Soft sweeps and beyond: Understanding the patterns and probabilities of selection footprints under rapid adaptation

Joachim Hermisson[1*] & Pleuni S Pennings[2*]

[1] Department of Mathematics and Max F. Perutz Laboratories, University of Vienna, joachim.hermisson@univie.ac.at

[2] Department of Biology, San Francisco State University, pennings@stsu.edu

[*] both authors contributed equally

March 6, 2017

Abstract

1. The tempo and mode of adaptive evolution determine how natural selection shapes patterns of genetic diversity in DNA polymorphism data. While slow mutation-limited adaptation leads to classical footprints of "hard" selective sweeps, these patterns are different when adaptation responds quickly to a novel selection pressure, acting either on standing genetic variation or on recurrent new mutation. In the past decade, corresponding footprints of "soft" selective sweeps have been described both in theoretical models and in empirical data.

2. Here, we summarize the key theoretical concepts and contrast model predictions with observed patterns in *Drosophila*, humans, and microbes.

3. Evidence in all cases shows that "soft" patterns of rapid adaptation are frequent. However, theory and data also point to a role of complex adaptive histories in rapid evolution.

4. While existing theory allows for important implications on the tempo and mode of the adaptive process, complex footprints observed in data are, as yet, insufficiently covered by models. They call for in-depth empirical study and further model development.

1 Hard and soft selective sweeps

The view of adaptation in molecular evolution has long focused almost exclusively on a mutation-limited scenario. The assumption in such a scenario is that beneficial mutations are rare, so that they are unlikely to be present in the population as standing genetic variation (SGV) or to occur multiple times in a short window of time. Mutation-limited adaptation therefore occurs from single new beneficial mutations that enter the population only after the onset of the selection pressure (e.g. due to environmental change). When such a beneficial mutation fixes in the population, it reduces the genetic diversity at linked neutral loci according to the classical model of a "hard" selective sweep (Maynard Smith and Haigh, 1974; Kaplan et al., 1989; Barton, 1998). If recurrent beneficial mutation is considered at all in a mutation-limited scenario, each such mutation is assumed to create a new allele. Single adaptive steps then either proceed independently of each other or compete due to clonal interference, where adaptation is slowed by linkage (Gerrish and Lenski, 1998; Desai and Fisher, 2007). Despite of evidence for rapid adaptation from quantitative genetics and phenotypic studies (Messer et al., 2016), SGV or recurrent origins of the same allele have long been ignored in molecular evolution.

Around a decade ago, several publications started to explore selective sweeps outside the mutation-limitated scenario (Innan and Kim, 2004; Hermisson and Pennings, 2005; Przeworski et al., 2005; Pennings and Hermisson, 2006a,b). These papers described novel patterns for genetic footprints termed "soft sweeps". It has since become clear that non-mutation-limited adaptation and soft sweeps are probably much more common than originally thought. However, recent years have also seen some lively debate about whether soft sweeps are everywhere (Messer and Petrov, 2013) or rather a chimera (Jensen, 2014). Some aspects of this dispute root in diverging interpretation of the available data, others go back to conceptual differences. With this review, we aim to provide an overview and intuitive understanding of the relevant theory and the patterns that are observed in model species.

Definitions

Selective sweeps refer to patterns in genomic diversity that are caused by recent adaptation. Characteristic patterns (footprints) arise if rapid changes in the frequency of a beneficial allele, driven by positive selection, distort the genealogical history of samples from the region around the selected locus. We can therefore understand sweep footprints via properties of their underlying genealogy or *coalescent* history (Wakeley, 2008). In this context, the key genealogical implication of mutation-limited adaptation is that, at the selected locus itself, the time to the most recent common ancestor (MRCA) of the sample, T_{MRCA}, is shorter than the time that has elapsed since the onset of the new selection pressure, T_S. For recent adaptation, we thus obtain coalescent histories that are much shorter than the expected neutral coalescent time T_N ($2N_e$ generations for a pair of lineages in a diploid population of effective size N_e). It is this shortened genealogy that is responsible for the characteristic footprint of a hard sweep (see the Footprints section below).

- We define a hard sweep based on the sample genealogy of the beneficial allele, requiring that (i) $T_{\mathrm{MRCA}} \ll T_N$ (recent adaptation), *and* (ii) $T_{\mathrm{MRCA}} \leq T_S$ (a single recent ancestor), see Fig. 1A.

If adaptation is not mutation limited because the beneficial allele was already present in the population prior to the onset of selection, or because the mutation occurs recurrently, both assumptions of a hard sweep can be violated: The time to the MRCA typically predates the onset of the selection pressure, $T_{\mathrm{MRCA}} > T_S$, and can approach the neutral coalescent time, $T_{\mathrm{MRCA}} \sim T_N$. Still, if adaptation is recent, we can observe a characteristic footprint: a soft selective sweep.

- For a soft sweep, we require that (i) $T_S \ll T_N$ (recent adaptation), and (ii) $T_{\mathrm{MRCA}} > T_S$ (more than a single recent ancestor) for the sample genealogy of the beneficial allele.

There are two different ways how soft sweep genealogies can come about, which lead to different patterns (Hermisson and Pennings, 2005).

1. *Single-origin soft sweeps* refer to a genealogy that traces back to a single mutational origin of the beneficial allele, but comprises multiple copies of this allele in the SGV when positive selection sets in at time T_S, see Fig. 1B.

2. *Multiple-origin soft sweep*: In this case, the sample genealogy of the beneficial allele comprises multiple origins of this same allele from recurrent mutation (or from some other source like migration). These origins can occur prior to the onset of selection (standing variation), but also only after that (new variation), Fig. 1C.

For both, hard and soft sweeps, the definitions apply irrespective of whether the beneficial allele has reached fixation (complete sweep) or still segregates in the population (partial sweep), as long as we restrict the sample to carriers of the beneficial allele only. Note that, for a given sweep locus, we may observe a soft sweep in one sample, but a hard sweep in a different sample. The probability to observe a soft sweep increases with sample size, up to the whole population. We discuss below under which conditions sample size has (or has not) a major influence on this probability.

Two non-exclusive processes can lead to soft sweeps, which both capture essential aspects of non-mutation-limited adaptation: adaptation from SGV and adaptation from recurrent mutation. However, there is

no one-to-one correspondence of a process and the type of selective sweep that results, see Fig. (2). In particular, adaptation from SGV can either lead to a hard sweep (if a single copy from the standing variation is the ancestor of all beneficial alleles in the sample), or to a soft sweep, which can either be *single-origin* or *multiple-origin*. The notion of a "soft sweep" as defined here following previous work (e.g. Hermisson and Pennings, 2005; Messer and Petrov, 2013; Jensen, 2014; Berg and Coop, 2015) is therefore not synonymous for "adaptation footprint from SGV". Sweep types refer to classes of patterns that result from characteristic coalescent genealogies, not to evolutionary processes. This leaves us the task to explore the *probability* of each sweep type under a given evolutionary scenario and to use this information for statistical inference of process from pattern.

2 Footprints of hard and soft sweeps

In order to distinguish footprints of hard and soft sweeps, we need to understand how these footprints are shaped by the different characteristic features of the underlying genealogies. We assume the following model. A diploid population of constant size N_e experiences a new selection pressure at time T_S. The derived allele A is generated from the wildtype a with fitness 1 by recurrent mutation of rate u. The frequency of A is $x = x(t)$. Prior to T_S, A is neutral or deleterious with fitness disadvantage $1 - s_d$, $s_d \geq 0$. After time T_S, allele A is beneficial with fitness $1 + s_b$. For simplicity, we assume codominance.

Footprints of hard sweeps

The hallmark of a hard sweep genealogy is a very recent common ancestor of all beneficial alleles in the sample. Among carriers of the beneficial allele, ancestral variation prior to the onset of selection can only be preserved if there is recombination between the polymorphic locus and the selection target. Between the closest recombination breakpoints to the left and to the right of the selected allele, we obtain a core region without ancestral variation. In the flanking regions, ancestral variation is main-

tained on some haplotypes due to recombination. These *recombination haplotypes* generate characteristic signals in the site-frequency spectrum.

We can understand the characteristics of hard sweep footprints from the shape of typical genealogies as sketched in Figure 1A: The genealogy directly at the selected site is "star-like", with all coalescence events happening in a short time interval when the frequency x of the beneficial allele is very small. This is because all ancestors (between now and the MRCA) in a hard sweep genealogy must carry the beneficial allele. The number of potential ancestors at any given point in time is $2N_e x$. For small x, the coalescence probability $1/(2N_e x)$ thus becomes very large (Barton, 1998).

