s)$

Reduction in heterozygosity immediately following the sweep: $\dfrac{H_h}{H_0} = 1 - f_s^2$

Heterozygosity t generations after sweep completed: $\dfrac{H(t)}{H_0} = 1 - f_s^2\, e^{-t/(2N_e)}$

Reduction in STR copy-number variation immediately following the sweep: $\dfrac{V_h}{V_0} = 1 - \beta\, f_s^2$

STR copy-number variation t generations after a sweep: $\dfrac{V(t)}{V_0} = 1 - \beta\, f_s^2\, e^{-t/(2N_e)}$

Summary: Signatures of a Hard Sweep

The key summary parameter for the potential impact of a sweep is the fraction $f_s = \Delta_q/\delta_q$ of original haplotypes that stay intact following a sweep. If $f_s \simeq 1$, the sweep has a major impact on the structure of variation at neutral sites, while if $f_s \simeq 0$, it has essentially no impact.

Table 8.1 summarizes both expressions for f_s and the population-genetic impact of a sweep on a linked neutral site. Table 8.2 summarizes more subtle signatures of a sweep beyond the simple reduction in variation. As detailed in the next chapter, all of the observations listed in Table 8.2, either singularly or in combination, have been used as the basis of tests of ongoing/recent selection. It is important to stress that the results in these two tables are *restricted to hard sweeps*, wherein the favorable allele is only present as (at most) a few copies at the start of selection. As is now shown, under soft sweeps, many of these signals are either muted or washed out entirely.

Table 8.2. Population-genetics theory predicts the following patterns associated with a hard sweep:

A recent or ongoing sweep leaves several potentially diagnostic signals:

(1) *An excess of sites with rare alleles (in either the folded or unfold frequency spectrum)*

(2) *An excess of sites with high frequency* derived *alleles in the unfold frequency spectrum*

(3) *Depression of genetic variation, often* asymmetrically, *around the site of selection*

Signatures in the spatial pattern of LD differ during the sweep and after its completion:

(4a) When a favorable allele is at moderate frequencies (a partial sweep), we see *an excess in LD throughout the region surrounding the sweep*

(4b) Following fixation of the favorable allele, we see *an excess in LD on either side of the site, but a* depression *in LD around the site*

Finally,

(5) *Signatures of a sweep are very fleeting, remaining on the order of* $0.5N_e$ *generations for signature (1),* $0.4N_e$ *generations for (2),* $1.4N_e$ *generations for (3), and* $0.1N_e$ *generations for (4b)*

SOFT SWEEPS AND POLYGENIC ADAPATION

While a hard sweep starts with selection on a single haplotype, a soft sweep refers to situations where, at the start of the sweep, multiple haplotypes contain the favored allele (Figure 8.4). Under a single-origin soft sweep, a single copy of the mutation arises in an environment that does not yet favor it, drifting around before an environmental change places all of the haplotypes associated with it under selection. Under a multiple-origin soft sweep, the favored allele consists of a collection of *independent* origins. These independent copies can arise in standing variation before the allele became favored and/or can arise *during* the sojourn to fixation for this allele. Finally, one can have **polygenic adaptation** (Pritchard and Di Rienzo 2010; Pritchard et al. 2010) occurring through the fixation of a large number of alleles of much smaller effect throughout the genome. In the extreme, adaptation occurs by modest allele-frequency change (as opposed to fixation), resulting in partial weak sweeps over a large number of loci, leaving essentially no signature in the neutral background variation around the selected polygenes.

Sweeps Using Standing Variation

The hard sweep model implies a lag in adaptation, with populations experiencing a new environmental challange having to wait for favorable mutations to arise in order to respond.

Conversely, artificial selection for just about any trait in an outbred population generates an immediate response (Chapter 18), showing that a large reservoir of **standing** (or preexisting) **variation** exists for most traits. Thus, hard sweeps are expected to be more frequent when standing variation is likely to be small, such as inbred or small populations. However, in outbred populations that suddenly experience a new environment, much of the initial response might arise from standing variation, although new mutations can also play a critical role in the continued response once this initial variation is depleted (Chapter 26).

Example 8.5. The threespine stickleback (*Gasterosteus aculeatus*) is a species (or species complex) of small fish widespread throughout the Northern Hemisphere in both freshwater and marine environments. The marine form is usually armored with a series of over 30 bony plates running the length of the body, while exclusively freshwater forms (which presumably arose from marine populations following the melting of the last glaciers) often lack some, or all, of these plates. Given the isolation of the freshwater lakes, it is clear that the reduced armor phenotype has independently evolved multiple times. Colosimo et al. (2005) showed that this parallel evolution occurred by repeated fixation of alleles at the *Eda* gene involved in the ectodysplasin signaling pathway. Surveying populations from Europe, North America, and Japan, they found that most nuclear genes showed a clear Atlantic/Pacific division. Conversely, at the *Eda* gene, low armored populations shared a more recent history than full-armored populations, independent of their geographic origins, presumably reflecting more recent ancestry at the site due to the sharing of a common allele. In marine populations, low-armored alleles at *Eda* are present a low (less than five percent) frequency. Presumably, these existing alleles were repeatedly selected following the colonization of freshwater lakes from marine founder populations. Barrett and Schluter (2008) and Messer and Petrov (2013) review a number of other examples of adaptation from pre-existing mutations.

The molecular signature resulting from a sweep using standing variation has been examined by Innan and Kim (2004) and Przeworski et al. (2005). Innan and Kim were interested in domestication, clearly a radical change in the environment to a new selection regime. As might be expected, the reduction in diversity is much less than for a hard sweep, because the time to most recent common ancestor for the favorable allele significantly predates the start of selection. Innan and Kim found that if the frequency of the allele at the start of selection was greater than five percent, at best only a weak reduction in background variation is generated by a sweep. However, domestication usually involves a strong bottleneck, which can result in a preexisting allele being reduced to one (or a very few) lineages that survived the bottleneck before being selected, generating a more hard-sweep pattern. A potential example of this is the maize domestication gene *tb1*. While this locus shows a classic hard-sweep pattern (Figure 8.2), the domestication allele was created by an insertion of the retroposon *Hopscotch* that predated domestication by at least 20,000 years (Studer et al. 2011). Jaenicke-Després et al. (2003), using DNA from ancient maize samples, suggested that selection started to occur on *tb1* roughly 4,400 years ago.

Przeworski et al. also found a critical dependence on the initial frequency, suggesting that as long as it was below $1/(4N_e s)$, the signal was the same as for a hard sweep. With higher initial allele frequencies, the situation is more complex. In some settings, the result is simply a weaker footprint but with the normal features of a sweep (reduced diversity, excess of rare alleles, excess of high-frequency derived alleles). However, in some cases, a weak sweep can result in an *excess* of *immediate* frequency alleles. In still other settings, essentially no detectable pattern is seen in the reduction of diversity, the frequency spectrum, or the distribution of LD. In particular, if the new environment favors an *ancestral* allele, especially one at high frequency, there will be no discernible change over the background pattern (Przeworski et al. 2005). The salient point is that selection on standing variation need not leave a hard-sweep signature, and significant ongoing/recent selection can easily be missed, even with strong selection.

How Likely is a Sweep Using Standing Variation?

Both Hermisson and Pennings (2005) and Przeworski et al. (2005) used population-genetic models to examine the likelihood of a sweep from standing variation. To consider the probability of such an event over a series of replicate populations, suppose $\phi(x)$ denotes the distribution for the frequency x for the soon-to-be favored allele **A**, and $U(x)$ its probability of fixation under the new environment given x. The probability $\Pr_{sv}$ that a sweep occurs using standing variation at this locus is simply

$$\Pr_{sv} = E\left[U(x)\right] = \int_{1/(2N)}^{1-1/(2N)} U(x)\phi(x)dx \tag{8.15a}$$

The limits on the integral confine us to considering only segregating alleles. Przeworski et al. (2005) assumed $\phi(x)$ was given by the neutral Watterson distribution (Equation 2.34a), while Hermisson and Pennings (2005) considered a more general setting, where the genotypes $aa : Aa : AA$ have fitnesses of $1 : 1-2h_d s_d : 1-2s_d$ in the old environment and $1 : 1+2hs : 1+2s$ in the new. This allows for the allele to be either neutral ($s_d = 0$) or deleterious ($s_d > 0$) before being favored. Assuming selection-drift-mutation equilibrium on the allele prior to it become favored, $\phi(x)$ is a function of N_e, the selection parameters (h_d, s_d), and the mutation rate μ_b to this allele, and can be obtained using diffusion machinery (Chapter 7, Appendix 1). Likewise, the fixation probability under the new fitnesses can also be obtained using diffusion results (Chapter 7). Putting these together, Hermisson and Pennings find that

$$\Pr_{sv} \approx 1 - e^{-\theta_b \, \ln(1+R)}, \quad \text{where} \quad R = \frac{2h\alpha_b}{2h_d\alpha_d + 1} \tag{8.15b}$$

with $\alpha_b = 4N_e s$ and $\alpha_d = 4N_e s_d$ are the scaled strengths of selection in the new and old environments, respectively, and $\theta_b = 4N_e\mu_b$ the scaled mutation rate.

The alternate scenaro to starting a sweep from standing variation is to wait for new favorable mutations to first arise and subsequently become fixed. Recall from Chapter 7 that the fixation probability of a single new mutation is roughly $4hs(N_e/N)$, so roughly $N/(4N_e hs)$ such mutations must appear to have a reasonable chance of one becoming fixed. The expected number of such beneficial mutations that arise each generation is $2N\mu_b$, giving

$$[4hs(N_e/N)]\,[2N\mu_b] = 2hs(4N_e\mu_b) = 2hs\theta_b \tag{8.16a}$$

as the expected number of destined-to-become-fixed mutations that arise each generation. Before proceeding, it is useful to consider the number τ of generations on a scale of $T = \tau/(2N_e)$, so that $\tau = 2N_eT$, with $T = 1$ corresponding to $2N_e$ generations. On this scale, the expected total number of beneficial mutations that have appeared by time T is $T \cdot 2N_e \cdot 2hs\theta_b = Th\alpha_b\theta_b$. Hence, the probability that at least one favorable mutation destined to become fixed appears by generation T is just one minus the probability that none do, which from the Poisson is

$$\Pr_{new}(T) = 1 - e^{-Th\alpha_b\theta_b} \tag{8.16b}$$

as obtained by Hermisson and Pennings (2005). When $\alpha_b\theta_b$ is small, the waiting time for a destined-to-become-fixed mutation is quite long. In such cases, mutation is the rate limiting step for adaptation. For example, suppose that adaptation can only occur through mutation at one of five nucleotide sites, generating an additive allele ($h = 1/2$) with a selective advantage of one percent ($s = 0.01$). In humans, assuming a historical value of $N_e = 50,000$ and a per site mutation rate of 10^{-9}, we have $h\alpha_b = (1/2)4N_e s = 2 \cdot 5 \times 10^4 \cdot 0.01 = 1000$, while $\theta_b = 4N_e\,\mu_b = 4 \cdot 5 \times 10^4 \cdot [5 \cdot 10^{-9}] = 0.001$, giving $h\alpha_b\theta_b = 1000 \cdot 0.001 = 1$. Solving

$$0.5 = 1 - e^{-T_{0.5} h\alpha_b\theta_b} = 1 - e^{-T_{0.5}},$$

gives $T_{0.5} = 0.69$ or $0.69(2N_e) = 1.38N_e = 69,000$ generations. Further, once such a destined-to-become fixed mutation arises, it still takes (on average) $2\ln(4N_e s)/s$ generations (for an additive allele) to become fixed (Equation 8.3d), which is roughly an additional 660 generations. More generally, the total waiting time until the *fixation* of a favorable (additive) allele (in generations) is approximately

$$t_{fix} = \frac{1}{s\,\theta_b} + \frac{2\ln(4N_e s)}{s} = s^{-1}\left[\theta_b^{-1} + \ln(4N_e s)\right], \tag{8.16c}$$

where the first term is the mean waiting time for the first appearance of a successful mutation and the second its fixation time. Karasov et al. (2010) develop a similar expression.

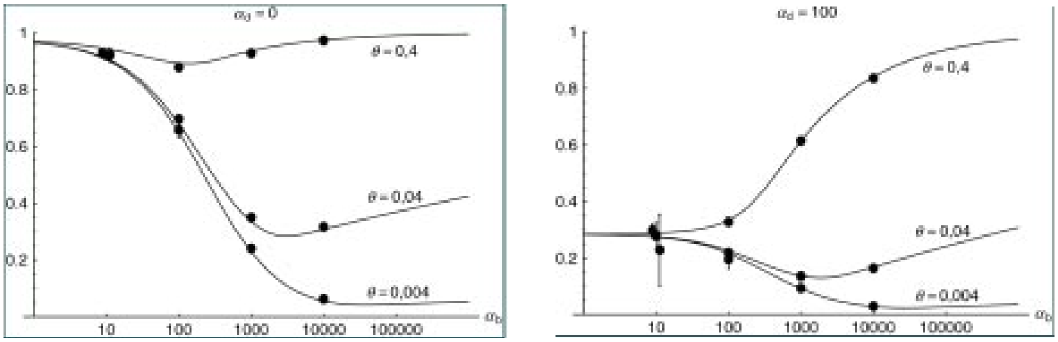

Figure 8.8. Plots of the probability (vertical axis) of a selected sweep from standing variation, *given* that a sweep has occured by $0.1N_e$ generations since the change in the environment (Equation 8.17). This is a function of the beneficial mutation rate θ_b (separate curves within each graph) and the scaled strength of selection α_b (horizontal axis). **Left:** The allele is neutral in the old environment ($\alpha_d = 0$). **Right:** The allele is deleterious in the old environment ($\alpha_d = 100$). After Hermisson and Pennings (2005).

If we *condition* on a sweep occuring, the probability $P_{ex}(T) = \Pr(\text{existing} \,|\, \text{Sweep by generation } T)$ it is from an existing allele is

$$P_{ex}(T) = \frac{\Pr_{sv}}{\Pr_{sv} + (1 - \Pr_{sv})\Pr_{new}(T)} = \frac{1 - \exp\left[-\theta_b\,\ln(1+R)\right]}{1 - \exp\left\{-\theta_b[\ln(1+R) + Th\alpha_b]\right\}} \tag{8.17}$$

which follows because $\Pr_{sv}$ is the probability that, in the absence of any mutation, a variant segregating at the start of the new selection regime is fixed, while the probability that the fixation occurs via a new mutation (arising by time T) is $(1 - \Pr_{sv})\Pr_{new}(T)$, the first term accounting for the probability that no segregating variant is fixed. For sufficiently large T, $\Pr_{new}(T) = 1$ and Equation 8.17 reduces to $\Pr_{sv}$ (Equation 8.15b), which sets the *lower limit* on the probability that a fixed favorable mutant was preexisting in the population before the start of selection. Figure 8.8 plots Equation 8.17 at $0.1N_e$ generations ($T = 0.05$) after an environmental shift. When both θ_b and α_b are high, most sweeps are from existing variation. This is true even when the allele is deleterious before the shift. When θ_b is small, most sweeps are from new mutations unless α_b and α_d are both small. The reason is that adaptation is unlikely with small α_b, and most of the adaptation that occurs results from alleles at relatively high frequency (and hence α_d small) before the start of selection.