The average fixation time of a beneficial allele with selection coefficient s_b is $\approx 2\log[4N_e s_b]/s_b$ generations (van Herwaarden and van der Wal, 2002; Hermisson and Pennings, 2005). All recombination events that can restore ancestral variation in a sample need to occur during this phase. For strong selection, this time is short so that we obtain a broad core region without any variation – other than new mutations that occur during or after the sweep. Its width is inversely proportional to the recombination rate r times the fixation time and thus roughly $\sim s_b/r$ (Kaplan et al., 1989). Due to the star-like genealogy, recombination during the selective phase will typically only affect single lineages, which (back in time) have not yet coalesced with other lineages (Fig. 1A). Recombination therefore introduces single long branches into the genealogy of a linked neutral locus, which reach far back into the time prior to the selective phase. Neutral mutations that occur on these branches either lead to low-frequency variants or high-frequency derived variants, depending on whether they occur on the internal or external long branch (Fig. 1A). This explains the typical site-frequency spectrum in the flanking regions of hard sweeps with an excess of high- and low-frequency alleles (Braverman et al., 1995; Fay and Wu, 2000). The recombination haplotypes create positive linkage disequilibrium (LD) in the flanking regions of a hard sweep. However, since separate recombination events are needed to both sides, there is (on average) no LD across the selected site (Kim and Nielsen, 2004; Stephan et al., 2006).

Footprints of single-origin soft sweeps

For single-origin soft sweeps, like for hard sweeps, all beneficial alleles in the sample trace back to a single origin. Coalescent histories at the selected site are therefore confined to the descendants of the founding mutation (the colored region in Fig. 1B). In contrast to hard sweeps, however, the MRCA predates the onset of selection, $T_{\mathrm{MRCA}} > T_S$ (measured from "Now"). This is possible if adaptation occurs from SGV. If the allele is neutral or only slightly deleterious as it arises, it can stay in the population for a long time, its frequency governed by drift. Only when the allele becomes beneficial, it quickly rises in response to selection.

The pattern of a single-origin soft sweep depends on the age of the MRCA relative to the onset of selection. If T_{MRCA} is not much older than T_S, the pattern will look like a hard sweep. This is typically the case if the selected allele is strongly deleterious in the ancestral environment, because deleterious alleles do not stay in the population for a long time. On the other hand, if T_{MRCA} *is* much older than T_S, a distinct footprint of a single-origin soft sweep will be visible. If the fitness of the selected allele is $1-s_d$ before T_S and $1+s_b$ after T_S, this will be the case for alleles with a weak fitness trade-off, $s_d \ll s_b$. Hermisson and Pennings (2005) define a parameter of relative selective advantage,

$$R = \frac{s_b}{s_d + 1/4N_e} \tag{1}$$

and estimate that a clear pattern of a single-origin soft sweep (relative to a hard sweep with strength s_b throughout) can be expected for $R \gtrsim 100$.

Alleles with a weak trade-off (large R) have a larger chance to be picked up from SGV and then will typically lead to soft sweeps (see the Probabilities section below). Patterns have been described in detail for adaptation from neutral variation ($s_d = 0$) by Przeworski et al. (2005), Peter et al. (2012) and Berg and Coop (2015). The essential difference relative to hard sweeps is that genealogies are no longer star-like because the frequency x of the (later) beneficial allele changes only slowly during the "standing phase" prior to T_S. The order of coalescence, recombination, and neutral mutation events during this phase is unaffected by x. As a consequence,

early mutation and recombination events often affect multiple individuals and lead to intermediate-frequency polymorphism in a sample (Fig. 1B). Berg and Coop (2015) show that the number and frequencies of these early recombination haplotypes follow the Ewens sampling distribution (Ewens, 1972). The net effect is a weakening of the sweep signal, especially close to the selected site (where the hard sweep signal is strongest): the core region is narrower and nearby flanking regions are not strongly dominated by low-frequency variants.

Footprints of multiple-origin soft sweeps

A soft sweep genealogy with multiple origins of the beneficial allele extends into the part of the population that carries the ancestral allele (Fig. 1C), where coalescence happens on a neutral time scale. This leads to the key characteristic of this sweep type: a pattern of extended haplotypes (one per origin) that stretch across the selected site. In contrast to the recombination haplotypes of a hard sweep, haplotypes corresponding to different mutational origins are typically observed at intermediate frequencies in population samples. Like the early recombination haplotypes of a single-origin soft sweep, they follow a Ewens distribution (Pennings and Hermisson, 2006a).

The difference in expected frequencies of recombination haplotypes of hard sweeps and mutation haplotypes of multiple-origin soft sweeps can be understood as follows. Imagine we follow a lineage back in time. In each generation, it "picks" an ancestor from the $2N_e x$ beneficial alleles in that generation. The probability that this ancestor is a new beneficial mutant is $2N_e u(1-x)/2N_e x \approx \Theta/(4N_e x)$ for small x. This is proportional to the coalescence probability ($\sim 1/(2N_e x)$) in the same generation. Both coalescence and beneficial mutation typically occur very early in the selected phase when x is small. However, since their *relative* frequency is independent of x, their order along the genealogy is not affected. Consequently, mutation events will often happen on internal branches, leading to intermediate frequency haplotypes (Fig. 1C). In contrast, the probability of "picking" a recombinant ancestor, $2N_e x(1-x)r/2N_e x = (1-x)r$,

depends much weaker on x. During a sweep, recombination therefore tends to happen before mutation and coalescence (going back in time). It thus typically affects single external branches, leading to low-frequency polymorphism.

Box: Inference of hard and soft sweeps

Can we identify patterns of soft sweeps? Clearly, only recent adaptive events leave a detectable footprint at all, hard or soft. Signals in the site frequency spectrum (like the excess of rare alleles that is picked up by Tajima's D, Tajima, 1989) typically fade on time scales of $\sim 0.1N_e$ generations, while signals based on LD or haplotype statistics only last for $\sim 0.01N_e$ generations (Przeworski, 2002; Pennings and Hermisson, 2006b). For a clear footprint, selection must be strong ($4N_e s_b \gg 100$). Even then, soft sweeps can be hard to distinguish from neutrality if they are "super soft", i.e. if there are very many independent origins of the beneficial allele, or if its starting frequency in the SGV is high ($\gtrsim 20\%$, Peter et al., 2012; Berg and Coop, 2015).

For robust inference of selection against neutrality, we need a test statistic with consistently high power for hard and soft sweeps. As expected from the patterns described above, and as demonstrated (Pennings and Hermisson, 2006b; Ferrer-Admetlla et al., 2014), tests based on the site-frequency spectrum (looking for low- or high-frequency derived alleles) have low power to detect soft sweeps, while haplotype tests can detect both types of sweeps (Garud et al., 2015). In contrast to single-origin soft sweeps (which always leave a weaker footprint), the power to detect multiple-origin soft sweeps can be higher than the power to detect completed hard sweeps due to the conspicuous haplotype structure directly at the selected site (Pennings and Hermisson, 2006b).

Given that selection has been inferred, can we distinguish soft from hard sweeps? For soft sweeps with a single origin, this is difficult (Berg and Coop, 2015). Tests based on a combination of summary statistics have been developed by Peter et al. (2012) and by Schrider and Kern (2016a). Both tests have reasonable power to distinguish soft sweeps for strong

selection and a high starting frequency ($5 - 20\%$) of the selected allele. Clear empirical examples usually also rely on other evidence, complementing the footprint (Barrett and Schluter, 2008): e.g., a source population is known with the selected allele in the SGV (e.g. marine and freshwater sticklebacks, Colosimo et al., 2005), or a known and very recent selection pressure does not leave enough time for the allele to increase from a single copy to the frequency observed today (e.g. adaptation to HIV in humans, Novembre and Han, 2012).

Chances for a distinction of sweep types are better for soft sweeps with multiple origins. A tailor-made summary statistic, $H12$, based on the two largest haplotype classes has been developed by Garud et al. (2015). It has a high power to distinguish hard and soft sweeps especially for recent or partial sweeps. $H12$ has also been included into the deep learning algorithm by Sheehan and Song (2016) that is able to distinguish hard and soft sweeps.

3 Probabilities: When to expect soft sweeps

If adaptation is strictly mutation limited, all selective sweeps are necessarily hard. Without mutation limitation, soft sweeps can originate either from SGV or from recurrent new mutation (Fig. 2). Below, we discuss the probability for soft sweeps in both scenarios.

Sweeps from standing genetic variation

When is adaptation from SGV likely? When does it produce hard or soft sweeps? In an early approach to address these questions, Orr and Betancourt (2001) used the following argument. Let x_0 be the frequency of the A allele in the population at time T_S when A turns beneficial. Assuming independence, each of the $2N_e x_0$ mutant copies has the chance $2s_b$ to escape stochastic loss. We can then approximate the distribution of the number X of *successful* A copies by a Poisson distribution with parameter $2s_b \cdot 2N_e x_0$. The probability for a *standing sweep* (at least one

successful mutant) is

$$P_{\text{sgv}}[x_0] = 1 - P[X = 0] = 1 - \mathrm{e}^{-4N_e x_0 s_b}\,. \tag{2}$$

The conditional probability of a soft sweep (at least two successful copies), given that there is a standing sweep at all, follows as

$$P_{\text{mult}}[x_0] = \frac{1 - P[X = 0] - P[X = 1]}{1 - P[X = 0]} = \frac{1 - (1 + 4N_e x_0 s_b)\mathrm{e}^{-4N_e x_0 s_b}}{1 - \mathrm{e}^{-4N_e x_0 s_b}}\,. \tag{3}$$

Clearly, x_0 is a crucial parameter in these equations, but which value does it take? Orr and Betancourt (2001) and Jensen (2014) assume that the population is in mutation-selection balance prior to T_S and use the deterministic approximation $x_0 = u/s_d$ for the frequency of A. With $\Theta = 4N_e u$, we obtain $4N_e x_0 s_b = \Theta s_b/s_d$, which implies that hard sweeps (from SGV) dominate over soft sweeps (from SGV) as long as $\Theta s_b/s_d \leq 1.14$ (Orr and Betancourt, 2001). This implies, in particular, that hard sweeps from SGV always dominate for small Θ. However, as shown in Fig. 3A, this prediction is not correct.

The problem of the argument is that x_0, the frequency of mutant allele A at time T_S, is not a fixed value. Really, it is a stochastic variable and follows a distribution. As long as $\Theta < 1$, this distribution is L-shaped with a maximum at 0. This means that in many cases, the A allele will not be present in the population, but in some cases, it will be present at a frequency that is much higher than u/s_d (see figure 3B). Obviously, if the adaptive allele is not present in a population, a standing sweep cannot occur. On the other hand, if many copies of the allele are present, it is likely that a standing sweep will happen and that such sweep will be soft. The result is that, when the L-shaped distribution of x_0 is taken into account, we get fewer sweeps in total, but those that do occur are more likely to be soft. We can compare how the two approaches predict the probability of a standing sweep (soft or hard) from mutation-selection balance. Using the deterministic assumption $4N_e x_0 s_b = \Theta s_b/s_d$, we find

$$P_{\text{sgv}}^{\text{det}} = 1 - \exp\Big(-\Theta\frac{s_b}{s_d}\Big). \tag{4}$$

When taking into account the distribution of x_0, we obtain (Hermisson

and Pennings, 2005, Eq. 8)

$$P_{\text{sgv}} \approx 1 - \exp\Big(- \Theta \log[1+R]\Big), \tag{5}$$

where R is the relative selective advantage defined in Eq. (1). Since $\log[1+R] \leq R < s_b/s_d$, we have $P_{\text{sgv}} < P_{\text{sgv}}^{\text{det}}$. Next, we compare the two approaches as they predict the probability of a soft sweep, given that a standing sweep from mutation-selection balance has happened. The deterministic approximation reads

$$P_{\text{mult}}^{\text{det}} = \frac{1-(1+\Theta s_b/s_d)\mathrm{e}^{-\Theta s_b/s_d}}{1-\mathrm{e}^{-\Theta s_b/s_d}} \approx \frac{\Theta s_b}{2s_d}\,. \tag{6}$$

Accounting for the distribution of x_0, we find (Hermisson and Pennings, 2005, Eq. 18)

$$P_{\text{mult}} \approx 1 - \frac{\Theta R/(1+R)}{(1+R)^{\Theta}-1}\,. \tag{7}$$

We can condition on the case of a single mutational origin of the allele by taking the limit $\Theta \to 0$, where (7) simplifies to

$$P_{\text{mult}} \approx 1 - \frac{R}{(1+R)\log[1+R]} \tag{8}$$

For $R \gtrsim 4$, we obtain more (single origin) soft sweeps than hard sweeps in the population, even for low Θ values, for which the deterministic approximation suggests that soft sweeps are exceedingly rare (Fig. 3A). For neutral standing variation ($s_d = 0$), this requires only weak positive selection ($N_e s_b > 1$). For deleterious standing variation ($s_d \gg 1/N_e$), we need $s_b/s_d > 4$. For biologically relevant parameters, this condition can deviate from the deterministic prediction by several orders of magnitude. For example, using the deterministic approximation, Jensen (2014) argues that soft sweeps are only likely if $s_b/s_d > 100$ when $\Theta = 10^{-2}$ ("*Drosophila*") and if $s_b/s_d > 10000$ when $\Theta = 10^{-4}$ ("humans") (Fig. 1 in Jensen, 2014). Our Fig. 3A shows that even for a low ratio $s_b/s_d = 10$, soft sweeps dominate for all Θ values.

Sweeps from independent mutational origins

When will the genealogy of a beneficial allele contain multiple independent origins? A rough argument (Hermisson and Pennings, 2005; Messer and

Petrov, 2013) shows that the probability of a multiple-origins soft sweep mainly depends on the mutation parameter Θ: A new beneficial allele establishes in the population within $\sim \log[4N_e s_b]/s_b$ generations. During this time, further copies of the allele arise at rate $\sim 2N_e u$ and establish with probability $2s_b$. Establishment of a second beneficial mutation during the sweep thus becomes likely if $2N_e u 2s_b \log[4N_e s_b]/s_b = \Theta \log[4N_e s_b] \gtrapprox 1$. The basic message of this argument is correct. Detailed derivations based on coalescent theory (Pennings and Hermisson, 2006a) show an even weaker dependence on s_b. For strong selection, $2N_e s_b \gg 1$, the probability for more than a single origin of the beneficial allele in a sample of size n derives to

$$P_{\text{ind}}(n) \approx 1 - \prod_{i=1}^{n-1} \frac{i}{i+\Theta} \tag{9}$$

which simplifies to $P_{\text{ind}}(n) \approx \big(0.577 + \log[n-1]\big)\,\Theta$ for small $\Theta \ll 1$ and to $P_{\text{ind}}(n) \approx 1 - (1/n)$ for $\Theta = 1$.

Since Eq. (9) for P_{ind} depends only on Θ, but not on any selection parameter, it also applies if the selection pressure changes during the course of adaptation. In particular, the same result holds for adaptation from SGV (where selection changes from negative or neutral to positive) and for adaptation from recurrent new mutation after time T_S only. Indeed, if Θ is large enough that adaptation acts on multiple origins of the same allele, the distinction of "standing variation" versus "new mutation" gets blurred – adaptive material is immediately available in both cases (Messer and Petrov, 2013). While the number of origins of a beneficial allele and their distribution in the sample are almost entirely independent of selection, other aspects of the sweep signature depend on selection strength. In particular, the signature is wide if and only if s_b and s_d are strong.

$P_{\text{ind}}(n)$ also depends only weakly on sample size n. This is because beneficial alleles from different origins are typically at intermediate frequencies in the population. Indeed, a multiple-origin soft sweep in the entire (panmictic) population will usually be visible already in a sample of moderate size. For $n = 100$, multiple-origin soft sweeps start to appear for $\Theta > 0.01$ ($> 5\%$ soft), become frequent for $\Theta > 0.1$ ($\approx 40\%$ soft) and dominate for $\Theta \geq 1$ ($\geq 99\%$ soft).

The probabilities of adaptation from SGV (5) and from multiple mutational origins (9) are both governed by the population mutation parameter $\Theta = 4N_e u$. Before we can assess Θ in natural populations, however, the factors that enter this composite parameter require some elaboration.

Understanding N_e in Θ: the effective population size

We have, so far, not made any distinction between the census size of a population and its so-called effective size. For a better understanding, we need to do this now. The expected number of new copies of a mutant allele A that enters a diploid population each generation is twice its census size times the mutation rate, $2Nu$. For species as abundant as microbes, fruitflies, or humans, almost any adaptive mutation that is possible will appear many times within a short time interval. However, what matters for adaptation are only those mutants that escape genetic drift. The establishment probability of a mutant with selection coefficient s_b is roughly $2s_bN_e/N$. Here, the effective population size N_e is an (inverse) measure of the strength of genetic drift: strong drift (small N_e) leads to a higher chance of loss due to random fluctuations. We then see that the rate of *successful* mutants, $2Nu \cdot 2s_bN_e/N = 4N_e u s_b$, depends only on N_e, while the census size drops out.