Example 8.6. Suppose $N_e = 10^6$ and the per-site mutation rate throughout the genome is $\theta = 4N_e\mu = 0.01$. For a beneficial mutation that can only occur by a change to a specific nucleotide at a specific site, 1/3 of mutations at that site are beneficial, giving $\theta_b = 0.0033$. For an additive allele ($h = 1/2$) with $s = 10^{-4}$, we have $\alpha_b = 4 \cdot 10^6 \cdot 10^{-4} = 400$. If this mutation was neutral before being favored, $\alpha_d = 0$, $R = 2h\alpha_b = 400$ and Equation 8.15b gives

$$\Pr_{sv} \approx 1 - e^{-\theta_b \, \ln(1+R)} = 1 - e^{-0.0033 \, \ln(1+400)} = 0.013$$

Hence, there is only a 1% chance that a sweep occurs at this locus in the absence of new mutation. Now suppose that we examine this population at $T = 0.5$ (N_e generations). The probability that at least one such mutation destined to become fixed arises by this time is

$$\Pr_{new}(T) = 1 - e^{-Th\alpha_b\theta_b} = 1 - e^{-0.5 \cdot (1/2) \cdot 400 \cdot 0.0033} = 0.281$$

Provided we see a sweep at this locus by N_e generations, the probability it was due to an existing allele present at the time the environment shifted is

$$\pi_{ex} = \frac{\Pr_{sv}}{\Pr_{sv} + (1 - \Pr_{sv})\Pr_{new}(T)} = \frac{0.013}{0.013 + (1 - 0.013)0.281} = 0.05,$$

giving only a five percent chance that the fixed favorable allele was present in the population at the start of selection.

Peter et al. (2012) developed an ABC-based approach (Appendix 3) that combines several tests of selection in an attempt to distinguish between sweeps from *de novo* mutation and standing variation. Their key idea is that an adaptation from standing variation involve alleles that experienced drift before becoming favored, while a *de novo* mutation was always favored. The fit of several summary statistic of selection (Chapter 9) is then compared with various models assuming drift (or no drift) at some point in the history of that allele. Simulation studies showed that the test had little power unless selection was strong ($4N_es > 100$). When applied to seven of the strongest known sweeps in human, two genes most likely were from standing variation, and three more fit the hard sweep model. The other two were equivocal. It should be noted that the strong sweep signal used to ascertain these candidates biases this set towards hard sweeps.

Recurrent Mutation of the Favorable Allele Cannot be Ignored

In their analysis of the effects of sweeps from standing variation, both Innan and Kim (2004) and Przeworski et al. (2005) assumed a *single origin* of the favorable mutation. Likewise, while the analysis leading to Equation 8.17 does consider recurrent mutation, it simply allows new copies of the favorable allele to arise by mutation once selection starts and keeps track of how long one must wait until a destined-to-be fixed copy arises. It ignores any ongoing mutation either during the fixation of a pre-existing copy of the favorable allele or following the introduction of a favorable allele that is destined to become fixed. Further, independent mutations with similar phenotypic effects can arise, and these can interfere with each other. In a geographically-structured population, this can mimic signals of local adaptation (Ralph and Coop 2010).

If the copies of the favorable allele segregating in a population before the start of selection have *multiple origins*, this is a game-changer as new mutations (on random backgrounds) of the favorable allele, in addition to recombination, can scramble the selected allele over haplotypes. Likewise, even when a sweep *starts* as a single favorable allele on its way to fixation, *additional* new copies can arise by mutation during the sojourn of the original copy, potentially diffusing any pattern from the sweep over multiple haplotypes.

Pennings and Hermisson (2006a,b) approached this problem by considering the number of independent lineages of the favorable allele that are expected to be observed in a sample of n sequences following a sweep. Their rather remarkable result is that, to first order approximation, this is a function of θ_b, and *not* the strength of selection α_b. In particular, an upper bound for the probability of a multiple-origin soft sweep (two or more independent lineages in our sample of size n) is

$$\Pr(\text{soft}\,|\,n) \le \theta_b \left(\sum_{i=1}^{n-1} \frac{1}{i} \right) \approx \theta_b[0.577 + \ln(n-1)] \tag{8.18}$$

They also show that the number of distinct lineages in the sample approximately follows Ewens' (1972) sampling distribution (Equation 2.30a) using θ_b for θ. A more detailed analysis suggests the following general rules: If $\theta_b < 0.01$, multiple-origin soft sweeps are rare (even in a large sample), they are somewhat common for $0.01 \le \theta_b \le 1$, and almost certain for $\theta_b > 1$.

Orr and Betancourt (2001) also examined this problem, but from the perspective of standing variation alone, asking if **Haldane's sieve**, wherein dominant alleles are postulated to be more likely to contribute to selection response than recessive alleles (Turner 1981, Charlesworth 1992), is correct. They were also interested in the number of original copies that leave descendants in the fixed population. Assuming adaptation from standing variation alone, they found that dominance has little effect if the dominance relationship is roughly the same under the old deleterious and new favorable environments. Recessive deleterious alleles are at higher frequency than dominant deleterious alleles, which compensates for their lower probability of fixation in the new environment. Further, they showed that $\lambda = \theta_b s_b / s_d$ is the critical parameter in determining the number of independent lineages that leave descendants in the fixed population. When $\lambda > 1.26$, or

$$\theta_b s_b / s_d > 1.26, \tag{8.19}$$

the fixed collection of favorable alleles is more likely to contain multiple, as opposed to a single, lineages. If s_b and s_d are roughly the same magnitude, their effect cancels, again showing the strong dependence of a multiple-origins soft sweep on the value of θ_b.

Multiple-origin soft sweeps are therefore expected to occur under biologically realistic conditions. Pennings and Hermisson highlight two scenarios where this might be expected: very large effective population sizes and favored loss-of-function mutations. Under the latter scenario, the presence of numerous pathways by which function can be lost significantly increases the value of μ_b.

Example 8.7. *Caspase-12* (a cysteinyl asparate proteinase) is involved in inflammatory and innate immune response to endotoxins (Wang et al. 2006). In humans, most alleles are nulls and nucleotide diversity is sharply reduced (relative to levels in the chimp) around this locus, suggesting a selective sweep. Using the current frequency of roughly 0.9 for nulls, the authors estimate $s = 0.009$ (using Equation 8.3a) with the sweep favoring null alleles starting shortly before the out-of-African migration of modern humans. They hypothesize that null alleles were favored due to change in the environment increasing the odds of severe sepsis (bacterial infection of the blood) when this gene is active. Consistent with this hypothesis, two other primate genes related to sepsis are also pseudogenes in humans. Similar findings were also reported by Xue et al. (2006).

Given our above focus on the potentially important impact from new favorable mutations arising during a sweep, the diligent reader might wonder why we are ignoring neutral mutations at linked sites, which are expected to be far more common. The reason is that almost all new neutral mutations that appear as single copies are likely to be lost, while in a large population the odds are roughly $2s$ that a favorable (additive) allele will increase in frequency. How many such recurrent favorable mutations are expected to appear during the sojourn of the favored allele towards fixation? Recalling Equation 8.3d, the expected time for a single copy of the favorable allele to sweep through a population is $\tau \approx 2\ln(4N_e s)/s$. If N is the population size, then the expected number of new favorable mutations arising in a generation is $2N(1-x)\mu_b$, where x is the current frequency of the favorable allele. A rough approximation for the expected number of new favorable mutations that arise can be obtained by noting that the average frequency of a favored additive allele over its sojourn from near zero to near fixation is roughly 1/2. Hence

$$\begin{aligned} E(\text{new favorable mutations}) &\approx 2N(1/2)\mu_b\tau = (\theta_b/4)2\ln(4N_e s)/s \\ &= 2N_e\theta_b\ln(\alpha_b)/\alpha_b, \end{aligned} \tag{8.20a}$$

as obtained by Pennings and Hermisson (2006a). This is the *total* number of recurrent favorable mutations that arise, but each has only probability $2s$ of increasing. Hence, the expected number of new mutations that arise and increase in frequency (i.e., likely to become part of the fixed pool of the favorable allele after the sweep) is approximately $2s$ times our result in Equation 8.20a, giving

$$E(\text{new favorable mutations that increase}) \approx \theta_b\ln(4N_e s) \tag{8.20b}$$

Again, this is the number of favorable new mutations that increase in frequency during the sojourn of the initial allele to fixation, so that approximately $1 + \theta_b\ln(4N_e s)$ distinct lineages in the population are expected at fixation.

Example 8.8. To get a feel for the expected number of new favorable mutations that arise during a sweep, consider the values used in Example 8.6 ($N_e = 10^6$, $\theta_b = 0.0033$, $\alpha_b = 400$). From Equation 8.20a we expect

$$2N_e\theta_b\ln(\alpha_b)/\alpha_b = 2\times 10^6 \cdot 0.0033\ln(400)/400 \approx 90$$

new favorable mutations to arise, but the number we actually expect to increase in frequency (and hence contribute to the pool of favorable alleles following the sweep) is just

$$\theta_b\ln(4N_e s) = 0.0033\ln(400) = 0.02$$

Hence, even though a large number of favorable mutations arise, none really contribute to the sweep. This is consistent with the general rule that multiple-origin soft sweeps are unlikely when $\theta_b < 0.01$. Suppose we increase θ_b to 0.5, while keeping the other parameter values the same. Now roughly 15,000 recurrent favorable mutations are expected, three of which are expected to increase (and hence give a soft sweep) .

While the reader may feel that the critical parameter for observing a soft sweep ($\theta_b = 4N_e\mu_b$) is generally expected to be very small, recent results from *Drosophila* suggest that more caution is in order. A common view is that the target site for a beneficial mutation is small (only one or a few sites can change) and hence the small nucleotide mutation rates (10^{-9} to 10^{-8}) suggest that such events are highly unlikely. However, it may be that μ_b is much larger than we think. González et al. (2008) found that transposable genetic elements (TEs) can induce adaptation in *Drosophila melanogaster*. In a set of 909 TEs that inserted into new sites following the spread of this species out of Africa, at least 13 show signs of being adaptive (associated with signatures of partial sweeps). They suggest that the majority of these are likely due to regulatory changes. The much higher rate of TE mobilization (relative to nucleotide mutation rates) coupled with their much larger target of action (their insertion at a large number of sites can influence regulation), suggests that μ_b may often be much larger than one expects.

Even independent single-site mutations may be more likely than expected. A potential human example of this derives from the work of Enattah et al. (2007) on the lactase gene (*LCT*). Variants at this gene are correlated with lactase persistence (the ability to utilize milk as an adult) and hence are candidates for selection following the invention of dairy farming. They found that the T_{-12910} variant upstream of *LCT* appears to have at least two independent origins. In addition to the common northern European allele, an independent origin appears to have occurred in an isolated region in eastern Europe (west of the Urals and north of the Caucasus). Further, Tishkoff et al. (2007) found independent mutants at different sites in the *LCT* gene in African populations that also lead to lactase persistence.

The second component to θ_b is N_e. This, too, might be much larger than expected (perhaps approaching the population census size), at least during short windows in time. Recall (Chapter 3) that N_e is a harmonic mean, and hence very sensitive to bottlenecks, no matter how infrequent. Current estimates of N_e are often based on levels of nucleotide diversity, which are generated by the cumulative joint action of mutation and drift over rather long periods of time. Conversely, when a favorable mutation appears, it can sweep through a population very quickly (relative to the drift time scale of $4N_e$ generations), and hence the effective population size during the short window of their sojourn may be much higher.

Example 8.9. Karasov et al. (2010) examined *Drosophila melanogaster* mutations at the *Ace* gene, which codes for the neural signaling enzyme Acetylcholinesterae, a target for many commonly used insecticides. Single nucleotide changes at four highly conserved sites confer partial insecticide resistance, with combinations of these conferring significantly greater resistance. Single, double, and triple mutations are all found in natural populations. While one model is that these variants existed at the start of major insecticide use (the 1950's), the authors found that mutations in North American and Australia appeared to have arise *de novo* following the *D. melanogaster* migration out of Africa. Given that only 1000 to 1500 fly generations have elapsed since the widespread use of insecticides that target the *Ace* product, estimates of $\theta \sim 0.01$ based on nucleotide diversity (and hence a θ_b of 1/3 this value at each of the four sites) are not consistent with the independent origins of single, much less multiple, mutations in this gene over this short time scale. However, if the actual effective population size was 10^8 instead of the standard assumed value of 10^6 during the past 50 years, then $\theta_b \sim 1$, and such multiple independent origins by mutation are highly likely. The effective population size that matters for these mutations is that during their origin and spread, not that set by any history predating their appearance.

Signatures of a Soft Sweep

The effect of a single-origin soft sweep is to soften, perhaps even erase, most of the signatures expected under a hard sweep. If the original copy is at very low frequency at the start of selection, a hard-sweep signature can be generated. However, hard-sweep signatures quickly dissipate as this initial frequency increases. The situation is even more dramatic for multiple-origin soft sweeps (Pennings and Hermisson 2006b). For the heterozygosity following a sweep, Equation 8.8c now becomes

$$\frac{H_h}{H_0} \simeq 1 - \frac{1}{1+\theta_b}(4N_e s)^{-2c/s}, \tag{8.21a}$$

so that even with a completely linked site,

$$\frac{H_h}{H_0} \simeq 1 - \frac{1}{1+\theta_b} > 0. \tag{8.21b}$$

For example, if $\theta_b = 0.01$, then $H_h/H_0 \simeq 0.01$, while if $\theta_b = 0.5$, then $H_h/H_0 = 0.33$.

Example 8.10. The myostatin gene (*MSTN*) is a negative regulator of skeletal muscle growth. Mutations in this gene underlie the excessive muscle development in double-muscled (DM) breeds of cattle, such as Belgian Blue, Asturiana de los Valles, and Piedmontese. Wiener et al. (2003) compared microsatellite variation as a function of their distance from *MSTN* in DM and non-DM breeds. For DM breeds, measures of variation decreased relative to non-DM breeds as they approached the *MSTN* locus. While this approach clearly indicates a genomic region under selection, the authors expressed skepticism about its ability to fine-map the target of selection (i.e., localize it with high precision within this region). At first glance, this seems surprising given that *MSTN* variants have a major effect on the selected phenotype (beef production). However, the authors note that Belgian Blue was a dual purpose (milk and beef) breed until the 1950's, and that in both Belgian Blue and Piedmontese there are records of this mutation before World War One, predating the intensive selection on the double-muscled phenotype. By contrast, they found that the selective signal is stronger in Asturiana, where the first definitive appearance of the mutation was significantly later. Thus, in both Belgian Blue and Piedmontese selection on this gene resulted in a soft sweep (adaptation from preexisting mutations), while in Asturiana the time between the initial appearance of the mutation and strong selection on it was much shorter, resulting in a more traditional hard sweep (adaptation from a new mutation). O'Rourke et al. (2012) used haplotype homozygosity to estimate the age of the Belgian Blue mutation (*821dell11*) at between $\sim$ 200-400 years, and $\sim$ 200 years for the Piedmontese mutation (*C313Y*).

In addition to the reduction in the heterozygosity signal, Pennings and Hermisson find that soft sweeps also significantly depress any sweep signal in the site-frequency spectrum.

Indeed, even when $c = 0$, the folded frequency spectrum after a soft sweep can be very close to the neutral (Watterson) spectrum. However, soft sweeps do leave a strong (but very transient, roughly $0.1N_e$ generations) signature in linkage disequilibrium (LD). A lower number of haplotypes and a higher level of association between sites relative to drift are expected, at least during a short window following the sweep. Pennings and Hermisson found that the power of LD-based tests for detecting soft-sweeps is significantly enhanced by ignoring new mutations. They suggest that when a closely-related population/sister species is available, using only sites that are shared polymorphisms (and hence not recent mutations) in both population can improve power. While there can be a strong, albeit transient, signal in LD following a soft sweept, it is *quite different* from the LD signature from a hard sweep. Under the latter, LD is zero *across* the selected site following fixation, while under a soft-sweep, LD extends *through* a site. As discussed in Chapter 9, the ω^2 statistic (Equation 9.37), which contrasts LD on either side (but not across) a site can detect hard sweeps, but misses soft sweeps, while the Z_{nS} test (Equation 9.36b), which computes the average LD over all sites in a region misses hard sweeps but can detect soft and ongoing (i.e., partial) sweeps.