How should we determine N_e? Many factors influence the strength of genetic drift in natural populations (Charlesworth, 2009). Some factors, such as variance in offspring number or unequal sex-ratios, act on the population in every single generation; others recur every few generations, like seasonal fluctuations of the census size. If all relevant factors operate on shorter time scales than typical coalescence times, they can be subsumed in a single well-defined *coalescent effective population size* N_e (Sjödin et al., 2005). Since, by definition, the coalescent effective size is constant over time, it can readily be estimated from neutral polymorphism or diversity data using measures like Watterson's Θ_W.

Unfortunately, several factors that drive drift do not fit this scheme. These include fluctuations in population size over time scales comparable with coalescence times and non-recurrent/sporadic events, such as ongoing

411 population growth, single population bottlenecks, or episodes of linked
412 selection. If any such event occurs during the coalescent history of an
413 allele, it exerts a drift effect on its frequency distribution, but there is "no
414 meaningful effective population size" (Sjödin et al., 2005) to fully describe
415 this effect. Therefore, the question arises which value for N_e should be
416 used in the parameter $\Theta = 4N_e u$ that enters our formulas.

417 Even if it does not take a unique value, N_e is always a measure of
418 drift. In each particular case, we therefore should ask which aspect of
419 drift is relevant. For example, equilibrium levels of SGV are affected
420 by drift throughout the coalescent history of potential "standing" alleles.
421 If adaptation occurs from neutral variation, this history is long and a
422 long-term measure of genetic drift, such as neutral polymorphism Θ_W or
423 pairwise nucleotide diversity π can serve as a valid proxy for Θ in formulas
424 like Eq. (5).

425 For the probability of adaptation from multiple independent origins
426 (9), the decisive time period where drift is relevant is only the estab-
427 lishment phase of the beneficial allele (Pennings and Hermisson, 2006a).
428 This is a time window of length $\sim 1/s_b$, where a successful allele quickly
429 spreads. For very strong selection, this period can be as short as $10 - 100$
430 generations. We thus need an estimate for a "fixation effective size" (Otto
431 and Whitlock, 1997) or "short-term N_e" (Karasov et al., 2010; Barton,
432 2010), which can differ from the "long-term N_e": While factors like the
433 offspring variance enter also into a short-term N_e, any event that occurs
434 outside of the crucial time window has no effect. Unfortunately, there is no
435 easy direct measure of the short-term N_e from polymorphism data. In par-
436 ticular, if demographic factors are at play, using neutral diversity can lead
437 to great underestimates of N_e. To account for such major demographic
438 trends, one can use inference methods like PSMC (Li and Durbin, 2011)
439 or deep sequencing (e.g. Chen et al., 2015) to estimate a time-dependent
440 effective size $N_e(T_S)$ at the (putative) time of the adaptation. However,
441 due to limited resolution, these methods will never capture all confound-
442 ing demographic factors. As pointed out by Karasov et al. (2010), the
443 same holds true for effects of linked selection (recurrent sweeps). Since

unresolved factors usually (though not always) increase drift, estimates of $N_e(T_S)$ generally still underestimate the "true" short-term N_e.

Values of short-term N_e vary not only between populations, but also across adaptive loci along the genome. They depend on the exact timing of the selection windows relative to demographic events or episodes of linked selection. Genome-wide studies, e.g. in *D. melanogaster*, show high heterogeneity in diversity levels due to linked selection (Elyashiv et al., 2016). Finally, since stronger selection leads to shorter windows, the relevant short-term N_e may also depend on the selection strength (Otto and Whitlock, 1997). As Wilson et al. (2014) point out, this may lead to larger N_e (thus larger Θ and more soft sweeps) for strong adaptations with shorter establishment times.

Understanding u in Θ: the allelic mutation rate

For any adaptive mutant allele A, e.g. a resistant phenotype, we can ask how this allele can be produced from a wildtype by mutation of the underlying molecular genotype. Sometimes, A is highly specific and can only be generated by a unique (point-)mutation. However, often multiple mutations produce the same phenotype (e.g. Barroso-Batista et al., 2014, for *E. coli* adaptation in the mouse gut). This is a generic property of the genotype-to-phenotype map, which often maps whole classes of equivalent genotypes to the phenotype that is seen by selection. Redundancies already exist on the level of the genetic code, but also on any other level, both in the coding sequences or regulatory regions of single genes, and across genes and pathways. In this case, A has an extended *mutational target* (Pritchard et al., 2010), which is reflected by an increased allelic mutation rate u. For strict redundancy, we require that all mutations in the target of A have the same fitness effect. We further require that multiple redundant mutations do not increase fitness any further.

We need to distinguish mutational targets on two different levels. On the level of a single locus, u_l and $\Theta_l = 4N_e u_l$ measure the total rate of redundant mutations to produce allele A in a single recombinational unit (technically, we need that recombination during the selective phase

is unlikely). Mutations contributing to Θ_l interfere both due to complete linkage (as in clonal interference) and due to epistasis. Without epistasis, the dynamics of adaptation under recurrent mutation is driven by *mutation stacking* and is dominated by the haplotype most loaded with beneficial mutations (Desai and Fisher, 2007). Negative epistasis among redundant mutations erases the fitness advantage of stacking and fundamentally changes the dynamics of adaptation, leading to soft sweeps instead of clonal interference. (We note that our definition of a mutational target deviates from the one by Messer and Petrov (2013), who also include adaptations to unrelated selection pressures.)

On the genome level, u_g and $\Theta_g = 4N_e u_g$ cover all redundant mutations that produce an A phenotype across all loci. Thus, $\Theta_g = m\Theta_l$ for m equivalent loci. Mutations for A on different genes may be unlinked, but still interfere due to negative epistasis. Without epistasis, the dynamics across different loci decouples, reproducing hard or soft single-locus sweeps. With epistasis, patterns and probabilities are no longer independent and new phenomena can arise. Currently, the corresponding patterns of polygenic adaptation (Pritchard et al., 2010) that can be distributed across many loci remain largely unexplored (see Berg and Coop, 2014, for the case without epistasis). However, simulations show that often a single locus emerges that contributes most to the effect. In our simulations below, we study how a multi-locus target $\Theta_g > \Theta_l$ affects the single-locus probabilities of hard *vs.* soft sweeps at this focal locus.

Θ_l and Θ_g affect the probabilities of hard and soft sweeps in different ways. The probability of a multiple-origin soft sweep, Eq. (9), depends only on the mutation rate Θ_l of the corresponding locus. In contrast, the probability for a single-origin soft sweep *vs.* a hard sweep depends primarily on the genome-wide rate Θ_g. Indeed, the probability of adaptation from SGV depends on whether some of this variation, genome-wide, is picked up by selection. Therefore, Θ_g should be used in Eq. (5). If adaptation happens from SGV, the *conditional* probability for adaptation from multiple SGV copies at a locus, Eq. (7), depends on Θ_l, but becomes independent of Θ_l in the limit of a single mutational origin (Eq.

8). Although hard and soft sweeps are single-locus footprints (defined via a single-locus genealogy), their probabilities thus depend on both Θ_l and Θ_g. Note that a short-term N_e is relevant for recurrent mutation and thus Θ_l, while a long-term N_e is relevant for total levels of variation and thus Θ_g.

Depending on the nature of the adaptation, allelic mutation rates u_l and u_g can vary widely. On the low end of the scale are phenotypes that are only generated by a single base substitution. In this case, $u_g = u_l$ can be as low as $\mu/3$, where μ is the point mutation rate (assuming that all three base substitutions occur at equal rates and ignoring variance of μ along the genome). On the locus level, mutational targets can comprise 10s and maybe 100s of point mutations (and also include insertions/deletions). High values for u_l should be expected especially for adaptive loss-of-function mutants and for some (cis-) regulatory mutations, as in the case of the Lactase gene (see below). For alleles with a polygenic target, u_g could in turn be an order of magnitude larger than the locus rate u_l.

It is worth noting that neither u_l nor u_g are closely related to the so-called distribution of fitness effects (DFE), which, among other things, informs us about the proportion of beneficial mutations among all mutations that hit some target (Jensen, 2014). The DFE divides the total mutation rate into classes of deleterious, neutral, and beneficial and generally finds that beneficial mutations are only a small fraction. In contrast, u_l and u_g ask for the total mutation rate to specific beneficial allele A of which we know that it exists. As such, it is not affected by the presence of further, neutral or deleterious mutations on the same target.

Simulation results

Figure 4 shows percentages of soft and hard sweeps for beneficial alleles with single-locus or multiple-locus mutation targets and with strong or weak fitness trade-off in the ancestral environment. For soft sweeps, we distinguish single- and multiple-origin types. For hard sweeps, we specifically record cases that derive from a single ancestor in the SGV. Assume

first that adaptation is locus-specific and can only occur at a single gene, $\Theta_g = \Theta_l$ (Fig. 4, top row). This may often be the case for resistance mutations.

- For mutations with a strong trade-off, $s_d \geq s_b$, hard sweeps dominate for $\Theta_l < 0.1$, while multiple-origin soft are most prevalent for $\Theta_l > 0.1$. Adaptations from SGV are only likely for $\Theta_l \gg 0.1$ where they produce multiple-origin sweeps. Single-origin soft sweeps or hard sweeps from SGV are very rare already for $s_d = s_b$, $R \approx 1$ (Fig. 4A). For adaptations with even stronger trade-off (e.g. $s_d = 5s_b$, as suggested by Orr and Betancourt, 2001; Jensen, 2014), they are virtually absent.
- For mutations with a weak trade-off, $s_d \ll s_b$, the probability for adaptation from SGV is higher (cf Fig. 4B with $R \approx 100$). We now find a few single-origin soft sweeps for intermediate Θ values, while hard sweeps from SGV remain rare.

The single locus results also describe adaptation of a polygenic trait as long as mutations at different loci do not interact. This is different if the beneficial allele has a mutational target across several loci, $\Theta_g > \Theta_l$, where mutations interact by negative epistasis. Fig. 4, (bottom row) shows how this affects probabilities of soft and hard sweeps, assuming 10 identical loci, $\Theta_g = 10\Theta_l$. Although adaptation can occur, in principle, by small frequency shifts at many loci, we usually observe a *major sweep locus* that experiences a frequency shift of $> 50\%$. For small Θ_g, adaptation is often even entirely due to a single locus. The figure shows the sweep type at a major sweep locus and leaves all areas white where no such locus exists.

- The effect of a multi-locus target is an increase in the probability of adaptation from SGV (Fig. 4C/D). Hard sweeps from new mutations are pushed back to lower values of Θ_l. In a parameter range with intermediate Θ_l between 0.01 and 0.1, hard sweeps or single-origin soft sweeps from SGV can occur. For a strong trade-off (Fig. 4C), these sweeps still produce essentially hard sweep patterns. We thus obtain the same distribution of patterns as in the single-locus case (compare with Fig. 4A).

- The combination of a multi-locus target and a weak trade-off creates the largest parameter space for adaptation from SGV (Fig. 4D). For $s_b \gg s_d$, almost all sweeps from SGV are soft (single- or multiple origin). The expected footprint of all these sweeps is clearly distinct from a hard sweep.

- Finally, for high mutation rates $\Theta_g \gtrsim 3$, we obtain cases of polygenic adaptation with small frequency shifts at many loci or super-soft sweeps with very large starting allele frequency in the SGV. In these cases, no major sweep locus exists (white areas in the figure).

Across all cases, the probability for a multiple-origin soft sweep depends only on the locus mutation parameter Θ_l, but not on Θ_g. It also does not dependent on selection strengths s_b and s_d, on epistasis among loci, or on the presence or absence of SGV. In contrast, the probabilities for hard sweeps and single-origin soft sweeps depend on on both, Θ_l and Θ_g, and on selection strength. For adaptations with a single-locus target, $\Theta_g = \Theta_l$, or strong trade-off, $s_d \geq s_b$, both of these sweep types are rare (Fig. 4A-C). This only changes for beneficial alleles with a multi-locus mutation target, $\Theta_g \gg \Theta_l$ and weak trade-off, $s_d \ll s_b$, where single-origin soft sweeps dominate for intermediate Θ_l values (Fig. 4D).

Our analysis of the multi-locus case is necessarily limited. While we assume full redundancy, mutations at different loci are often only partially redundant. They can have variable fitness effects and negative epistasis (which is a natural consequence of stabilizing selection) can be weaker. It remains to be explored how these factors influence the probabilities of sweep types at single loci, as well as the tendency for polygenic footprints with small allele frequency changes at many loci. Another factor that is ignored here, but which increases the percentage of sweeps from SGV among all observed sweeps comes to play if we condition on a short time period between the onset of selection and the time of observation (Hermisson and Pennings, 2005; Pritchard et al., 2010; Berg and Coop, 2015). Further complications, e.g. due to population structure, are briefly discussed below. Any comprehensive analysis of the probabilities of sweep types in genome scans needs to take these factors into account.

4 Complications

Population structure

Most natural populations are spatially extended and show at least some degree of geographic structure. While our genealogy-based definition of sweep types applies in the same way, the probabilities to observe hard or soft sweeps can depend on the strength of structure and also on the sampling strategy. Consider a spatially structured population that experiences a common novel selection pressure. If genetic exchange between different parts is weak, adaptation across the whole range may require independent origins of the adaptive allele in different geographic regions, a phenomenon also called *parallel adaptation* (Ralph and Coop, 2010). Thus, spatial structure can increase the probability of multiple-origin soft sweeps in global population samples, but not necessarily in local samples from a single region.

When does population structure favor soft sweeps? – If migration is long-ranged (like in an island model), only very strong structure (weak migration) will increase the probability of a soft sweep in global samples (Messer and Petrov, 2013). Consider two islands of size N_e, each with $\Theta = 4N_e u \leq 0.01$ (assuming $\Theta = \Theta_l = \Theta_g$), such that soft sweeps within islands are unlikely. In this regime, an adaptive mutation will fix on one island before an independent mutation can establish on the other island. If migration is larger than mutation, $4N_e m > \Theta$, it is more likely that the second island adapts from a beneficial migrant than from independent mutation, which results in a hard sweep. Soft sweeps are only likely if $4N_e m < \Theta$ (less than a single migrant per 100 generations) (Messer and Petrov, 2013). Note however, that effective isolation is only required during a short time window. This could increase the probability of spatial soft sweeps during periods of fragmentation, e.g. in glaciation refugia.

If migration is short-ranged and causes isolation by distance, geographic structure has a stronger effect (Ralph and Coop, 2010, 2015a). Consider a population of size N_e in continuous space with average dispersal distance σ. In such models, adaptations spread from their point of ori-

gin in so-called Fisher waves that proceed with constant speed $v \sim \sigma\sqrt{s_b}$. In populations that extend over d dispersal distances, the time-lag due to the finite wave speed enables parallel adaptation with multiple waves if $\Theta s_b \gtrsim v/(d\sigma) = \sqrt{s_b}/d$ (Ralph and Coop, 2010; Messer and Petrov, 2013). If parallel waves are driven by different origins of the same beneficial allele at the same locus, they constitute a spatial soft sweep. With large diameter d, spatial soft sweeps can occur for lower Θ than multiple-origin soft sweeps in the panmictic case (9), especially for strong selection (large s_b) or for alleles with weak trade-off (small s_d) that adapt from SGV (cf Ralph and Coop, 2015a).

Patterns of spatial soft sweeps – Whenever spatial structure and isolation by distance cause soft sweeps, despite low Θ, we expect to find hard sweeps in local samples, but soft sweeps in global samples. However, as Ralph and Coop (2010) observe, ongoing migration can blur this signal and patterns can look increasingly soft in local samples, too.

Vice-versa, there is also a parameter range where population structure is not causal for multiple origins, but still leads to geographic sweep patterns. Consider a population with $\Theta \sim 1$, such that multiple-origin soft sweeps are likely. Now, assume that this population is divided into k islands. We still have multiple origins globally, but if $\Theta/k \ll 1$ multiple origins on a single island are unlikely. Assume that islands are connected by migration of strength $1 \ll 4N_e m \ll 2N_e s_b$. In that case, gene flow will erase any trace of structure at neutral sites. However, migration during the selection window of $\sim 1/s_b$ generations is limited. Samples taken from an island where a beneficial mutation has occurred will thus be dominated by descendants from this mutation and may show patterns of local hard sweeps.

In natural populations, further factors can influence probabilities and patterns of soft sweeps. For example, heterogeneous spatial selection promotes parallel adaptation if gene flow between "adaptive pockets" is hampered by negative selection in interjacent regions (Ralph and Coop, 2015b). In contrast, gene surfing during range expansions may hamper soft sweeps, because successful copies of the beneficial allele need to orig-

672 inate in a narrow region near the expansion front (Gralka et al., 2016).
673 Currently, patterns of spatial soft sweeps have only partially been char-
674 acterized and remain an active field for study.