Under a soft-sweep (especially when $\theta_b > 1$), there is no single dominant haplotype as would be expected under a hard-sweep. However, Garud et al. (2014) note that there may instead be a *few* dominant haplotypes (provided the sweep is not too soft, e.g., $\theta_b < 10$), and suggested a simple modification of a standard hard-sweep test to detect soft sweeps. If p_i denotes the frequency of the ith haplotype in our sample (ranked, so that $i = 1$ is the most frequent), then the **haplotype homozygosity** $H_1 = \sum p_i^2$ should be excessive relative to its value under neutrality under a hard sweep. Garud et al. suggest a modified haplotype homozygosity statistic H_{12} lumping the first two haplotypes into a single class,

$$H_{12} = (p_1 + p_2)^2 + \sum_{i>3} p_i^2 = H_1 + 2p_1p_2$$

finding that this test has reasonable power to detect *both* hard and soft sweeps. The reason for the former is that there is one dominant haplotype under a hard sweep, so adding on the frequency of the next most common has very small effect. They also suggested a test to distinguish between soft and hard sweeps. Again, the logic is that under a soft sweep, we expect a few haplotypes to be common, but only a single haplotype under a hard sweep. Their $H_2 = \sum_{i>1} p_i^2$ statistic is simply haplotype homozygosity with the most common haplotype removed, with the ratio of H_2/H_1 expected to be very small under a hard sweep, but modest under a soft sweep. When Garud et al. applied this method to the 50 largest sweeps detected using $H12$ in a North American sample of *Drosophila melanogaster*, all showed much stronger support (Bayes factors > 10, see Appendix 2) for soft, rather than hard, sweeps.

Polygentic Sweeps

The strength of signal left by a hard sweep is a function of the strength of selection, with any signal significantly diminished under soft-selection scenarios. This suggests that weak selection at a number of loci (especially if standing variation is used and/or the underlying loci have large mutational targets) is the worst-case scenario for detecting recent/ongoing selection. Unfortunately, this appears to be *exactly* the situation for many, indeed perhaps most, quantitative traits. As detailed in Chapter 18, just about any trait in an outbred populations shows some, and usually rather significant, response to artificial selection. Given the immediate nature of response, standing genetic variation underlies almost all initial response to selection on complex traits, although contributions from new mutations becomes increasingly important over time (Chapter 26).

Recalling that Equation 5.21, the strength of selection $s = \bar{\imath}(a/\sigma_z)$ on a QTL allele underlying a complex trait is a function of the strength of selection on that trait ($\bar{\imath}$, the within-generation change in the mean, expressed in standard deviations) and the fractional contribution of the allele to overall trait variation (a/σ_z, the additive effect for that allele, scaled in phenotypic standard deviations). With modest selection on the trait (a 0.1 change in phenotypic standard deviations within a generation) and a modest contribution from an underlying QTL (an effect of 0.01 standard deviations), $s = 0.001$. Assuming a recombination fraction of 1 cM/Mb, Equation 8.6a suggests that a sweep at this locus should cover roughly $0.02\,(0.001/0.01) = 0.002$ Mb = 2,000 bases. This is a small track, and yet it is the *best case* scenario, a hard sweep. Under a soft sweep this signal is further degraded. Moreover, for most complex traits, the situation (from the standpoint of detecting sweeps) is even worse. Polygenic response occurs through the *joint* response over a number of loci, allowing for substantial change in the trait mean with only modest change in allele frequencies at the underlying loci (Chapter 24). Thus, significant response in the mean value of a trait can occur through modest changes over a number of loci of small effect using standing variation. Further, it is generally assumed that QTLs have a large mutational target, as subtle changes in regulation likely result in subtle changes in the contribution of a locus to trait value.

Given these concerns, Pritchard and Di Rienzo (2010) and Pritchard et al. (2010) suggest that such polygenic adaptation is likely to leave little, if any, signal under traditional approaches (also see Chevin and Hospital 2008; Pavlidis et al. 2012). This prediction was observed in a scan of selection on cattle by Kemper et al. (2014). Strong signals were seen around major loci that define breeds – *polled* (the absence of horn) and coat color loci. Five major loci with strong effects on production traits (stature, milk production, muscle mass) also showed signatures, but not as strong. However, no significant traditional signals were observed around a large number of small effect QTLs directly involved in production traits under strong selection.

How might “**polygenic sweeps**” be detected? An interesting suggestion comes from Hancock et al. (2010a, b), who looked for subtle allele frequency shifts that were concordant for human populations in similar environments, but different geographic regions (also see Fumagalli et al. 2011). Such approaches clearly have power issues (being a function of the number of independent replicates under the same environmental conditions) and also rely on the same alleles responding in the same environmental conditions. However, one interesting study is Turchin’s et al. (2012) comparison of height-QTL allele frequencies in Northern versus Southern Europeans. Of 139 variants associated with height, the allele of larger effect was more common in Northern Europeans in 85 out of 139 cases ($p = 0.01$), with a highly-significant average frequency increase of around 1.2%. While the effects for any single locus is quite small (average effects are $\leq 10^{-2}$ - 10^{-3} SD, and an allele frequency change of ~ 0.01), and hence have low power when considered separately, when combined as a group there is a strong signal of likely polygenic selection. Further, s on these underlying loci is $< 10^{-4}$.

Under what situations might one expect hard sweeps versus polygenic adaptation? In reality, given the vast reservoir of standing variation for most traits, a shift to a new environment will likely have an initial polygenic response, but a major allele or major mutation could still occur and have very dramatic effects. Thus, hard sweeps are expected in situations where very little standing variation for the trait is present, as might occur for traits with a long history of consistent directional selection. In such cases, further response might be mutation-limited. However, it can also be the case that low standing variation is in part due to one (or more) alleles of major effect that are at low frequency because of deleterious fitness effects (Lande 1983). A sudden shift in the environment may result in their major effect on a trait under selection overcoming any adverse effects on other traits, giving the allele a net selection advantage, which can result in a hard-sweep (from standing variation) if the allele is sufficiently rare (we revisit this topic in Chapter 25).

GENOME-WIDE IMPACT OF REPEATED SELECTION AT LINKED SITES

Up to this point, our focus has been on the local impact of a single sweep. There is a much broader picture as well — *recurrent* hitchhiking events can have profound implications on the entire genome, a topic first introduced in Chapter 3. Indeed, Maynard Smith and Haigh (1974) proposed that **recurrent selective sweeps** could depress variation throughout a genome, potentially providing a solution to the vexing observation that levels of polymorphism (expected value $4N_e\mu$ under the equilibrium drift model) don't seem to scale with N_e (a second potential fractor is the decrease in μ with increasing N_e, see Chapter 4). Large-scale sequencing has lead to the current view that recurrent selection at linked sites does indeed have a profound effect on many, perhaps most, genomic regions, reducing standing levels of variation by lowering N_e. Such a reduction can also elevate the fraction of new mutations that are effectively neutral, potentially increasing the substitution rates in these regions. The current debate is what fraction of these genome-wide effects is due to recurrent sweeps (adaptive evolution) versus background selection against deleterious mutations (purifying selection). Our goal in this section is to introduce the basic theoretical results, many of which follow from the above machinery for sweeps, as well as examine the current empirical data.

Effects of Recurrent Sweeps

In a region with low recombination, even weak selection at a distant location can have an impact. In the extreme where an entire genome has *no* recombination (such as a bacteria or an organelle), a single advantageous mutation can sweep a single genotype to fixation. Laboratory populations of bacteria often show the phenomena of **periodic selection** (Atwood et al. 1951a,b; Kock 1974; Dykhuizen 1990; Guttman and Dykhuizen 1994), wherein genetic diversity builds up slowly over time only to be rapidly removed before starting all over again. Presumably, this is due to the periodic fixation of newly-appearing favorable mutations, which generate sweeps that fix single chromosomal lineages. In this setting, the standing levels of variation are a function of the frequency of sweeps. If sweeps are sufficiently common, the population never has a chance to reach mutation-drift equilibrium following each sweep, while if rare, the population may be at mutation-drift equilibrium much of the time. Thus, the rate of adaptation at least partly determines the amount of neutral variation, a theme returned to throughout this section.

On a less dramatic scale, telomeric and centromeric regions of chromosomes typically show reduced levels of recombination, while very small chromosomes (such as the fourth of *D. melanogaster*) may have essentially no recombination. Studies in *D. melanogaster* show that regions with reduced recombination also have reduced genetic variation (Aguadé et al. 1989; Berry et al. 1991; Begun and Aquadro 1991). The fact that between-species divergence does not appear to be depressed in these regions suggests that a reduction in the mutation rate is not the culprit, and the initial interpretation was that this pattern is generated, as with periodic selection, by recurrent sweeps of favorable mutations reducing linked neutral variation.

For a population of constant size undergoing periodic hard sweeps, Wiehe and Stephan (1993) found that the equilibrium level of heterozygosity is a decreasing function of the rate of adaptation λ. This is the number of beneficial mutations destined to become fixed that arise each generation, namely the product of the number of new beneficial mutations arising each generation $2Nu_b$ times their individual fixation probabilities $2sN_e/N$, or

$$\lambda = (2Nu_b)(2sN_e/N) = 4N_e s u_b \tag{8.22a}$$

Wiehe and Stephan found that the expected heterozygosity, as measured by nucleotide diversity π, at linked neutral sites is approximately

$$\frac{\pi}{\pi_0} \simeq \frac{c}{c + \lambda\,\gamma\,\kappa} = \frac{1}{1 + \lambda\,\gamma\,\kappa/c} \tag{8.22b}$$

where $\pi_0 = 4N_e\mu$ is the average heterozygosity at a single site for an equilibrium neutral population under no sweeps, c is the per-nucleotide recombination rate over the region of interest, $\gamma = 2N_e s$ the scaled strength of selection, λ the per-nucleotide adaptive substitution rate, and the constant $\kappa \simeq 0.075$. For modest values of c (relative to $\lambda\gamma\kappa$), Equation 8.22b is approximately

$$\frac{\pi}{\pi_0} \simeq 1 - \frac{\lambda\gamma\kappa}{c}. \tag{8.22c}$$

A change of variables gives the linear **Stephan regression**,

$$y = \pi_0 - (\lambda\gamma\kappa)\,x, \quad \text{where } y = \pi, \text{ and } x = \pi_0/c \tag{8.22d}$$

whose intercept and slope estimate π_0 and $\lambda\gamma\kappa$, respectively (Stephan 1995).

Fitting Equation 8.22b, Wiehe and Stephan (1993) obtained an estimate of $\lambda\,\gamma \simeq 1.3 \times 10^{-8}$ based on 17 loci in medium to high recombinational backgrounds in *D. melanogaster*. For a modest recombination rate of 1 cM per megabase, $c = 0.01/10^6 = 10^{-8}$, this value of $\lambda\,\gamma$ implies

$$\frac{\pi}{\pi_0} \simeq \frac{10^{-8}}{10^{-8} + 1.3 \times 10^{-8} \cdot 0.075} = 0.911$$

or roughly a 9 percent reduction in background heterozygosity. For small recombination rates, say 0.1 cM per megabase ($c = 10^{-9}$) standing levels of variation are reduced by 49 percent, while in a region of high recombination (2.5 cM/Mb, $c = 2.5 \times 10^{-8}$), the reduction in π is only 3.7 percent. Hence, in regions of low recombination, recurrent selective sweeps can have a dramatic effect on standing levels of variation. Additional estimates of $\lambda\gamma$ are summarized in Table 8.3.

A Few Large or Many Small Sweeps?

Since reduction in heterozygosity from sweeps is a function of the product $\lambda\gamma$, the same average reduction in π could be caused by either a few large sweeps (λ small, γ large) or many smaller sweeps (λ large, γ small) as long as their product is held constant. With rare, strong sweeps, there would be dramatic reduction in variation over a fairly large region, but many regions would see little effect, as no recent sweep would have occurred in their vicinity. Conversely, with many weaker sweeps, most regions would be influenced, but each by a smaller amount. While the expected value of π is the same under both models, the variance in π is expected to be much greater under rare strong sweeps (Jensen et al. 2008).

Example 8.11. As summarized in Table 8.3, for a set of X-linked genes in *D. melanogaster*, Andolfatto (2007) and Jensen et al. (2008) obtained estimates for the per-nucleotide adaptation rate λ of 7.5×10^{-10} and 4.2×10^{-11} (respectively). Consider a region of length 100 kb. Under Andolfatto's estimate, the per generation rate of adaptive substitutions over a region of this size is $10^5 \cdot 7.5 \times 10^{-10} = 7.5 \times 10^{-5}$ or one sweep roughly every 13,300 generations. Under Jensen's estimate, a sweep influencing this region occurs roughly every 238,000 generations. Potential reasons for such a vast difference in the estimates will be examined shortly.

Distinguishing between the strong and weak selection scenarios requires an independent estimate of either λ or γ in addition to an estimate of $\lambda\gamma$. Methods to accomplish this are more fully developed in Chapter 10, but one approach is as follows. Suppose L sites are examined between two populations that separated t generations ago, and a total of D sites show divergence, giving $d = D/L$ as the per-site divergence. Ignoring multiple mutations at the same site, if α denotes the fraction of all divergent sites that are adaptive, $d\alpha$ is the per-site number of adaptive divergences, which occured over $2t$ generations. This gives the rate as $\lambda = d\alpha/(2t)$. Estimates of t are possible from several sources, but estimates of the adaptive fraction α seem more elusive. However, as detailed in Chapter 10, for coding regions estimates of α follow by noting that the ratio of the number of silent to replacement polymorphic sites should equal the ratio of the number of silent to replacement substitutions under drift. An excess of replacement substitutions presumably reflects the role of adaptive evolution, and the amount of excess allows an estimate of α (e.g., Example 10.1), and hence of λ.

Table 8.3. Estimates of the rates of adaptive evolution λ and its components at the molecular level for several *Drosophila* species and for the aspen tree (*Populus tremula*). The species listed provided the polymorphism data, while an outgroup was used for some estimates of λ (Equation 10.11a). Methods for estimating individual components of the product $\lambda\gamma$ (the scaled strength of selection $\gamma = 2N_e s$, the rate of adaptive substitutions per base pair per generation λ, and the average strength of selection of a beneficial mutation s) are more fully developed in Chapter 10.

Organism	$\lambda\,\gamma$	γ	s	λ	Reference
D. melanogaster	3.9×10^{-7}	34,400	2.0×10^{-3}	6.0×10^{-11}	Li and Stephan 2006
	5.1×10^{-8}	74	2.3×10^{-5}	7.0×10^{-10}	Bachtrog 2008
	2.6×10^{-8}	35	1.2×10^{-5}	7.5×10^{-10}	Andolfatto 2007
	4.0×10^{-7}	10,000	2.0×10^{-3}	4.2×10^{-11}	Jensen et al. 2008
				1.8×10^{-11}	Smith & Eyre-Walker 2002
				3.6×10^{-11}	Andolfatto 2005
	1.3×10^{-8}				Wiehe & Stephan 1993
		10			Schneider et al. 2011
D. simulans	1.1×10^{-7}	30,000	1.0×10^{-2}	3.6×10^{-12}	Macpherson et al. 2007
D. miranda	1.2×10^{-6}	3,100	2.7×10^{-3}	4.0×10^{-10}	Bachtrog 2008
P. tremula	1.5×10^{-7}				Ingvarsson 2010
Humans				2.3×10^{-12}	Example 10.12

Letting d_a denote the per-site rate of amino-acid divergence, and substituting $\lambda = d_a\alpha/(2t)$ into Equation 8.22b gives the **Andolfatto regression**,

$$\pi \simeq \pi_0 \frac{c}{c + \lambda\,\gamma\,\kappa} = \pi_0 \frac{c}{c + [\alpha\gamma\,\kappa/(2t)]\,d_a} = \frac{\pi_0}{1 + \beta x} \tag{8.23a}$$

where $x = d_a/c$ and $\beta = \alpha\gamma\,\kappa/(2t)$ (Andolfatto 2007). As shown in Figure 8.9, the per-site amino acid divergence d_a for each gene is scaled by its local rate of recombination c, leaving a regression between observed nucleotide diversity π and d_a/c. The resulting regression parameters become π_0 and $\alpha\gamma\,\kappa/(2t)$, which (with estimates of α and t in hand) returns γ. Alternatively, using $\gamma = 2N_e s$, we can rewrite this regression as

$$\pi = \pi_0 \frac{c}{c + \alpha s[\kappa N_e/t]\,d_a} \tag{8.23b}$$

returning an estimate of αs scaled by the divergence time in N_e units. Chapter 10 reviews other approaches for estimating α and/or γ from joint polymorphism and divergence data at single loci.