675 **Soft sweeps turning hard?**

676 Sweeps that appear soft for a recent adaptation may turn hard if we take
677 a sample at some later time. This occurs if descendants from only a single
678 copy of the beneficial allele dominate at that later time, either because of
679 genetic drift or due to ongoing selection.

680 Under neutrality, the average time to fixation or loss of an allele at
681 frequency x is $\approx 4N_e\big(x\log x^{-1}+(1-x)\log(1-x)^{-1}\big)$ generations (Ewens,
682 2004). Soft sweeps typically lead to intermediate haplotype frequencies,
683 but even for a major haplotype frequency of 99% the expected fixation
684 time is $> 0.1N_e$ generations. I.e., over time scales where sweep patterns
685 are generally visible, a soft sweep will usually not "harden". Demographic
686 events like bottlenecks increase drift and lead to accelerated hardening.
687 Still, a single bottleneck needs to be very strong to erase a soft sweep
688 pattern. By genetic drift alone, soft sweep signals will rather fade than
689 turn hard.

690 Fitness differences among soft sweep haplotypes can occur either be-
691 cause different mutations in the target of the beneficial allele are not fully
692 redundant or because of selection on linked variation. As long as these
693 differences are small relative to the primary effect of the allele, patterns
694 of recent soft sweeps appear to be remarkably stable. Pennings and Her-
695 misson (2006a) show that the patterns and probabilities of multiple-origin
696 soft sweeps do not change much for fitness differences up to 50% of the
697 primary effect if the sample is taken at fixation of the primary allele.
698 However, this no longer holds if fitness differences are of the same order
699 as the primary effect or if sampling occurs at a later time. For fitness
700 differences of $2N_e\Delta s_b \geq 100$, hardening occurs within the time window of
701 $\sim 0.1N_e$ generations where selection footprints are visible. An example
702 of such hardening in *Plasmodium* is described in the section on Microbes.
703 A related issue arises if a locus is hit by recurrent sweeps. Indeed, both

a stepwise approach of a new optimum and compensatory mutation at the same gene are well-documented adaptive responses to a new selection pressure. If adaptation is not mutation limited, these steps can occur in quick succession. Due to such "stacking" of soft sweeps (repeated Ewens sampling), we readily obtain a hardening of the pattern, unless the locus mutation rate is as high as $\Theta_l > 1$, where many haplotypes survive each sweep.

Importantly, a hardened soft sweep is not the same as a classical hard sweep, neither its biological interpretation (see the *Plasmodium* example below), nor its pattern. Both stages of the process are visible in time-series data and potentially also in the final pattern, e.g., if a strong (soft) resistance adaptation is followed by a weaker (hardening) compensatory mutation at the same gene. Theory and statistical methods to deal with such complex cases are currently lacking.

Hard sweeps looking soft?

In a recent paper, Schrider et al. (2015) suggested that patterns that look like soft sweeps could result from flanking regions of hard sweeps (soft shoulder effect). It is easy to see what could drive a spurious signal: for many summary statistics of polymorphism data, like π or the number of haplotypes, both, soft sweeps and flanking regions of hard sweeps result in weaker footprints than the core regions of hard sweeps. Selection scanners that base their prediction on a single local window can confuse one for the other. However, Schrider et al. (2015) do not provide an example where a pattern has been misclassified because of the shoulder effect. It seems that problems can be avoided if scans pre-select loci with strong signals (Garud et al., 2015; Garud and Petrov, 2016) or if selection footprints are evaluated together with their local genomic context (Sheehan and Song, 2016; Schrider and Kern, 2016a).

Another concern that has been voiced is that allelic gene conversion during a hard sweep could mimic a soft sweep pattern (Schrider et al., 2015). Indeed, gene conversion at the selected site can change the pattern of a hard sweep (Pennings and Hermisson, 2006b; Jones and Wakeley,

2008): A conversion event with short conversion tract around the selected allele can place this allele onto another genetic background. Like for a multiple-origin soft sweep, the beneficial allele becomes associated with several ancestral haplotypes that stretch across the adaptive target. LD across the selected site will be positive, in contrast to a classical hard sweep, and in accordance with a soft sweep. However, important differences remain. Like single crossing-over, gene conversion is a recombination event. As explained in the Footprints section, recombination during a hard sweep typically leads to recombination haplotypes at a low frequency in the sample (see also Jones and Wakeley, 2008), in contrast to the intermediate-frequency haplotypes created by recurrent mutation at a soft sweep (Pennings and Hermisson, 2006b).

We can apply this argument to the polymorphism pattern in three immunity genes in *Drosophila simulans* reported by Schlenke and Begun (2005). All three genes show extreme values of positive LD, caused by invariant haplotypes that extend across the gene. In each case, two haplotypes are found at an intermediate frequency in a combined Californian sample. Whereas it appears unlikely that such a pattern, repeated across three genes, has been created by gene conversion, it is perfectly expected under the scenario of multiple-origin soft sweeps. Clearly, gene conversion can also be excluded if different redundant mutations at the same locus contribute to a soft sweep pattern (as for the lactase case discussed below).

A case where a hard sweep will look soft results if selection acts on a fully recessive allele. A recessive allele behaves essentially neutrally as long as its frequency is smaller than $x < x_0 = 1/\sqrt{2N_e s_b}$. Its trajectory resembles the one of an allele that derives from neutral SGV with starting frequency x_0 at time T_S. Since the impact of selection on the sweep depends only on the shape of the trajectory, both footprints are indistinguishable. Indeed, equivalent models have been used to describe recessive hard sweeps (Ewing et al., 2011) and single-origin soft sweeps from SGV (Berg and Coop, 2015). As stressed by Berg and Coop (2015), both scenarios can only be distinguished if additional biological informa-

tion about dominance is available. Note, finally, that the combined effect of a recessive hard sweep or single-origin soft sweep *and* gene conversion may indeed create a pattern that mimics a multiple-origin soft sweep.

5 Evidence

How do our theoretical expectations of sweep types compare with observed data? Evidence for both soft and hard sweeps exist for many species. In this section, we focus on three exemplary cases: *Drosophila*, humans, and microbial adaptation.

Drosophila

Estimates of Watterson's Θ_W per base pair from the least constrained regions (short introns) in African *D. melanogaster* and *D. simulans* (Parsch et al., 2010; Andolfatto et al., 2010) result in lower bounds $\Theta_l \geq \Theta_W/3 \approx 0.008$ (for *mel*), and $\Theta_l \gtrapprox 0.013$ (for *sim*), respectively. These values are right at the boundary where multiple-origin soft sweeps start to appear in larger samples (4% resp. 6% in a sample of size 100, Eq. 9). Given that short-term N_e likely exceeds the values from polymorphism data and mutational targets may often consist of at least a few sites, $\Theta_l \sim 0.1$ is a plausible estimate for the per-locus rate. We thus expect to see a mixture of hard sweeps and multiple-origin soft sweeps in data (Fig. 4). For adaptations with a weak trade-off, also single-locus soft sweeps can contribute, in particular if mutation targets are polygenic ($\Theta_g > \Theta_l$).

Recent genome scans by Garud and Petrov (2016) and Sheehan and Song (2016) for a Zambian *D. melanogaster* population confirm to this expectation. Both find a mix of hard and soft sweep signals for regions with the strongest evidence for recent selection. Garud et al. (2015) and Garud and Petrov (2016) report much higher rates for soft sweeps for *melanogaster* from North Carolina. According to their test, all top 50 signals are consistent with soft sweeps rather than hard sweeps. Where this difference comes from is yet unclear. Potential causes include more recent selection in the American population and a larger short-term N_e,

but also confounding factors like inbreeding and admixture.

Similar to genome scans, also an inspection of selection footprints at genes with well-characterized function shows a mix of signals. However, clear patterns that are fully consistent with the simplest sweep models are rather the exception. Among the clearest examples are patterns consistent with multiple-origin soft sweeps in *D. simulans* immunity receptor genes (Schlenke and Begun, 2005, discussed above) and the loss of pigmentation in *D. santomea* due to inactivation of a cis-regulatory element at the *tan* locus (Jeong et al., 2008). The latter case is an example for an adaptive loss-of-function allele that can be produced by a large number of fully redundant mutations (three independent origins have been identified, two deletions and one double substitution). The adaptation underlies a species characteristic of *D. santomea* and is a rare case of a sweep (hard or soft) with clear phenotypic effect that is completed throughout the species.