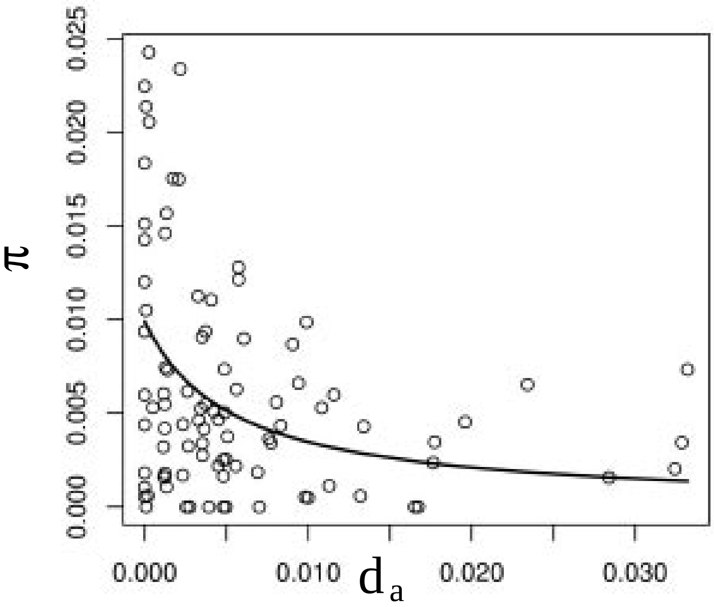

Figure 8.9. An example of Andolfatto's regression of the nucleotide diversity π on the per-site amino-acid divergence d_a in *Drosophila miranda*. The solid curve is the least-square fit of Equation 8.23a, which gives estimates of π_0 and γ (as α and t were independently estimated). After Bachtrog (2008).

A technically more demanding approach to using regressions based on π is to jointly estimate two of these three parameters (λ, γ, or $\lambda\gamma$) using the spatial pattern of genetic variation over a region (Macpherson et al. 2007; Jensen et al. 2008). Figure 8.10 shows the motivation for this idea. Jensen et al. (2008) noted that strong selection should produce a higher variance in π and other measures of genetic variation that are impacted by a sweep (Table 8.2), such the number S of segregating sites (Chapter 2), excessive of high-frequency derived alleles, and pairwise LD. Using simulations, they examined the behavior of the coefficient of variation (CV) for summary statistics for these quantities as a function of the size L of the unit of analysis. Over small regions (L of 500 to 1000 bp), there was little difference in the CV between the rare/strong versus frequent/weak sweep models, but as the size of the analysis region increased, so did the CV for strong, but not weak, selection.

Based on this observation, Jensen et al. (2008) developed an approximate bayesian approach that jointly considers the means and variances of summary statistics measuring these various sweep features (such as π and S for polymorphism levels, θ_H for departures in the site-frequency spectrum, and Z_{nS} for the local structure of LD; the latter two given by Equations 9.28a and 9.36b) to obtain separate estimates of λ and s from joint polymorphism-divergence data. As outlined in Appendix 3, the general approach for **approximate bayesian calculations** (or **ABC**) is to generate a posterior as follows. First, draw potential λ and s values from some prior, and then use these to generate a simulation of the sweep. The summary statistics of interest are them computed and if sufficiently close to the observed values, the joint λ and s values are kept, else they are rejected and new values drawn. This procedure is repeated several thousands of times to generate a joint empirical posterior distribution of λ and s values consistent with the observed data. They found that assuming a constant s value for each sweep results in an overestimation of s and underestimation of λ relative to allowing each new sweep to have an s value drawn from a distribution.

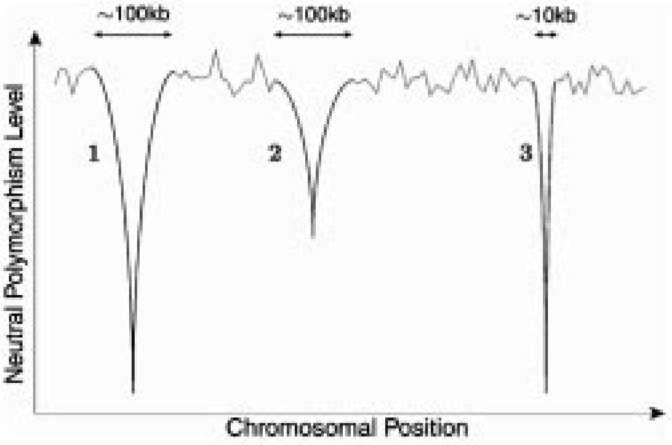

Figure 8.10. The pattern of nucleotide diversity over a large region may provide clues on the frequency and strength of past sweeps. Within this hypothetical region, three hard sweeps have occurred. Sweep 1 is a strong, recent sweep; 2 is a strong older sweep; and 3 a weak recent sweep. Strong sweeps (1) result in a depression in variation over a significant region. As the signal from a past sweep decays, its window of influence stays roughly the same size, but its impact within that window vanishes over time. An old strong sweep (2) leaves a weak signal of depressed variation over a fairly large region, while an old weak selection leaves a similar signal to (2), but over a much smaller region. After Macpherson et al. (2007).

Macpherson et al. (2007) also used spatial information, starting the with standard regression of π on c, (Equation 8.22b) which is a function of $\lambda\gamma$. They then introduced a new statistic Q_S, the ratio of a minimal estimate of heterozygosity within a window to the average heterozygosity over the region scanned by the windows. Their key insight was that the location for the minimal value corresponds very closely with actual selected site, and hence its value is *not* a function of the strength of selection (as recombination is very near zero, and hence all of the site is swept along, independent of the strength of selection). They showed that the expected value of Q_s is function of both $\lambda\gamma$ and λ, so that the joint pair of statistics Q_s and π allows for separate estimates of λ and $\lambda\gamma$.

As summarized in Table 8.3, while estimates of the product $\lambda\gamma$ for various studies in *Drosophila* are reasonably compatible, individual estimates of its components γ (or s) and λ can differ by several orders of magnitude. There are several potential reasons for this. Different studies of even the same species may use different populations as well as different sets of genes, such as autosomal (Macpherson et al. 2007) versus X-linked (Andolfatto 2007; Jensen et al. 2008; Bachtrog 2008). These authors use a variety of different methods, and this may be the major contributor to the significant disparity between studies. Estimates based on short regions (single genes) as the unit of analysis, such as those by Andolfatto (2007) and Bachtrog (2008), found small estimates of γ and s in *D. melanogaster* (γ between 35 and 74, s around 10^{-5}). Estimates based on much longer regions (10-100 kb), such as Macpherson et al. (2007) and Jensen et al. (2008) found much larger estimates of γ (10,000 to 30,000) and s

(0.002 to 0.01). Estimates obtained by Bachtrog (2008) for *D. miranda* using a number of small regions were intermediate, with $\gamma = 3000$, $s = 10^{-3}$.

Motivated by Figure 8.10, Sella et al. (2009) suggested that these estimates of γ and λ may actually be more compatible than their spread suggests. Weak selection leaves a strong signal over only a very small region, while strong selection leaves a signal over a much larger region. For example, using an average recombination rate of 1 cM/Mb ($c = 10^{-8}$), Equation 8.6a suggests that weak sweeps ($\gamma = 35$, $s = 10^{-5}$) only influence at most a few hundred bases, while strong sweeps ($\gamma = 10,000$, $s = 0.01$) can influence almost a hundred kilobases. Sella et al. suggest that methods using small regions (such as single genes) for their units of analysis are biased towards the detection of weak selection, while methods using much larger regions are biased towards strong selection. Under this view, weak selection accounts for most of the observed between-population divergence, while strong selection accounts for most of the reduction in heterozygosity. Sattah et al. (2011) found strong support for this view by examining the pattern of nucleotide variation around substitutions in *Drosophila simulans*. As expected from the neutral theory, there was a slight elevation in standing diversity around silent site substitutions (as would be expected given both scale with the local genomic mutation rate). After adjusting for this effect, they found a trough in nucleotide diversity around sites that resulted in amino acid substitutions, as would be expected given a sweep. Using a composite likelihood approach to test for a sweep given the local diversity pattern around a site (Chapter 9), they estimated that around 13% of the substitutions resulted in sweeps. A mixture model allowing for different strengths of selection fit the genomic data the best, and suggested that about 30% of the sweep sites (4% of the total substitutions) were under strong selection with a mean s value of $\sim 0.5\%$ while the remainder had a mean s of $\sim 4 \times 10^{-5}$.

Selective Interference and the Hill-Robertson Effect

The above results for recurrent sweeps assume that *concurrent* sweeps influencing the same region are rare. Given that the expected number of new beneficial mutations introduced into the population each generation scales of NU_b (with U_b the total beneficial mutation rate for the region of interest), this assumptions breaks down in sufficiently large populations. In this setting, multiple sites are segregating beneficial mutations, and sweeps can interfer with each other (Kim and Stephan 2003; Neher et al. 2010; Neher and Shariman 2010; Weissman and Barton 2012).

If sites with segregating beneficial mutations are loosely-linked, their effect is to lower the effective population size that a particular site experiences, making selection slightly less efficient (the Hill-Roberson effect, Chapter 7). A more dramatic effect occurs when some of these sites are tightly linked, in which case there can be strong **selective interference** among them. One example is **clonal interference** (Gerrish and Lenski 1988) seen in very large populations with complete linkage (i.e., bacteria). Competition occurs among the set of lineages (clones) carrying different beneficial mutations, as each such clone has a strong advantage against lineages lacking such mutations, but (at best) only a weak advantage against other beneficial lineages. In a sexual population, recombination can shuffle genomes to combine beneficial mutations, but the dynamics are very different in large asexual populations. Here NU_b is expected to be large, with multiple favorable mutations arising each generation. A genome with an existing beneficial mutation might not acquire *new* favorable mutations as quickly as other genomes and thus become lost (Desai and Fisher 2007). As a consequence, Neher et al (2010) show that while the appearance of favorable new mutations scales linearly with N (as NU_b), selective interference significantly *reduces* individual fixation probabilities as N increases, resulting in the rate of adaptive evolution scaling as $\log(NU_b)$.

Weissman and Barton (2012) extend this analysis to sexual species. Letting $\Lambda = L\lambda$ be the total rate of adaptation over a region of length L and total recombination rate $C = Lc$. If

$\Lambda_0 = L\lambda_0$ is the expected rate of adaptation if the effects of interference are ignored (Equation 8.22a), they found that the observed value Λ is approximately given by

$$\Lambda \simeq \frac{\Lambda_0}{1 + 2\Lambda_0/C}, \quad \text{or} \quad \lambda \simeq \frac{\lambda_0}{1 + 2\lambda_0/c}. \tag{8.24a}$$

Expressing this as

$$\Lambda/C \simeq \frac{\Lambda_0/C}{1 + 2\Lambda_0/C} \tag{8.24b}$$

shows for very large populations where $\Lambda_0/C \gg 1$ (i.e., $\lambda_0/c \gg 1$) that the upper limit of adaptation is given by $\Lambda \simeq C/2$ (or $\lambda \simeq c/2$). For example, for a one centimorgan region ($C = 0.01$), the total rate of adaptive substitutions over that region is bounded by $\Lambda \leq 0.005$, or the appearance of one successful beneficial mutation every 200 generations. There is an additional effect from selection at unlinked sites, which reduces the rate further by $\exp(-4\Lambda s)$, which (for small s) is negligible relative to the effects of linkage. Weissman and Barton used their results to examine the impact of selective interference on the levels of variation at linked neutral sites. For small values of λ_0/c, there is not much interference among sweeps, and Equation 8.22b is a good approximation for the loss of linked neutral variation. Conversely, when interference is strong, substantial neutral variation can still be maintained, as the density of sweeps (λ/c) approaches a limiting asymptotic value, and the coalescent times for pairs of neutral alleles (and hence the amount of variation maintained) is a function of sweep density.

This notion of inference among linked selected sites is not restricted to beneficial alleles. Indeed, as we will see later, it may have a significant impact when considering tightly linked weakly *deleterious* alleles. We have referred to situations where selection at one site influences the strength of selection at a second as the Hill-Robertson (HR) effect (Chapter 7). As the above discussion on selective interference illustrates, while the HR effect is usually regarded as a reduction in N_e, this is not the whole story (Felsenstein 1974; Comeron and Kreitman 2002; Neher and Shariman 2011; Schiffels et al. 2011), as selective interfere must also be considered. With free recombination between selected sites, gametes quickly arise that contain multiple favorable alleles, giving a wider range of fitness values and hence more efficient selection due to a higher additive variance in fitness. Conversely, with linkage, the production of gametes with multiple favored alleles is retarded (selection generates negative linkage disequilibrium among favorable alleles; Chapters 5, 16), resulting in a smaller variance in fitness among individuals in the population and hence less efficient selection.

Background Selection: Reduction in Variation Under Low Recombination or Selfing

Charlesworth et al. (1993) challenged the view that reduction of variation in regions of low recombination was evidence for periodic selective sweeps (and hence the frequent substitution of adaptive alleles). They noted that the exact same pattern can be generated by selection *against* new deleterious mutations. Hence, purifying selection can potentially account for this pattern of reduced variation without the need to invoke adaptive selection. This occurs because removal of new deleterious mutations lowers the effective population size, and in a sufficiently long region of low recombination, the number of targets for mutation may be large enough to generate a high total deleterious mutation rate and therefore a significant reduction in variation. They referred to this process as **background selection** (or **BGS**), which was introduced in Chapter 3. We review (and generalize) some of our Chapter 3 results here in order to more fully contrast BGS against recurrent sweeps.

Charlesworth et al. estimated the potential impact of BGS as follows. First, consider a neutral site completely linked to a region in which new deleterious mutations arise at rate U (per gamete). A key assumption is that these new mutations are sufficiently deleterious to be removed rapidly, so that the population is at an equilibrium with the removal of mutation-bearing chromosomes by selection balanced by the creation of new such chromosomes by mutation. Assuming that the fitness of a new deleterious mutation (in the heterozygous state) is $1 - hs$, and that fitness over loci is multiplicative, the expected number of deleterious mutations per gamete at the mutation-selection equilibrium is $U/(2hs)$ (Kimura and Maruyama 1966). Further, the number of mutations follows a Poisson distribution, so that the probability of a mutation-free gamete is given by the zero term of a Poisson,

$$f_0 = \exp\left(-\frac{U}{2hs}\right) \tag{8.25}$$

The effect of background selection is to reduce effective population size from N_e to $f_o N_e$, giving an expected reduction in neutral variation of $\pi/\pi_0 = f_0$ as well. Since *selfing* reduces the effects of recombination by decreasing the number of heterozygotes (and hence the chance for recombination), the effects of background selection can be quite significant in highly selfing plant populations. Charlesworth et al. (1993) noted that the reduction in strict selfers is given by Equation 8.25, with hs replaced by s, the selection against mutant homozygotes.

Hudson and Kaplan (1995) extended these results by allowing for recombination. For a neutral locus in the middle of a region of length L and total recombination frequency C,

$$\frac{\pi}{\pi_0} \simeq \exp\left(-\frac{U}{2hs + C}\right) \tag{8.26a}$$

where U is the total mutation rate within this region, with $U = L\mu$ and $C = Lc$, where (as above) μ and c denote the average per-nucleotide rates of mutation and recombination. When the total amount of recombination within the region is large relative to hs ($C \gg hs$),

$$\frac{\pi}{\pi_0} \simeq \exp\left(-\frac{u}{c}\right) \tag{8.26b}$$

Under these conditions, the decline in heterozygosity is independent of the strength of selection. Since $e^x \sim 1 - x$ for $|x| \ll 1$, it follows that for moderate to high recombination ($u/c \ll 1$) that

$$\frac{\pi}{\pi_0} \simeq 1 - \frac{u}{c} \tag{8.26c}$$

This is the same form ($\pi/\pi_0 = 1 - b/c$) as our moderate to high recombination result under recurrent selective sweeps ($b = \mu$ in Equation 8.26c, $b = \lambda\gamma\kappa$ in Equation 8.22c). As a consequence, for this range of recombination values, the regression of π on c cannot distinguish between hitchhiking and background selection. Hudson and Kaplan (1995) found that background selection provided a reasonable fit to the polymorphism data over most of the third chromosome of *D. melanogaster*, while Charlesworth (1996) found that the background selection model provides a good fit for most regions of the *D. melanogaster* genome. However, not surprisingly, Stephan (1995) found that the recurrent sweep model gave an excellent fit as well.