A well-resolved example for a partial hard sweep is the *Bari-Jheh* insertion in *D. melanogaster*, a gain of function mutation for protection to oxidative stress (Gonzáles et al., 2008; Guio et al., 2014). Since the allele has a large fitness trade-off, it is unclear whether it is *de-novo* or from SGV. Both cases would result in a hard sweep pattern. An example for a partial sweep from SGV is the *CHKov1* gene in *D. melangaster* (Aminetzach et al., 2005). At the gene, an exonic insertion that provides insecticide resistance has recently swept to high frequency. Strong divergence of the sweep haplotype shows that the allele must be very old and long predates the selection pressure. However, the pattern is not easily explained by a simple selection history. Indeed, the allele most likely has a history of pre-adaptation to viral infection (Magwire et al., 2011) and the signal is shaped by multiple selection episodes. Note that single-origin sweeps (soft or hard) are likely for exonic insertion adaptations, which are not easily replicated and plausibly have a very low allelic mutation rate.

A stepwise selection history is also evident for evolution of resistance to oranophosphate insecticides at the *Ace* locus in *D. melanogaster* (Menozzi et al., 2004; Karasov et al., 2010; Messer and Petrov, 2013). In response to a very recent selection pressure, the same base substitution has occurred

at least three times independently in different world regions (global soft sweep) and partially also within regions (local soft sweep). Resistance is reinforced by two further substitutions at the *Ace* locus. The haplotype pattern of the most resistant type shows signs of partial (but not complete) hardening.

Humans

For Humans, estimates of a long-term $N_e \sim 10^4$ (Takahata, 1993) are so heavily influenced by sporadic demographic events, like the bottleneck connected to out-of-Africa migration, that this number is almost useless for population genetic theory. More refined methods based on deep sequencing estimate changes in N_e from $\sim$ 14000 pre-agriculture (10000 years ago) to $\sim$ 500000 presently in Africa (Chen et al., 2015). With a mutation rate of 1 - $2 \cdot 10^{-8}$, this leads to a "recent" $\Theta \approx 10^{-3}$ for point mutations, consistent with estimates of Θ from singletons (Mathieson and McVean, 2014). Using similar assumptions about target sizes and short-term N_e as for *Drosophila*, we arrive at ~ 0.01 as a rough estimate for Θ_l or Θ_g. This is an order of magnitude lower than for *Drosophila* and in a range where hard sweeps from new mutations dominate (Fig. 4). However, there are reasons why adaptation from SGV may be more prevalent. First, population growth protects rare alleles from loss due to drift (Otto and Whitlock, 1997; Hermisson and Pennings, 2005). This enhances the probability of adaptation from SGV and is not captured by Eq. (5). Second, many selection pressures result from the dramatic changes in nutrition and population density since the advent of agriculture, or from pathogens that have spread in response to increased human density. Since these selection pressures are very young, almost all adaptive alleles that are now at high frequency must have emerged from SGV (Przeworski et al., 2005; Novembre and Han, 2012). Depending on their fitness trade-off or potential pre-adaptation in the ancestral environment, adaptations will display signals of either hard or soft partial sweeps.

Tests for candidate genes and genome scans that can distinguish hard and soft sweeps either find evidence for both types (Peter et al., 2012;

Schrider and Kern, 2016a) or predominantly soft sweeps (Schrider and Kern, 2016b). Detailed studies of individual genes support the role of SGV in recent adaptations. A clear example is the CCR5 mutation conferring resistance against HIV (reviewed by Novembre and Han, 2012). Other loci once again reveal more complex adaptive histories. E.g., the recent spread of a derived haplotype at a Serine protease inhibitor gene in Yorubans has been identified as resurrection of a pseudogene using pre-adapted SGV (Seixas et al., 2012). Adaptation to high altitude in Tibetans involves a variant of the hypoxia pathway gene EPAS1 that has introgressed into modern humans from Denisovans (Huerta-Sanchez et al., 2014).

A consistent finding is that recent sweeps in humans are never species-wide (Mallick et al., 2016), but rather partial and regional. Regional patterns arise not only as a response to regional selection pressures (e.g., adaptation to Lassa fever in regions where the disease is endemic, Andersen et al., 2012), but also because of parallel adaptation to the same selection pressure across geographic regions (Coop et al., 2009; Ralph and Coop, 2010, 2015a; Novembre and Han, 2012). The prime example is light skin pigmentation, a trait that has evolved several times independently in Europeans and Asians. Parallel adaptation in this case is facilitated by a larger genomic mutation target, comprising several genes. The pattern at each single gene locus is still consistent with a hard sweep, either from new mutation after the out-of-Africa migration or from a deleterious standing variant (Coop et al., 2009; Jablonski and Chaplin, 2012).

Given the low estimated value of Θ_l, maybe the most surprising finding is the evidence of multiple-origin soft sweeps – including two of the most prominent examples of recent adaptation in humans: Besides the lactase adaptation, which has a large mutation target (see separate box), this is the *Duffy* FY-0 mutation conveying resistance against Vivax malaria. For *Duffy*, the exact same point mutation has been found on three different haplotypes. Two haplotypes with linkage disequilibrium across the selected site are found at intermediate frequency in sub-Saharan Africa (Hamblin and Di Rienzo, 2000). One further independent origin has been described in Papua New Guinea (Zimmerman et al., 1999). There is evi-

dence for the influence of geographic structure for both, lactase and *Duffy* (global soft sweep), but both are also locally soft in Africa.

Box: The human lactase gene example

The lactase enzyme enables humans (and other mammals) to digest the milk sugar lactose and to consume milk without adverse side-effects. All humans produce lactase as infants, but in the ancestral wildtype the gene is down-regulated during childhood. However, around 61% of Europeans, 34% of Africans, and 28% of East Asians today are lactose tolerant also as adults (Ingram et al., 2009a). The trait has been attributed to derived allelic variants in an enhancer region of LCT (the lactase gene). To date, 5 different functionally confirmed mutations have been identified, all in a narrow region of around 100bp, 14000bp upstream of LCT (Tishkoff et al., 2007; Ingram et al., 2009a; Jones et al., 2013). They display clear geographic structure: lactose tolerance in Europe and central Asia is almost exclusively due to a single SNP, while a different causal SNP dominates among milk consumers in Tanzania and Kenya. Polymorphism data from LCT in these populations show haplotype patterns compatible with a strong partial sweep (Bersaglieri et al., 2004; Tishkoff et al., 2007) that is globally soft, but hard in each single region. However, more recent studies in Ethiopian and Sudanese lactose digesters identify multiple functional variants in single populations (Ingram et al., 2009b; Jones et al., 2013; Ranciaro et al., 2014): a local soft sweep from multiple mutational origins. Although the regulatory mechanism is not yet fully resolved, it seems that the associated SNPs are loss-of-function mutations (preventing down-regulation) that lead to a gain in enzyme activity (Ingram et al., 2009b). Parallel origin of the adaptive phenotype is facilitated by a large mutational target, leading to a large Θ_l for adult lactose tolerance.

Microbes

Some of the most compelling evidence for rapid adaptation comes from microbes, especially from pathogens that evolve drug resistance. We discuss two recombining pathogens in which selective sweeps have been studied:

HIV and the Malaria parasite *P. falciparum*.

Drug resistance in HIV typically evolves independently in each patient. Work on selective sweeps in HIV therefore focuses on within-patient populations that start off as susceptible and then acquire one or several resistance mutations. The identity of drug resistance mutations in HIV are well known for all commonly used antiretroviral drugs (Wensing et al., 2015). In the absence of treatment, within-patient HIV populations have large census sizes with around 10^8-10^9 active virus-producing cells (Haase, 1994; Coffin and Swanstrom, 2013). Since HIV has a high mutation rate of 10^{-7}- 10^{-5}, depending on mutation type (Abram et al., 2010; Zanini et al., 2016), almost all single nucleotide mutations are present as SGV at any time. With values of $\Theta_l \sim 10$ to 10^4, one would expect that adaptation occurs only via soft selective sweeps. However, the scenario may be different for patients under treatment. Although the effect of treatment on the effective population size is unknown, successful treatment can reduce the number of viral particles in the blood by several orders of magnitude.

Paredes et al. (2010) determined whether resistance mutations were already present as SGV in the viral population of 183 patients. Comparing blood samples from before the start of treatment and after treatment had failed, they found that patients with resistance mutations present as SGV were more likely to fail treatment due to resistance evolution (see also Wilson et al., 2016). Another study (Jabara et al., 2011) looked only at one patient, but did very deep and accurate sequencing. The same haplotypes carrying resistance mutations were observed at low frequency in SGV, and at much higher frequency after treatment.