In regions of very low recombination, some of the assumptions leading to Equation 8.25 can break down, and both Hudson and Kaplan (1995) and Charlesworth (1996) found that the BGS model gives a poor fit in such regions. Hudson and Kaplan were able to obtain a reasonable fit to the BGS model, but only by using much smaller selection coefficients than assumed for higher recombining regions on the same chromosome. The problem, as noted by Kaiser and Charlesworth (2009), is that the standard BGS model *overpredicts* the reduction in regions of very low recombination. They reasoned this might occur in regions where U is sufficiently large that multiple deleterious alleles are segregating at any given time. As with beneficial mutations, multiple segregating deleterious sites interfere with each other, reducing the efficiency of selection, and hence less reduction in variation at linked sites. Incorporating this effect into their simulation results gave reductions that were consistent with observed values in very low regions of recombination. We return shortly to the implications of such selection interference in regions of very low recombination.

The second issue is **Muller's ratchet** (Muller 1964; Felsenstein 1974): In a region of very low recombination, the class of chromosomes that carry no mutations may become lost due to drift. Without recombination, there is no way (other than by extremely fortuitous back-mutations) to recover mutation-free chromosomes, so a new class (say harboring just a single mutation) becomes the most fit. These, too, can eventually be lost by drift, leading to double-mutations now being the most fit class, turning the ratchet once again, and so on. The assumption leading to Equation 8.25 is that the zero class is at equilibrium (i.e., is unlikely to be lost in reasonable biological time). Gordo et al. (2002) relaxed this assumption. The approximate condition for the ratchet to operate (i.e., losing the zero class) is that $1/s \gg f_o N_e$, in which case the mean persistence time of a mutation-bearing chromosome ($\sim 1/s$) is larger than the average coalescent time of a mutation-free one ($\sim f_o N_e$). Hence, weak selection and/or small N_e is required for the ratchet to operate. Provided that $f_0 N_e s > 10$, the effective population size is well approximated by Equation 8.25. When the ratchet is operating, in addition to reducing the background variation, an excess of rare alleles is also generated, skewing the site-frequency spectrum. As we will see shortly, this has significant implications if one is tying to distinguish between BGS and recurrent sweeps.

Of course, one imagines that *both* background selection and recurrent sweeps are operating at some level. Kim and Stephan (2000) showed that Equation 8.22b can be modified to given the approximate diversity when both act as

$$\frac{\pi}{\pi_0} \simeq \frac{f_o c}{c + \lambda(c)\, f_0 \gamma\, \kappa} \tag{8.27}$$

where f_0 is the reduction from BGS (Equation 8.25 with complete linkage or Equation 8.26 with recombination), which is also the reduction in effective population size. This changes the scaled strength of selection from $\gamma = 2N_e s$ to $f_0\gamma = 2N_e f_0 s$. The more subtle correction is that the reduction in N_e from BGS (which changes with c) also changes the fixation probabilities for new favorable mutations, so that λ, the product of fixation probability times the number of new adaptive mutations arising per generation, is now a function of the recombination rate c (and indexed as such in Equation 8.27 to remind the reader of this fact). Kim and Stephan (2000) suggest that recurrent sweeps are likely more important in regions of very low recombination, while BGS is more dominant in high recombination regions. Of course, these two forces simply set the levels of background variation which can be significantly disrupted over a region by a very recent sweep. On a practical note, comparison of Equation 8.22b and 8.27 shows that ***ignoring background selection results in an inflated estimate of*** $\lambda\gamma$, and hence an inflated estimate of the rate of adaptation (Kim 2006).

Finally, it is important to stress that background selection is not strictly a phenomena of coding sequences. Indeed, the rather high rate of sequence conservation (and hence functional constraints) seen for noncoding DNA in *Drosophila* has important implications for background selection. Taking into account both its abundance and average level of constraint (Chapter 10), Andolfatto (2005) and Halligan and Keightley (2006) determined that noncoding DNA is likely a much large deleterious-mutational target (by at least a factor of two) than coding DNA. Further, and perhaps most striking, is the recent results from the ENCODE functional genomics project , which found that over 70% of the human genome is transcribed at one time or another, and 80% of the genome having sites that are "biochemically active", displaying some differential biochemical function (such as DNA modifications in specific tissues types). While some (unknown) fraction of these effects may simply be nonfunctional background noise, it is clear that at least some of what had historically been called "noncoding" DNA may play some role in gene regulation (ENCODE Project Consortium 2012), and thus be under some selective constraint. Fu and Akey (2013) offer further commentary on this issue.

Background Selection versus Recurrent Selective Sweeps

While both BGS and recurrent sweeps reduce neutral variation in regions of low recombination, they represent very different processes, purifying selection versus adaptive change. As such, evolutionary geneticists have spent considerable effort trying to distinguish between the two, but no clear answer has yet emerged (Hudson 1994; Andolfatto 2001; Sella et al. 2009; Stephan 2010b; Charlesworth 2009, 2012). As comparison of Equations 8.22c and 8.26c shows, for regions of moderate to high recombination, both processes predict a relationship of the form $\pi/\pi_0 = 1 - b/c$, where b is an unknown to be estimated. Hence, there is little resolution using the relationship between recombination and heterozygosity in moderate- to high-recombination genes. However, such is *not* the case for regions of low (but not too low) recombination. Innan and Stephan (2003) noted that in this region the regression of π on c is convex for recurrent sweeps and concave for BGS (compare Equations 8.22b and 8.26a). They applied this approach to a set of low-recombination X-linked genes in *D. melanogaster*, finding that recurrent sweeps gave a much better fit than BGS. However, when two highly selfing species of tomatoes (*Lycopersicon*) were examined, BGS provided the better fit. In humans, Hellmann et al. (2008) found that recurrent sweeps gave a better fit that BGS, but cautioned that this may simply be an artifact of the simplistic nature of the BGS model leading to Equation 8.26a (i.e., assuming no variation in s).

One distinct prediction between BGS and recurrent sweeps is the expected effect on the site-frequency spectrum. Under the "strong" version of BGS, deleterious mutations have strong effects ($4N_e s \ll -1$) and are quickly removed by selection. In this case, the effect is to simply lower N_e to $f_o N_e$, but not otherwise change the frequency spectrum (Charlesworth et al. 1993, 1995). Conversely, under selective sweeps, an excess of sites with rare alleles is expected (Braverman et al. 1995; Kim 2006). A negative value of Tajima's D statistic (Chapter 9) indicates an excess of rare alleles, and negative D values are often (but not always) associated with genes showing reduced variation in regions of low recombination in *Drosophila* (e.g., Langley et al. 2000). An interesting study is by Andolfatto and Przeworski (2001), who found a highly significant positive association ($r^2 = 0.31, p = 0.002$) between Tajima's D and recombination rate in a study of 29 *D. melanogaster* genes — as the recombination rate decreased, D became more negative. Such an observation is consistent with a recurrent sweep model, but not with a strong BGS model.

Nevertheless, while findings like this are suggestive of recurrent selection as opposed to BGS, they are not as conclusive as one might think. A model with *weakly* deleterious alleles can generate an excess of rare alleles (Tachida 2000; Comeron and Kreitman 2002; Comeron et al. 2008). While a weak BGS selection model will not generate a significant reduction in variability (Golding 1997; Neuhauser and Krone 1997; Przeworski et al. 1999), a process generating both strong and weak deleterious alleles could generate both a reduction in variability *and* a negative skew in the frequency spectrum (Gordo et al. 2002). Likewise, a more careful analysis of BGS under very low recombination shows that selective interference (Kaiser and Charlesworth 2009) can also generate negative D. More generally, BGS and recurrent sweeps are but two models of selection. Equally realistic models of linkage to sites experiencing fluctuating selection coefficients can generate the same patterns as sweeps (Gillespie 1997, 2000).

Sweeps, Background Selection, and Substitution Rates

Both recurrent sweeps and background selection are expected to lower the effective population size N_e, and hence reduce variation at tightly linked sites. Do these processes also influence the rate of divergence at such sites? For *strictly* neutral alleles ($s = 0$), changes in N_e have no effect on the substitution rate, as this is simply the neutral mutation rate μ (Chapter 2). However, when alleles have a *distribution* of fitness effects (s may be very small, but not zero), this is no longer true. Accepting the view that many mutations may be slightly deleterious (Ohta 1973, 1992, 2002), in smaller populations an allele can be **effectively neutral** ($4N_e|s| < 1$), while being selected against in larger populations (when $4N_e s \ll -1$). In genomic regions where the effect of recurrent sweeps and/or background selection is expected to be strong (such as regions of low recombination), an *increase* in the divergence rate might be expected, as the fraction of new mutations that are effectively neutral increases. Likewise, in such regions, the rate of adaptive changes may decrease, as weakly favorable mutations are overpowered by the effects of drift, reducing their fixation rates. This alteration of the substitution pattern (through fixation of a greater fraction of weakly deleterious alleles) and decrease in the substitution rate of adaptive changes (reduced fixation of weakly-favorable alleles) are both examples of the Hill-Robertson effect.

Example 8.12. Modern rice was domesticated from *Oryza rufipogon* to form the indica (*Oryza sativa indica*) and japonica (*O. sativa japonica*) lineages (Huang et al. 2011). Lu et al. (2006) examined the ratio of the replacement to silent substitution rates, K_a/K_s (Chapter 10), between both these two subspecies and an outgroup, *O. brachyantha*. In a comparison of over 15,000 genes, the K_a/K_s ratio for divergence between indica and japonica was 0.498. Conversely, in a comparison of roughly 5000 genes between japonica and the outgroup, K_a/K_s = 0.259, a highly significant difference. This increase in K_a/K_s between the domesticated lines occurs throughout the genome, with most regions showing elevated values when comparing the two modern cultivars. Regions of lower recombination showed the largest K_a/K_s values, with a highly significant negative regression of K_a/K_s on recombination rate. The authors interpreted these data as suggesting an increase in the fixation rate of deleterious alleles due to a decrease in N_e during the domestication of both of these lines. If the increase in K_a/K_s ratios was due to the accelerated fixation of favorable alleles, this ratio should *increase* with recombination rate, as the effective population size is higher in regions of higher recombination, increasing the fixation rate of favorable alleles. Conversely, the fixation rate of (slightly) deleterious alleles should increase with decreasing recombination, as the smaller N_e in these regions allows more of these alleles to behave as if effectively neutral. The initial founding of lines during the early phases of domestication reduced N_e, a process that authors suggest was exacerbated by strong selfing, and hence reduction of the effective amount of recombination throughout the genome. This, in turn, resulted in selective sweeps associated with the fixation of domestication genes influencing larger regions of the genome.

To support this view of an increased fixation rate for slightly deleterious alleles, the authors used a regression method developed by Tang et al. (2004) based on the relationships among the K_a/K_s ratios associated with the 75 possible single-base replacement changes (where a single nucleotide change in one codon coverts it to a replacement codon). Tang et al. showed that the general pattern over the genome is that the proportional relationships over the various K_a/K_s ratios for different codon pairs remains constant. This pattern was observed when comparing the divergence in the wild rices *rufipogon* and *brachyantha*. For indica and japonica, however, a disproportional amount of change involving radical amino acids replacements over conservative replacements was observed, with the authors estimating that a quarter of the replacement substitutions were likely deleterious.

As the above example highlights, the direction of a potential change in the rate of replacement substitutions as recombination decreases is a function of whether there are more weakly positively-selected alleles (rate goes down) or weakly negative-selected alleles (rate goes up). Betancourt and Presgraves (2002) found in a comparison of roughly 250 genes between *D. melanogaster* and *D. simulans* that the nonsynonymous divergence rate is reduced in regions of low recombination, consistent with reduced fixation of weakly positive alleles in these regions due to a reduction in N_e. However, their gene set contained a large number of male accessory gland proteins (*Acps*), which are rapidly evolving, and hence might have biased their results. When the *Acps* genes were removed from the analysis, there was no significant relationship between replacement rates and recombination. Among *Acps* genes, rapid protein evolution was largely confined to regions of high recombination, again consistent with a reduction in N_e retarding the rates of evolution for these genes. Conversely, Haddrill et al. (2007) examining genes in regions of *no* recombination in *D. melanogaster* and *D. yakuba* found *elevated* rates of replacement substitution (as seen in the rice example above), consistent with weakly deleterious alleles behaving as if efficiently neutral due to reduction in N_e in low recombination regions. Similarly, in comparisons between the small largely nonrecombinational "dot" chromosome of *D. americana* and its other autosomes, Betancourt et al. (2009) found an increased rate of replacement substitutions on the dot. Further, estimates of the fraction α of adaptive substitutions (using methods discussed in Chapter 10) were significantly smaller for the dot than for the other autosomes, suggesting that the increase in replacement substitutions was largely due to the fixation of slightly deleterious alleles. Finally, Bullaughey et al. (2008) found no effect of recombination on rates of protein evolution over the human, chimp, and rhesus macaque lineages. Genes in the regions of lowest recombination did not evolve at rates different from other genes. It is perhaps not surprising that no consistent result on divergence as a function of recombination rate has emerged, as the nature of any potential signal depends on the distribution of selection coefficients relative to the reduction in N_e in low recombination regions.

Since both BGS and recurrent sweeps can reduce diversity in regions of low recombination, we can also ask the related question of whether the amount of *divergence* at a site influences the amount of linked neutral variation. Under a strictly neutral model (Chapter 2), the amount of divergence and polymorphism is a function of the mutation rate, so that sites with higher divergence should also display higher levels of polymorphism (when jointly compared in the same population giving all sites a common N_e). Under recurrent sweeps, if a gene shows a high rate of divergence, this might imply more frequent sweeps, and therefore lower diversity due to the local reduction in N_e that accompanies these sweeps. Such a negative correlation between synonymous nucleotide diversity and the substitution rate at replacement sites was seen in *Drosophila melanogaster* (Andolfatto 2007), *D. simulans* (Macpherson et al. 2007), *D. miranda* (Bachtrog 2008; Jensen and Bachtrog 2010), *D. pseudoobscura* (Jensen and Bachtrog 2010), European rabbit (*Oryctolagus cuniculus*, Carneiro et al. 2012), European aspen (*Populus tremula*, Ingvarsson 2010), and humans (Cai et al. 2009).

These last authors suggested that selection at linked sites in humans appears to reduce nucleotide diversity by 6% genome-wide and 11% in the gene-rich half of the genome. McVicker et al. (2009) obtained even higher values in humans, between 19 and 26% for autosomes and between 12 and 40% on the X. One reason for this apparent discrepancy is that Cai et al. specifically excluded regions immediately adjacent to genes, which likely are under some of the strongest selection. However, for both these studies, the authors caution that the reduction could be due to recurrent sweeps, BGS, or (most likely) a combination of both. More recently, several groups have suggested that the reduction in nucleotide diversity in humans is better explained by BGS (Hernandez et al. 2011; Lohmueller et al. 2011; Alves et al. 2012). Hernandez et al. found that classic sweeps were rare in recent human history. These authors found local troughs in nucleotide diversity around amino-acid substitutions, but these essentially the same as seen around synonymous substitutions, more consistent with BGS than recurrent sweeps. This is in sharp contrast to the findings of Sattah et al. (2011) in *Drosophilia* where the pattern was quite different, and consistent with recurrent sweeps. Lohmueller et al. noted that the correlation between neutral diversity and nonsynonymous divergence would be more negative under recurrent sweeps than that value observed in humans, with a model based on BGS fitting the pattern better

Finally, a region with a very low level of divergence may be under strong constraints, with most new mutations being deleterious. Under the BGS model, regions that are slow evolving should also have reduced nucleotide diversity, reflecting a lower local value of N_e. McVicker et al. (2009) scanned for conserved genomic regions using humans and four other primates. Surprisingly, less than 25% of such detected sequences corresponded to coding regions. Using adjacent less-conserved sites as neutral proxies, they found that neutral diversity is lower around highly conserved sites. Of course, such a pattern could easily be generated under the neutral theory by a simple reduction in the mutation rate, decreasing both variation and divergence. As a control for this, the authors examined whether the divergence in these presumed neutral regions with reduced human diversity also showed reduced divergence between human and dog. While there was a slight reduction, it only accounted for the small part of the overall trend. Hence, reduced mutation rates are likely not sufficient to account for this observation.

Sweeps, Background Selection, and Codon Usage Bias

One of the most sensitive indicators of localized changes in N_e is provided by the behavior of sites under very weak selection ($N_e|s| \sim 1$). Under this setting, weakly favorable sites are still selected for, while weakly deleterious sites are selected against. However, a small decline in N_e (be it from recurrent sweeps ad/or background selection), or s (from interference among multiple segregating selected alleles) can make a significant fraction of these weakly-selected sites behave neutrally.

Although synonymous codons are typically used as proxies for neutral sites, the observation of **codon usage bias** (the nonrandom use among the set of all synonymous codons for a given amino acid) in many organisms shows that this is only approximately correct. In reality, synonymous sites often appear to be under very weak selection for **optimal** (or **preferred**) codons, which are more frequent than expected from genome nucleotide frequencies. As potential sites under very weak selection, synonymous codons may be rather sensitive (at least in some species) to subtle changes in N_e. We first examine the evidence suggesting selection on synonymous codons and the genomic patterns of codon usage before considering what this might tell us about selection at linked sites. We stress that local changes in N_e are expected to generate *subtle* signals at weakly-selected sites that can be detected only when one examines hundreds of genes.

The classic view of codon bias is that selection is likely to be stronger on more highly expressed genes, so that bias is expected to vary over genes. Further, the actual strength of selection, postulated to arise from improved transitional efficiency and accuracy due to the optimal codon matching the most abundant tRNA for that amino acid, is expected to be quite weak. So weak, in fact, that for an average gene, bias is expected to be significant only in organisms with large effective population sizes. While this general underlying theme holds, it is not the whole story. There is a general trend for codon usage bias to be more pronounced in organisms with larger census population sizes, but a surprising observation is that bacteria, yeast, and *Drosophila* all have roughly similar levels (Powell and Moriyama 1997), despite their perceived great differences in effective population size. This is tantalizingly reminiscent of Lewontin's (1974) observation that the level of average protein heterozygosity within a species (the surrogate for genetic variation at the time) is much narrower that than expected given the range of census population sizes.

One of the first studies to suggest that *segregating* synonymous alleles may be under selection was the work of Akashi (1995) in *Drosophila*. By using an outgroup, Akashi polarized segregating alleles, determining which was ancestral allele (fixed in a sister species) and which is the derived new mutation. For a particular amino acid that shows codon usage bias, **preferred codons** are those used more frequently than expected, while **unpreferred codons** are used less frequently. Segregating and fixed differences were then placed into two categories: those involving a preferred codon that mutated to an unpreferred one (denoted by $P \rightarrow U$), and those involving an unpreferred codon mutating to a preferred codon ($U \rightarrow P$). Under the expectation that $P \rightarrow U$ alleles are slightly selected against, and $U \rightarrow P$ weakly selected for, Akashi compared the divergence to polymorphism ratio of $P \rightarrow U$ to that for $U \rightarrow P$. If unpreferred codons are selected against, we expect a higher ratio of polymorphism (ratio of segregating U versus P alleles) to divergence (ratio of fixed U versus P mutations), as alleles under weakly deleterious selection can segregate, but are unlikely to be fixed. A significantly higher ratio was indeed seen in both *D. simulans* and *pseudoobscura* (Akaski and Schaeffer 1997), but an excess of unpreferred fixations was seen in *D. melanogaster*, suggesting far weaker codon selection on the 28 *melanogaster* genes examined, which the authors attributed to the three to six fold reduction in N_e in *D. melanogaster* relative to *simulans*.

Example 8.13. Maside et al. (2004) examined codon usage in *D. americana*, a member of the *virilis* species group. Using *virilis* as an outgroup, they observed 84 synonymous substitutions (fixed differences or divergence) between the two species and 144 segregating synonymous sites within *americana*. Classifying these as either a $P \rightarrow U$ or $U \rightarrow P$ showed the following pattern:

	Substitutions	Polymorphic (*americana*)	Polymorphism/Divergence
$P \rightarrow U$	52	124	2.38
$U \rightarrow P$	32	20	0.62

This roughly four-fold higher polymorphism to divergence ratio for the putative deleterious mutations $P \rightarrow U$ is highly significant (Fisher's exact test gives $p = 6.4 \times 10^{-5}$). Further, if this class is indeed deleterious, we would expect these mutations to be at lower frequencies in the sample than $U \rightarrow P$ mutations, and such a significant difference was observed. This difference in the site-frequency spectrum was first noticed by Akashi (1999) for *D. simulans*, which was shifted towards lower frequencies for unpreferred mutations and towards higher frequencies for preferred mutations.

Given the above evidence for selection against unpreferred codons, how strong is selection? Using the Poisson random field (PRF) method for analysis of the pattern of fixed differences and polymorphic sites (examined in detail in Chapter 10), estimates of $N_e|s| \sim 1$ were obtained for *simulans* and *pseudoobscura* (Akashi 1995; Akaski and Schaeffer 1997). An alternative approach to estimate $N_e|s|$ follows from Equation 7.36, which gives Li's (1987) expression for the expected frequency $\widetilde{p}$ of a preferred allele at the mutation-selection-drift equilibrium. In the notation of this chapter (where we use $2\gamma = 4N_e s$ for the strength of selection, as S is used to denote the number of segregating sites), this becomes

$$\widetilde{p} \simeq \frac{\exp(2\gamma)}{\exp(2\gamma) + \zeta} \tag{8.28}$$

where γ is the scaled strength of selection for preferred codons and $\zeta = \mu_{P \rightarrow U} / \mu_{U \rightarrow P}$ measures any mutation bias (also see Bulmer 1991; McVean and Charlesworth 1999, 2000; Zeng and Charlesworth 2009, 2010; Zeng 2010). If ζ is known, Equation 8.28 can be used to directly estimate γ for a given synonymous codon set (averaged over genes). Maside et al (2004) offered an alternative (but related) procedure that does not involve estimating ζ. They showed that the fraction p_U of segregating sites where the *derived* allele is the unpreferred synonymous codon (i.e., $P \rightarrow U$ mutations as opposed to the dervied allele being a $U \rightarrow P$ mutation) in a sample of n alleles can be expressed as a function of γ alone, namely

$$p_U = \frac{\exp(2\gamma)}{\exp(2\gamma) + I(n, -\gamma)/I(n, \gamma)} \tag{8.29a}$$

where

$$I(n, \gamma) = \int_0^1 [\, 1 - x^n - (1 - x)^n \,] \, \frac{1 - e^{-2\gamma(1-x)}}{x(1-x)(1 - e^{-2\gamma})} \, dx \tag{8.29b}$$

The term in the brackets is the probability of a polymorphic sample given n sequences (Equation 2.36b), while the second term is the density for the allele frequency of a gene under additive selection (Equation 10.14a). Since Equation 8.29a gives the expected probability that a segregating synonymous site has a $P \rightarrow U$ mutation, the probability that we see k such sites over all S segregating sites follows a binomial distribution, $k \sim \text{Binom}(p_U, S)$, where S is the sample size and p_U the success parameter. The resulting log-likelihood becomes

$$\ln(L) = k \ln(p_U) + (S - k) \ln(1 - p_U) \tag{8.29c}$$

Here S, k, n are the observed values, and one plots $\ln(L)$ as a function of γ to find the ML estimate. If one assumes the same γ value over a set of codons, the total likelihood is just the product of Equation 8.29c over all appropriate sites. Using this approach, which measures contemporaneous selection coefficients (unlike PRF estimates which use divergence data, and hence are influenced by historical selection), Maside et al. obtained estimate of $N_e|s| \simeq 0.65$ in *D. americana*.

Thus, for several *Drosophila* species, the strength of selection on synonymous codon usage is roughly $N_e|s| \simeq 1$, offering the possibility that small localized genomic changes in N_e can significantly impact on codon bias. The prediction is that ***codon bias is reduced in regions where*** N_e ***is lowered***, reducing the strength of selection. Three observations offer support for this, with bias being less extreme: 1) in regions of low recombination, 2) for genes that are rapidly diverging, and 3) in the middle of long exons. We examine each of these observations in turn. Note that most of these observations come from *Drosophila*, which seems to have the requisite $N_e|s| \sim 1$ weak selection condition on synonymous codons. Organisms where the scaled strength of selection is weaker (i.e., those with much smaller N_e) or much stronger (i.e., those with much larger N_e) might not show these trends, as an order of magnitude change in the N_e for a genomic region will still leave drift overpowering selection (small baseline N_e) or selection still overpowering drift (large baseline N_e).

There are numerous reports of codon bias depending, to some extent, on recombination rates in *Drosophila*. Kliman and Hey (1993) examined roughly 400 loci in *D. melanogaster*, finding that codon bias is reduced in regions of low recombination. The relationship was not linear, rather was only apparent for genes in the lowest regions of recombination. Marais et al. (2001; Duret and Galtier 2009) suggested this relationship results from a mutation bias towards G and C bases (which are commonly used in the optimal codon) in regions of high recombination. However, a more detailed analysis by Hey and Kliman (2002) looking at 13,000 genes in *melanogaster* again found a weak, but significant, positive correlation between bias and recombination rate, although the roughly 9000 genes in region of modest to high recombination rate ($c > 1.5$ cM/Mb) showed no association. They further showed that subtle differences in how recombination is measured could account for the negative result of Marais et al. Similarly, Haddrill et al. (2007) found essentially no codon bias for genes in *melanogaster* and *yakuba* residing in regions with no recombination, and Betancourt et al. (2009) found that a significantly smaller fraction of genes on the small ("dot") chromosome of *D. americana* used optimal codons relative to sites on larger chromosomes.

In addition to these regional effects over the scale of a small chromosomal segment, there are also reports of effects on much finer scales, namely gene-by-gene and even different regions within the same gene. Genes undergoing multiple sweeps (and hence higher rates of substitutions) might be expected to have lower effective population sizes, and hence less codon bias. In a study involving roughly 250 genes, Betancourt and Presgraves (2002) found those with higher replacement rates tended to show less codon usage bias in both *melanogaster* and *simulans*. Maside et al. (2004) examined over 600 *melanogaster* genes, also finding a negative association between rates of replacement substitution and codon bias. However, they also noted that both codon bias and replacement rates are correlated with gene expression, so perhaps the latter is the driver for the correlation. Andolfatto (2007) found both reduced codon bias, as well as reduced synonymous-site diversity, in rapidly evolving proteins in a survey of roughly 140 proteins on the high-recombination region of the X chromosome from an African population of *D. melanogaster*, and similar results are reported by Bachtrog (2008) for *D. miranda*. While most observations are restricted to *Drosophila*, Ingvarsson (2010) found a weakly negative (but not significant) relationship between codon bias and protein evolution rates in European aspen (*Populus tremula*).

On an even finer scale are reports for a correlation between codon bias and gene length in *Drosophila* (Comeron et al. 1999). For short genes (less than 750 bp), tighter linkage results in reduced bias. This effect is less for genes with longer coding regions. Moreover, the length of a coding region is negatively correlated with bias (longer genes have less bias) over all recombination values. Strikingly, Comeron and Kreitman (2002) found that codon bias decreases in the *middle* of long exons, which likely accounts for the reduced bias over longer genes. A more detailed analysis by Qin et al. (2004) showed that codon bias decreases at the ends, as well as the middle, of long genes in *Drosophila*, while yeast and several species of bacteria showed no such pattern. Comeron and Guthrie (2005) used Equation 8.29 to estimate the strength of selection γ on synonymous codons on long versus short genes, finding the former had significantly reduced γ values. Consistent with relaxation of selection, longer exons also had higher rates of synonymous substitution, as would be expected if reduction in N_e made weakly-deleterious synonymous mutations behave in a more neutral fashion.

All of these signals of reduction in N_e resulting in more neutral patterns of codon usage are consistent with the effects of selection at linked sites. Both recurrent sweeps and background selection could generate the reduction in bias in regions of low recombination. Likewise, lower codon bias for genes with high replacement substitution rates is consistent with recurrent sweeps (Kim 2004). The most interesting observations, however, are the very fine scale differences, in particular the decrease in bias in the middle of long exons. Loewe and Charlesworth (2007) suggest that background selection could generate such a pattern, with the edges of exons being linked to fewer regions under selection, and hence experiencing a lower total deleterious mutation rate U. Regions in the middle of exons can have deleterious mutations arise for some distance on both sides of them, increasing their U value, creating a local decrease in N_e. These very-fine scale effects are very sensitive to recombination. Hey and Kliman (2002) found that, when measured by number of genes per kilobase, density had no effect on codon bias. However, very tightly spaced genes did show decreased bias, showing that the potential linked effects of selection operates over very short distances.

One explanation for these very short range effects is *interference among selected sites*. Background selection and recurrent sweeps typically assume alleles are under strong selection, so they have only a short persistence time in the population. Conversely, alleles under weak selection segregate for longer periods of time, allowing for multiple segregating mutations of weak effect within a gene. In such cases, selection at other sites may generate some interference, reducing the efficiency of selection. For example, if selected mutations are in negative LD, this reduces the additive variance in fitness (Chapter 16) and hence the efficiency of selection (Chapters 5, 6). Likewise, within a set of alleles that are nearly selectively equivalent, drift can occur, reducing the efficiency of selection on any particular allele. This reduction in efficiency from selective interference has been called **small-scale Hill-Robertson** (Comeron et al. 1999), **weak selection Hill-Robertson interference** (McVean and Charlesworth 2000), and **interference selection** (Comeron and Kreitman 2002). For example, if multiple weak positively-selected alleles are segregating in a tightly linked region (such as multiple preferred codons within an exon), they mutually interfere with each other, resulting in weaker section and a smaller codon usage bias. The same is true for a collection of weakly deleterious alleles. The key is extremely tight linkage. Simulation studies (Comeron and Kreitman 2002; Comeron et al. 2008) show that interference selection can indeed produce a decrease in codon bias in the middle of long exons, with bias decreasing with the number of selected sites. Its effect, however, is *extremely* local, except in regions of very, very low recombination. McVean and Charlesworth (2000) show that interference selection can also account for the puzzling observation of the relative insensitivity of bias to changes in N_e (provided it is sufficiently large) seen in cross-species comparisons (Powell and Moriyama 1997). When interference selection is present, it tends to moderate the effects of selection, so that the expected bias is relatively similar over several orders of magnitude in N_e. While BGS and recurrent sweeps reduced codon bias in regions of low recombination, they found that interference selection can reduce bias even in genes in regions of moderate recombination, because there is still tight linkage over very small regions which might be segregating multiple sites under weak selection. As noted by Comeron and Kreitman (2002), exons and their adjacent control regions are prime candidates for interference selection as the physical clustering of functional sites offers the possibility of weak selection over a number of tightly linked sites.

Despite these general patterns, there are still unexplained aspects of codon bias. In particular, the observation that X-linked genes tend (as a group) to have higher codon bias than autosomal genes in *D. melanogaster* is intriguing (reviewed by Campos et al. 2012). Give that one would expect the effective population size of X-linked genes to be lower than those for autosomes ($\sim 3/4$), the above agruments suggest that X-linked bias should be reduced. Campos et al. examined whether increased recombination on the X might counter this, by reducing Hill-Robertson effects from selection at linked sites. The argument for higher average recombination on the X is that they spend 2/3 of their time in females, which have normal recombination, and 1/3 of their time in males with no recombination. Conversely, autosomes spend equal amounts of time in both sexes, resulting in a 33% higher expected effective recombination rates on the X. When comparing sets of X and autosomal genes with similar effective recombination rates, Campos et al. still observed higher X-linked bias, suggesting that this pattern results from stronger selection for preferred codons for X-linked genes in *melanogaster*, rather than localized changes in effective population size (also see Zeng and Charlesworth 2010).

A Paradigm Shift Away from the Neutral Theory of Molecular Evolution?

As ably summarized by Charlesworth (2010), molecular population genetics has a rich and dynamic history, with roots back to Darwin. Its current focus traces back to the **neutral theory of molecular evolution**, born in the late 1960's in response to the higher than expected levels of protein polymorphism found in natural populations (Kimura 1968; King and Jukes 1969), and gained strength through the 1980's as more molecular data became available (Kimura 1983). Under its initial version, the vast majority of new mutations were either assumed to be rather strongly deleterious or neutral. As selection is expected to rapidly remove the former, such mutations contribute little to the levels of polymorphism and even less to divergence. Under this theory, advantageous mutations can indeed occur, but are assumed to be extremely rare, and rapidly fixed (or lost) when they arose, resulting again in little impact of polymorphisms and (at best) modest impact on divergence. Given these assumptions, it then follows that most fixed differences between populations/species and most segregating variation within a population/species is largely due to neutral variation. Ohta (1973, 1992, 2002) presented an important modification, the **nearly-neutral theory**, allowing for slightly deleterious alleles, which could be close to being effectively neutral, and hence contribute significantly to polymorphisms and some to divergence (reviewed by Akashi et. al. 2012). A key prediction from either version of the neutral theory is that regions with fewer functional constraints (and therefore a higher fraction of neutral or nearly neutral mutations) evolve faster. This prediction is strongly supported by the observation of slower substitution rates at replacement sites, faster rates at synonymous sites, and ever faster rates in pseudogenes.

The key feature of all versions of the neutral theory is that while purifying selection can be very common, *adaptive evolution at the molecular level is rare*, so that most segregating alleles and most fixed sites are, at best, effectively neutral. The flood of molecular data from the genomics era now calls this key assumptions into question (Hahn 2008; Rockman 2012), and leads to a more nuanced view of the nature of stochastic changes in allele frequencies that can dominate weakly-selected sites (but see Nei at el. 2010 for a contrary opinion). As discussed in Chapter 10, the estimates of high α values (fraction of replacement substitutions that are adaptive) in some species is strongly at odds with the view of neutral or nearly neutral theories. A second potential problem is genomic effects from selection at linked sites, the most celebrated of which is the correlation between recombination rates and levels of variation. If due to background selection, this observation is still consistent with the classic neutral theory, with selection generating this correlation as a consequence of removing new deleterious mutations. However, if periodic selective sweeps generate this pattern, then much of the genome is impacted by *positive selection*, either directly or indirectly through the effects of selection at linked sites. Finally, observations consistent with selective constraints on silent sites and even noncoding DNA in some species (reviewed in Chapter 10) is also somewhat problematic for the neutral theory. This is especially true given the recent findings that upwards of 80% of the human genome may have some functional role (ENCODE Project Consortium 2012), at least as indicated by differential expression or modification of sites over various human cell lines. This is almost certainly an inflated figure (Fu and Akey 2013), as, for example, a cell type might have reduced genome-wide stringency on control of transcription, or methylation, resulting in a large number of sites that would be differently modified (or transcribed) over cell types, and hence be deemed functional. However, it does caution one that a much larger collection of sites in the human genome may be functionally in play. While the removal of new deleterious mutations falls under the neutral theory umbrella, the converse, fixation of slightly favored sites (such as the fixation of a silent mutation to a preferred codon), is an example of positive selection. The inescapable conclusion is that weak selection is occurring throughout the genome, and patterns of variation are shaped by selection at linked sites. These effects can be over quite small scales, on the level of differences between the ends and middle of a long exon, presumably due to interference among weakly selected sites.

The great irony of a deeper appreciation for how rampant selection (and especially weak selection) is throughout the genome is that it likely makes *more* alleles behave as if they are effectively neutral. Kimura's original grand vision of the role of selection acting as a giant filter, through which only neutral and a very few advantageous alleles pass, now appears to being replaced by the role of selection throughout the genome making weakly selected alleles behave in a more neutral fashion. Chapter 3 introduced Gillespie's (1997, 2000) concept of **genetic draft**, where the frequency of neutral alleles fluctuate more from their random association with ongoing sweeps than they do from drift. While drift may be important in small populations, in sufficiently large populations, sweep-generated stochastic fluctuations of allele frequencies can still overpower weak selection. A combined analysis of drift and draft by Schiffels et al. (2011) found that when the absolute value of a selection coefficient of an allele is less that $1/(2N) + \lambda$, it is effectively neutral, a feature they refer to as **emergent neutrality**. When N is small, the rate of adaptation λ (which scales in NU_b for small-medium values of N) is likely small, and drift dominates. Conversely, when N is very large, $1/N$ can be dominated by λ. For very large N, interference among sweeps strongly constrains λ, which makes the transition from scaling as NU_b to scaling as $\log(NU_b)$. Thus, at *every* population size stochastic fluctuations in allele frequencies are important, but the forces underlying them differ. Drift dominates at small populations, draft in large populations. Once the draft stage is reached, since λ increases within N (albeit very slowly), *more* alleles become effectively neutral as population size increases. When coupled with selective-inference among weakly-selected sites, this leads to a view where much of the variation in a large population may be effectively neutral. However, is not due to a lack of positive selection, but rather a *consequence* of it.

Aguadé, M., N. Miyashita, and C. H. Langley. 1989. Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. *Genetics* 122: 607–615. [8]

Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* 139: 1067–1076. [8]

Akashi, H. 1999. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. *Gene* 238: 39–51. [8]

Akashi, H., N. Osada, and T. Ohta. 2012. Weak selection and protein evolution. *Genet.* 192: 15–31. [8]

Akashi, H., and S. W. Schaeffer. 1997. Natural selection and the frequency distribution of "silent" DNA polymorphisms in *Drosophila*. *Genetics* 146: 295–307. [8]

Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver, D. A. Nickerson, and L. Kruglyak. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2: e286. [8]

Alves, I., A. S. Hanulova, M. Foll, and L. Excoffier. 2012. Genomic data reveal a complex making of humans. *PLoS Genet.* 8: e1002837. [8]

Andolfatto, P. 2001. Adaptive hitchhiking effects on genome variability. *Curr. Opin. Genet. Develop.* 11: 635–641. [8]

Andolfatto, P.. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437: 1149-1152. [8]

Andolfatto, P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res.* 17: 1755–1762. [8]

Andolfatto, P., and M. Przeworski. 2001. Regions of lower crossing over harbor more rare variants in African populations of *D. melanogaster. Genetics* 158: 657–665. [8]

Atwood, K. C., L. K. Schneider, and F. J. Ryan. 1951a. Periodic selection in *Escherichia coli. Proc. Natl. Acad. Sci. USA* 37: 146–155. [8]

Atwood, K. C., L. K. Schneider, and F. J. Ryan. 1951b. Selective mechanisms in bacteria. *Cold Spring Harbor Symp. Quant. Biol.* 16: 345–355. [8]

Bachtrog, D. 2008. Similar rates of protein adaptation in *Drosophila miranda* and *D. melanogaster*, two species with different current effective population sizes. *BMC Evol. Biol.* 8: 334. [8]

Barrett, R. D. H., and D. Schluter. 2008. Adaptation from standing genetic variation. *Trend. Ecol. Evol.* 23: 38–44. [10]

Barton, N. H. 1995. Linkage and the limits to natural selection. *Genetics* 140: 821–841. [8]

Barton, N. H. 1998. The effect of hitch-hiking on neutral genealogies. *Genet. Res. Camb.* 72: 123–133. [8]

Barton, N. H. 2000. Genetic hitchhiking. *Phil. Trans. R. Soc. Lond.* B 355: 1553–1562. [8]

Begun, D. J., and C. F. Aquadro. 1991. Molecular population genetics of the distal portion of the *X* chromosome in *Drosophila*: Evidence for genetic hitchhiking of the *yellow-achaete* region. *Genetics* 129: 1147–1158. [8]

Berry, A. J., J. W. Ajioka, and M. Kreitman. 1991. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics* 129: 1111–1117. [8]

Betancourt, A. J., and D. C. Presgraves. 2002. Linkage limits the power of natural selection in *Drosophila. Proc. Natl. Acad. Sci. USA* 99: 13616–13620. [8]

Betancourt, A. J., J. J. Welch, and B. Charlesworth. 2009. Reduced effectiveness of selection caused by a lack of recombination. *Curr. Biol.* 19: 655–660. [8]

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140: 783–796. [8]

Brinkman, M. A., and K. J. Frey. 1977. Yield component analysis of oat isolines that produce different grain yields. *Crop Sci.* 17: 165–168. [8]

Bullaughey, K., M. Przeworski, and G. Coop. 2008. No effect of recombination on the efficacy of natural selection in primates. *Genome Res.* 18: 544–554. [8]

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129: 897–907. [8]

Cai, J. J., J. M. Macpherson, G. Sella, and D. A. Petrov. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genetics* 5: e1000336. [8]

Campos, J. L., K. Zeng, D. J. Pakrer, B. Charlesowrth, and P. R. Haddrill. 2012. Codon usage bias and effective population sizes on the X chromosome versus autosomes in *Drosophila melanogaster. Mol. Biol. Evol.* 30: 811–823. [8]

Carneiro, M., F. W. Albert, J. Melo-Ferreira, N. Galtier, P. Gayral, J. A. Blanco-Aguiar, R. Villafuerte, M. W. Nachman, and N. Ferrand. 2012. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. *Mol. Bio. Evol.* 29: 1837–1849. [8, 10]

Charlesworth, B., 1992. Evolutionary rates in partially self-fertilizing species. *Am. Nat.* 140: 126–148. [8]

Charlesworth, B. 1996. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet. Res.* 68: 131–149. [8]

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* 10: 195–205. [8]

Charlesworth, B. 2010. Molecular population genomics: a short history. *Genet. Res.* 29: 397–411. [8]

Charlesworth, B. 2012. The effects of deleterious mutations on evolution at linked sites. *Genetics* 190: 5-22. [8]

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134: 1289–1303. [8]

Charlesworth, D., B. Charlesworth, and M. T. Morgan. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* 141: 1619–1632. [8]

Chevin, L.-M., S. Billiard, and F. Hospital. 2008. Hitchhiking both ways: effects of two interfering selective sweeps on linked neutral variation. *Genetics* 180: 301–316. [8]

Chevin, L.-M., and F. Hospital. 2008. Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genet.* 180: 1645—1660. [8]

Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Vilarreal Jr., M. Dickson, J. Grimwood, J. Schmutz, R. M. Myers, D. Schluter, and D. M. Kingsley. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* 307: 1928–1933. [8]

Comeron, J. M., and T. B. Guthrie. 2005. Intragenic Hill-Robertson interference influences selection intensity on synonymous mutations in *Drosophila*. *Mol. Biol. Evol.* 22: 2519–2530. [8]

Comeron, J. M., and M. Kreitman. 2002. Population, evolutionary and genomic consequences of interference selection. *Genetics* 161: 389–410. [8]

Comeron, J. M., M. Kreitman, and M. Aguadé. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics* 151: 239–249. [8]

Comeron, J. M., A. Williford, and R. M. Kliman. 2008. The Hill-Roberston effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity* 100: 19–31. [8]

Duret, L., and N. Galtier. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Ann. Rev. Genomics Hum Genet.* 10: 285–311. [8]

Durrett, R., and J. Schweinsberg. 2004. Approximating selective sweeps. *Theor. Pop. Biol.* 66: 129–138. [8]

Dykhuizen, D. E. 1990. Experimental studies of natural selection in bacteria. *Ann. Red. Ecol. Syst.* 21: 373–398. [8]

Enard, W., M. Przeworski, S. E. Fisher, C. S. L. Lai, Vi. Wiebe, T. Kitano, A. P. Monaco, and S. Pääbo. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 418: 869–872. [8]

Enattah, N. S., and 26 others. 2007. Evidence of still-ongoing convergence evolution of the lactase persistence T_{-13910} alleles in humans. *Amer. J. Hum. Gen.* 81: 615–625. [8]

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74. [8]

Etheridge, A., P. Pfaffelhuber, and A. Wakolbinger. 2006. An approximate sampling formula under genetic hitchhiking. *Ann. Appl. Prob.* 16: 685–729. [8]

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* 3: 87–112. [8]

Ewing, G., J. Hermisson, P. Pfaffelhuber, and J. Rudolf. 2011. Selective sweeps for recessive alleles and for other modes of dominance. *J. Math. Biol.* 63: 399–431 [8]

Fay, J. C., and C.-I. Wu. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405–1413. [8]

Felsenstein, J. 1974. The evolutionary advantage of recombination. *Genetics* 78: 737-756. [8]

Fu, W., and J. M. Akey. 2013. Selection and adaptation in the human genome. *Ann. Rev. Genomics Hum. Genet.* 14: 467—489. [8, 9]

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. [8]

Garud, N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov. 2014. Recent selective sweeps in *Drosophila* were abundant and primarily soft. *ArXiv* 1303.0906. FLAG UPDATE REF [8, 9]

Gillespie, J. H. 1997. Junk ain't what junk does: neutral alleles in a selected context. *Gene* 205: 291–299. [8]

Gillespie, J. H. 2000. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* 155: 909–919. [8]

Golding, G. B. 1997. The effect of purifying selection on genealogies. *In* P. Donnelly and S. Tavare (eds), *Progress in population genetics and human evolution*, pp. 271–285. Springer-Verlag. New York. [8]

González, J., K. Lenkov, J. M. Macpherson, and D. A. Petrov. 2008. High rate of recent transposable element-induced adaptation in *Drosophila melanogaster*. *PLoS Bio.* 6: e251. [8]

Gordo, I., A. Navarro, and B. Charlesworth. 2002. Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* 161: 835–848. [8]

Guttman, D. S., and D. E. Dykhuizen. 1994. Detecting selective sweeps in naturally occurring *Escherichia coli. Genetics* 138: 993–1003. [8]

Haddrill, P. R., D. L. Halligan, D. Tomaras, and B. Charlesworth. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biology* 8: R18. [8]

Hahn, M. W. 2008. Toward a selection theory of molecular evolution. *Evolution* 62: 255–265. [8]

Halligan, D. L., and P. D. Keightley. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* 16: 875–884. [8]

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. [8]

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. [8]

Hellmann, I., Y. Mang, Z. Gu, P. Li, F. M. de la Vega, A. G. Clark, and R. Nielsen. 2008. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res.* 18: 1020–1029. [8]

Hermisson, J., and P. S. Pennings. 2005. Soft sweeps: population genetics of adaptation from standing genetic variation. *Genetics* 169: 2335–2352. [8]

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, G. Sella, and M. Przeworski. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331: 920–924. [8]

Hey, J., and R. M. Kliman. 2002. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics* 160: 595–608. [8]

Hill, W. G., and A. Robertson. 1966. The effects of linkage on limits to artificial selection. *Genet. Res.* 8: 269–294. [8]

Huang, X., and 34 others. 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490: 497–501. [8, 9]

Hudson, R. R. 1994. How can the low levels of DNA sequence variation in regions of the *Drosophila* genome with low recombination rates be explained? *Proc. Natl. Acad. Sci. USA* 91: 6815–6818. [8]

Hudson, R. R., and N. L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111: 147–164. [8]

Hudson, R. R., and N. L. Kaplan. 1988. The coalescent process in models with selection and recombination. *Genetics* 120: 831–840. [8]

Hudson, R. R., and N. L. Kaplan. 1995. Deleterious background selection with recombination. *Genetics* 141: 1605–1617. [8]

Ingvarsson, P. K. 2010. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula. Mol. Biol. Evol.* 27: 650–660. [8]

Innan, H., and Y. Kim. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proc. Natl. Acad. Sci. USA* 101: 10667–10672. [8]

Innan, H., and W. Stephan. 2003. Distinguishing the hitchhiking and background selection models. *Genetics* 165: 2307–2312. [8]

Innan, H., and F. Tajima. 1997. The amounts of nucleotide variation within and between allelic classes and the reconstruction of the common ancestral sequence in a population. *Genetics* 147: 1431–1444. [8]

Jaenicke-Després, V., E. S. Buckler, B. D. Smith, M. T. P. Gilbert, A. Cooper, J. Doebley, and S. Pääbo. 2003. Early allelic selection in maize as revealed by ancient DNA. *Science* 302: 1206-1208. [8]

Jensen, J. D., and D. Bachtrog. 2010. Characterizing recurrent positive selection at fast-evolving genes in *Drosophila miranda* and *Drosophila pseudoobscura*. *Genome Biol. Evol.* 2: 371–378. [8]

Jensen, J. D., K. R. Thornton, and P. Andolfatto. 2008. An approximate bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genetics* 4: e1000198. [8]

Jensen, J. D., K. R. Thornton, C. D. Bustamante, and C. F. Aquadro. 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics* 176: 2371–2379. [8]

Jensen, M. A., B. Charlesworth, and M. Kreitman. 2002. Patterns of genetic variation at a chromosome 4 Locus of *Drosophila melanogaster* and *D. simulans. Genetics* 160: 493–507. [8]

Kaiser, V. B., and B. Charlesworth. 2009. The effects of deleterious mutations on evolution in non-recombining genomes. *Trends Genet.* 25: 9–12. [8]

Kaplan, N. L., T. Darden, and R. R. Hudson. 1988. The coalescent process in models with selection. *Genetics* 120: 819–829. [8]

Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The "hitchhiking effect" revisited. *Genetics* 123: 887–899. [8]

Karasov, T., P. W. Messer, and D. A. Petrov. 2010. Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLoS Gene.* 6: e1000924. [8]

Kemper, K. E., S. J. Saxton, S. Bolormaa, B. J. Hayes, and M.E. Goddard. 2014. Selection for complex traits leaves little or no classic signatures of selection. *BMC Genom.* 15: 246. [8, 9]

Kim, Y. 2004. Effects of strong directional selection on weakly selected mutations at linked sites: implications for synonymous codon usage. *Mol. Biol. Evol.* 21: 286–294. [8]

Kim, Y. 2006. Allele frequency distribution under recurrent selective sweeps. *Genetics* 172: 1967–1978. [8]

Kim, Y., and R. Nielsen. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167: 1513–1524. [8]

Kim, Y., and W. Stephan. 2000. Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* 155: 1415–1427. [8]

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160: 765–777. [8]

Kim, Y., and W. Stephan. 2003. Selective sweeps in the presence of interference among partially linked loci. *Genetics* 164: 389–398. [8]

Kimura, M. 1968. Evolutionary Rate at the Molecular Level. *Nature* 217: 624–626. [8]

Kimura, M. 1983. *The neutral theory of molecular evolution*, Cambridge Univ. Press, Cambridge. [8]

Kimura, M., and T. Maruyama. 1966. The mutational load with epistatic gene interactions in fitness. *Genetics* 54: 1337–1351. [8]

King, J. L., and T. H. Jukes. 1969. Non-Darwinian evolution. *Science* 164: 788—798. [8]

Kliman, R. M., and J. Hey. 1993. Reduced natural selection associated with low recombination in *Drosophila melanogaster. Mol. Biol. Evol.* 10: 1239–1258. [8]

Koch, A. L. 1974. The pertinence of the periodic selection phenomenon to prokaryote evolution. *Genetics* 77: 127–142. [8]

Kojima, K., and H. E. Schaffer. 1967. Survival process of linked mutant genes. *Evolution* 21: 518–531. [8]

Kreitman, M., and R. R. Hudson. 1991. Inferring the evolutionary histories of the *Adh* and *Adh-dup* Loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics* 127: 565–582. [8]

Kurtz, T. G. 1971. Limit theorems for sequence of jump Markov processes approximating ordinary differential equations. *J. Appl. Prob.* 8: 344–356. [8]

Lande, R. 1983. The response to selection on major and minor mutations affecting a metrical trait. *Heredity* 50: 47–65. [8]

Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen, and J. M. Braverman. 2000. Linkage disequilibria and the site frequency spectra in the *su(s)* and *su*(W^a) regions of *Drosophila melanogaster* X chromosome. *Genetics* 156: 1837–1852. [8]

Lewontin, R. C. 1974. *The genetic basis of evolutionary change.* Columbia University Press, New York. [8]

Li, H., and W. Stephan. 2005. Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics* 171: 377–384. [8]

Li, H., and W. Stephan. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila. PLoS Genetics* 2: e166. [8]

Li, W.-H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* 24: 337–345. [8]

Loewe, L., and B. Charlesworth. 2007. Background selection in single genes may explain patterns of codon bias. *Genetics* 175: 1381–1393. [8]

Lohmueller, K. E., A. Albrechtsen, Y.Li, S. Y. Kim, T.Korneliussen, N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, A. F. Feder, N. Grarup, R. Jørgensen, T. Jiang, D. R. Witte, A. Sandbaek, I. Hellmann, T. Laurtitzen, T. Hansen, O. Pederson, J. Wang, and R. Nielsen. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet.* 7: e1002326. [8]

Lu, J., T. Tang, H. Tang, J. Huang, S. Shi, and C.-I. Wu. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet.* 22: 126–131. [8]

Macpherson, J. M., G. Sella, J. C. Davis, and D. A. Petrov. 2007. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. *Genetics* 177: 2083–2099. [8]

Marais, G., D. Mouchiround, and L. Duret. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc. Natl. Acad. Sci. USA* 98: 5688–5692. [8]

Maside, X., A. W. Lee, and B. Charlesworth. 2004. Selection on codon usage in *Drosophila americana. Curr. Biol.* 14: 150–154. [8]

Maynard-Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favorable gene. *Genet. Res.* 23: 23–35. [8]

McVean, G. 2007. The structure of linkage disequilibrium around a selective sweep. *Genetics* 175: 1395–1406. [8]

McVean, G. A. T., and B. Charlesworth. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genet. Res.* 74: 145–158. [8]

McVean, G. A. T., and B. Charlesworth. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* 155: 929–944. [8]

McVicker, G., D. Gordon, C. Davis, and P. Green. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics* 5: e1000471. [8]

Messer, P. W., and R. A. Neher. 2012. Estimating the strength of selective sweeps from deep population diversity data. *Genetics* 181: 593–605. [8]

Messer, P. W., and D. A. Petrov. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends Genet.* 28: 659–669. [8]

Muller, H. J. 1964. The relation of recombination to mutational advance. *Mutat. Res.* 106: 2-–9. [8]

Neher, R. A., and B. I. Shariman. 2011. Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics* 188: 975–996. [8]

Nei, M., Y. Suzuki, and M. Nozawa. 2010. The neutral theory of molecular evolution in the genomic era. *Ann. Rev. Genom. Human Genet.* 11: 265–289. [8]

Neuhauser, C., and S. M. Krone. 1997. The genealogy of samples in models with selection. *Genetics* 145: 519–534. [8]

Norman, M. F. 1974. A central limit theorem for Markov processes that move by small steps. *Ann. Prob.* 2: 1065–1074. [8]

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246: 96–98. [8]

Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Ann. Rev. Ecol. Syst.* 23: 263–286. [8]

Ohta, T. 2002. Near-neutrality in evolution for genes and gene regulation. *Proc. Natl. Acad. Sci. USA* 99: 16134–16137. [8]

O'Rourke, B. A., P. L. Greenwood, P. F. Arthur, and M. E. Goddard. 2012. Inferring the recent ancestry of *myostatin* alleles affecting muscle mass in cattle. *Anim. Genet.* 44: 86–90. [8]

Orr, H. A. and A. J. Betancourt. 2001. Haldane's sieve and adaptation from the standing genetic variation. *Genetics* 157: 875–884. [8]

Otto, S. P., and N. H. Barton. 1997. The evolution of recombination: Removing the limits to natural selection. *Genetics* 147: 879–906. [8]

Pavlidis, P., D. Metzler, and W. Stephan. 2012. Selective sweeps in multilocus models of quantitative traits. *Genet.* 192: 225—239. [8]

Pennings, P. S., and J. Hermisson. 2006a. Soft sweeps II – Molecular population genetics of adaptation from recurrent mutation or migration. *Mol. BIol. Evol.* 23: 1076–1084. [8]

Pennings, P. S., and J. Hermisson. 2006b. Soft sweeps III – The signature of positive selection from recurrent mutation. *PLoS Genetics* 2: e186. [8]

Perlitz, M., and W. Stephan. 1997. The mean and variance of the number of segregating sites since the last hitchhiking event. *J. Math. Biol.* 36: 1–23. [8]

Peter, B. M., E. Huerta-Sanchez, and R. Nielsen. 2012. Distinguishing between selective sweeps from standing variation and from a *de novo* mutation. *PLoS Genet.* 8: e1003011. [8]

Pfaffelhuber, P., B. Haubold, and A. Wakolbinger. 2006. Approximate genealogies under genetic hitchhiking. *Genetics* 174: 1995–2008. [8]

Pfaffelhuber, P., A. Lehert, and W. Stephan. 2008. Linkage disequilibrium under genetic hitchhiking in finite populations. *Genetics* 179: 527–537. [8]

Pfaffelhuber, P., and A. Studeny. 2007. Approximating genealogies for partially linked neutral loci under a selective sweep. *J. Math. Biol.* 55: 299–330. [8]

Powell, J. R., and E. M. Moriyama. 1997. Evolution of codon usage bias in *Drosophila. Proc. Natl. Acad. Sci. USA* 94: 7784–7790. [8]

Pritchard, J. K., and A. Di Rienzo. 2010. Adaptation — not by sweeps alone. *Nature Rev. Genet.* 11: 665–667. [7]

Pritchard, J. K., J. K. Prickrell, and G. Coop. 2010. The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Cureent Biol.* 20: R208–R215. [7]

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* 160: 1179–1189. [8]

Przeworski, M. 2003. Estimating the time since the fixation of a beneficial allele. *Genetics* 164: 1667–1676. [8]

Przeworski, M., B. Charlesworth, and J. D. Wall. 1999. Genealogies and weak purifying selection. *Mol. Biol. Evol.* 16: 246–252. [8]

Przeworski, M., G. Coop, and J. D. Wall. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59: 2312–2323. [8]

Qin, H., W. B. Wu, J. M. Comeron, M. Kreitman, and W.-H. Li. 2004. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics* 168: 2245–2260. [8]

Ralph, P., and G. Coop. 2010. Parallel adaption: One or many waves of advance of an advantegous allele? *Genetics* 186: 647–668. [8]

Rockman, M. V. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* 66: 1–17. [8]

Santagio, E., and A. Caballero. 2005. Variation after a selective sweep in a subdivided population. *Genetics* 169: 475–483. [8]

Sattah, S., E. Elyashiv, O. Kolodny, Y. Rinott, and G. Sella. 2011. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. *PLoS Gen.* 7: e1001302. [8]

Schneider, A., B. Charlesworth, A. Eyre-Walker, and P. D. Keightley, P. D. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics* 189: 1427–1437. [8, 10]

Schiffels, S., G. J. Szöllősi, V. Mustonen, and M. Lässig. 2011. Emergent neutrality in adapative asexual evolution. *Genetics* 189: 1361–1375. [8]

Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto. 2009. Pervasive natural selection in the *Drosophila genome*? *PLoS Genetics* 5: e100049. [8]

Slatkin, N. 1995b. Hitchhiking and associative overdominance at a microsatellite locus. *Mol. Biol. Evol.* 12: 473–480. [8]

Slatkin, M., and T. Wiehe. 1998. Genetic hitch-hiking in a subdivided population. *Gene. Res.* 71: 155–160. [8]

Smith, N. G. C., and A. Eyre-Walker. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415: 1022–1024. [8]

Stephan, W. 1995. An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. *Mol. Biol. Evol.* 12: 959–962. [8]

Stephan, W. 2010a. Detecting strong positive selection in the genome. *Mol. Ecol. Res.* 10: 863–872. [8]

Stephan, W. 2010b. Genetic hitchhiking versus background selection: the controversy and its implications. *Phil. Trans. R. Soc.* B 365: 1245–1253. [8]

Stephan, W., Y. S. Song, and C. H. Langley. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* 172: 2647–2663. [8]

Stephan, W., T. H. E. Wiehe, and M. W. Lenz. 1992. The effect of strongly selected substitutions on neutral polymorphisms: analytical results based on diffusion theory. *Theor. Pop. Biol.* 41: 237–254. [8]

Strobeck, C. 1983. Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. *Genetics* 103: 545–555. [8]

Studer, A., Q. Zhao, J. Ross-Ibarra, and J. Doebley. 2011. Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nature Gen.* 43: 1160–1165. [8]

Tachida, H. 2000. Molecular evolution in a multisite nearly neutral mutation model. *J. Mol. Evol.* 50: 69–81. [8]

Tajima, F. 1989. Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595. [8]

Tang, H., G. J. Wyckoff, J. Lu, and C.-I. Wu. 2004. A universal evolutionary index for amino acid changes. *Mol. Biol. Evol.* 21: 1548–1556. [8]

Teshima, K. M., and M. Przeworski. 2006. Directional positive selection on an allele of arbitrary dominance. *Genetics* 172: 713–718. [8]

Thomson, G. 1977. The effect of a selected locus on linked neutral variation. *Genetics* 85: 753–788. [8]

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omra, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas. 2007. Convergence adaptation of human lactase persistance in Africa and Europe. *Nature Gen.* 39: 31–40. [8]

Turner, J. R. G., 1981. Adaptation and evolution in *Heliconius*: a defense of NeoDarwinism. *Annu. Rev. Ecol. Syst.* 12: 99–121. [8]

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. *Nature Gene.* 44 1015–1019. [8]

Wang, R.-L., A. Stec, J. Hey, L. Lukens, and J. Doebley. 1999. The limits of selection during maize domestication. *Nature* 398: 236–239. [8]

Wang, X., W. E. Grus, and J. Zhang. 2006. Gene losses during human origins. *PLoS Biol.* 4: e52. [8]

Weissman, D. B., and N. H. Barton. 2012. Limits to the rate of adaptive substitution in sexual populations. *PLoS Genetics* 8: e1002740. [8]

Wiehe, T. 1998. The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. *Theor. Pop. Biol.* 53: 272–283. [8]

Wiehe, T. H. E., and W. Stephan. 1993. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol. Biol. Evol.* 10: 842–854. [8]

Wiener, P., D. Burton, P. Ajmone-Marsan, S. Dunner, G. Mommens, I. J. Nijman, C. Rodellar, A. Valentini, and J. L. Williams. 2003. Signatures of selection? Patterns of microsatellite diversity on a chromosome containing a selected locus. *Heredity* 90: 350–358. [8]

Xue, Y., A. Daly, B. Yngvadottir, M. Liu, G. Coop, Y. Kim, P. Sabeti, Y. Chen, J. Stalker, E. Huckle, J. Burton, S. Leonard, J. Rogers, and C. Tyler-Smith. 2006. Spread of an inactive form of Caspase-12 in humans due to recent positive selection. *Amer. J. Hum. Genet.* 78: 659–670. [8]

Zeng, K. 2010. A simple multiallele model and its application to identifying preferred–unpreferred codons using polymorphism data. *Mol. Bio. Evol.* 27 1327–1337. [8]

Zeng, K., and B. Charlesworth. 2009. Estimating selection intensity on synonymous codon usage in a nonequilibrium population. *Genetics* 183: 651–662. [8]

Zeng, K., and B. Charlesworth. 2010. Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster. J. Mol. Evol.* 70: 116–128. [8]