There is also clear evidence for multiple-origin soft sweeps in HIV. The most palpable example comes from patients treated with the drug Efavirenz. The main resistance mutation against this drug is a K103N substitution in the reverse transcriptase gene. The wildtype codon for K (lysine) is AAA. Since codons AAT or AAC both encode N (asparagine), two redundant mutations create the resistant allele. For multiple-origin sweeps, we expect mixtures of AAT and AAC codons in drug resistant

populations. Indeed, this is frequently observed (Pennings et al., 2014). However, sometimes only one codon, AAC or AAT, is found. At least some of theses cases are likely hard sweeps. Secondary hardening due to fitness differences (codon bias) among AAC and AAT codons appears unlikely as both versions are observed. A possible explanation is hardening due to linked selection.

Another factor favoring hard sweeps is (temporary) treatment success. Indeed, some patients do not evolve drug resistance at all, or only after many years of treatment (Harrigan et al., 2005; UK Collaborative Group on HIV Drug Resistance, 2005). This suggests a role of *de novo* mutation in resistance evolution. Depending on the patient and the treatment, HIV populations may actually be mutation limited. This is corroborated by a recent analysis of patient derived sequences from the late 1980s until 2013 (Feder et al., 2016). The authors show that the reduction in diversity associated with the fixation of a resistance mutation increases as treatment becomes more effective. This suggests that better treatments lead to harder sweeps, likely because they reduce the population size and move HIV into a mutation-limited regime.

In *Plasmodium falciparum*, unlike in HIV, single drug resistant strains often spread across wide geographic areas. The evolution of drug resistance is thus mostly studied at the level of human populations, rather than the level of parasite populations inside single patients. Patterns of drug resistance in *P. falciparum* roughly come in two forms: When several substitutions are needed to create a resistant allele, observed sweeps are locally hard, but globally soft. When a single substitution or gene duplication event can create a resistant allele, very soft sweeps are observed, both locally and globally.

A striking example of a multiple-origin soft sweep is the evolution of artemisinin resistance due to mutations in the Kelch gene in South-East Asia. Anderson et al. (2017) analysed samples from western Thailand from 2001 (before the spread of artemisinin resistance) through 2014 (when Kelch mutations had nearly fixed). They found 32 different non-synonymous mutations in the Kelch gene, most of which create a resistant

phenotype, and estimate a large target size between 87 and 163bp for the resistance allele. However, for the most recent years 2013/14, the authors find that one of the resistance mutations (C580Y) has spread to high frequency (Fig. 1B in Anderson et al., 2017). This has led to partial hardening of the sweep and a lower diversity in samples, likely because of a fitness advantage of C580Y relative to the other resistance mutations. We may see complete hardening of this sweep in future samples. Therefore, as the authors point out, only the observation of a soft sweep in real time allows for the correct assessment of a high risk of resistance evolution. Retrospect analyses after secondary hardening and erroneous classification as a classical hard sweep would miss the essential information.

Another example of a multiple-origin soft sweep comes from resistance evolution to mefloquine and artemisinin due to gene amplification at *pfmdr1* (Nair et al., 2007). The authors find evidence for 5 to 15 independent origins of the gene duplication, based on an analysis of 5' and 3' breakpoints of the amplification events. Since the data was collected two decades after the initial spread of resistance, even more independent origins of this gene amplification may have existed originally.

One example of a locally hard, but globally soft sweep is observed at the *dhfr* gene. Mutations at this gene reduce susceptibility to pyrimethamine. Single, double, and triple mutant haplotypes occur in Africa. In contrast, only one predominant triple mutant haplotype was found in South-East Asia, displaying a classical (local) hard sweep pattern (Nair et al., 2003). The same triple mutant haplotype was found in East and South Africa. Since this haplotype differs from the susceptible or resistant types in Africa, the most parsimonious explanation is that it stems from South-East Asia (Roper et al., 2004). In another study, however, researchers found the same triple mutant on an unrelated genetic background in the West-African archipelago of São Tomé and Principe (Salgueiro et al., 2010), which makes this a globally soft sweep. Another example of a locally hard, but globally soft sweep is found in chloroquine resistance, which is conferred by several non-synonymous mutations in the chloroquine resistance transporter (*crt*) gene (Escalante et al., 2009;

Mehlotra et al., 2001).

6 Conclusions: Beyond "hard" and "soft"

In a mutation-limited world, recent adaptation leads to hard sweeps and leaves clear and well-understood footprints in genomic diversity. However, existing theory and data indicate that this is not the world we live in. More often than not, adaptation is rapid, resorting to genetic material from various sources, including standing variation, or the recurrent origin of beneficial alleles through mutation or migration.

From theory, we have seen that a single parameter, the population mutation rate $\Theta = 4N_e u$, is most important in separating the rapid world from the mutation limited one. However, this parameter is more complex than it may seem. It depends on the specific beneficial allele and its mutational target size. It also depends on the timing of the adaptive event and the corresponding short-term effective population size. This leads to a large variance for Θ, spanning regions of hard and soft sweeps in most species. Consequently, there is empirical evidence for soft sweeps even in humans, and for hard sweeps even in microbes, despite of huge differences in population size. What is more, the best resolved empirical cases tell us that real adaptive stories go beyond a simplistic hard-or-soft dichotomy.

If adaptation is not mutation limited, we see diversity also among selection footprints. There are classical hard sweeps and two types of "classical" soft sweeps, from a single origin or from multiple origins of the beneficial allele. However, more often then not, real patterns do not fit these model archetypes: For example, there are stacked soft sweeps with secondary hardening, soft sweeps with spatial components, adaptation from pre-adapted variation or from introgression, and there is polygenic adaptation.

There is little reason to argue that either hard sweeps or soft sweeps do not exist. But there is good reason to build on existing concepts and to go on and explore.

7 Simulation methods

We simulate a standard Wright-Fisher model of $2N_e$ haploids. Individual genotypes have m unlinked loci with two alleles each, a wildtype allele a_i and a mutant allele A_i, $1 \leq i \leq m$. Mutation from the wildtype to the mutant occurs at each locus with rate u. Although loci are bi-allelic, we distinguish each new mutational origin of a mutant A_i allele from previous mutants at the same locus and record its frequency separately. Let x_i be the total mutant frequency at locus i. Prior to T_S, mutant alleles are either neutral or deleterious with selection coefficient s_d. After T_S, each genotype with at least one beneficial allele has selection coefficient s_b. Assuming linkage equilibrium, this results in a marginal fitness of the wildtype of $1 + s_b(1 - \prod_{j \neq i}(1 - x_i))$. In the life cycle, mutation and selection is followed by independent multinomial sampling at each locus. Prior to T_S, we run the simulation model for $10/s_d$ generations to approach mutation-selection-drift balance. After T_S, selection turns positive and we follow the trajectories of mutant alleles across all loci. Simulations are stopped once the frequency of the wildtype genotype (with allele a_i at all loci) is $< 5\%$. We record the sweep pattern at the locus with the highest frequency of the beneficial allele at this time, given that this locus has experienced an allele frequency shift of at least 50%.

8 Acknowledgments

We thank Graham Coop, Alison Feder, Nandita Garud, Ilse Höllinger, Sebastian Matuszewski, Philipp Messer and an anonymous referee for comments on the manuscript.

References

M. E. Abram, A. L. Ferris, W. Shao, W. G. Alvord, and S. H. Hughes. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *J. Virol.*, 84(19):9864–9878, 2010.

Y. T. Aminetzach, J. M. Macpherson, and D. A. Petrov. Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. *Science*, 309:764–767, 2005.

K. G. Andersen, I. Shylakhter, S. Tabrizi, S. R. Grossman, C. T. Happi, and P. C. Sabeti. Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Phil. Trans. R. Soc. B*, 367:868–877, 2012.

T. J. C. Anderson, S. Nair, M. McDew-White, I. H. Cheeseman, S. Nkhoma, F. Bilgic, R. McGready, E. Ashley, A. P. Phyo, N. J. White, and F. Nosten. Population parameters underlying an ongoing soft sweep in Southeast Asian malaria parasites. *Mol. Biol. Evol.*, 34:131–144, 2017.

P. Andolfatto, K. M. Wong, and D. Bachtrog. Effective population size and the efficacy of selection on the X chromosomes of two closely related *Drosophila* species. *Genome Biol. Evol.*, 3:114–128, 2010.

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

J. Barroso-Batista, A. Sousa, M. Lourenco, M.-L. Bergman, D. Sobral, J. Demengeot, K. B. Xavier, and I. Gordo. The first steps of adaptation of *Escherichia coli* to the gut are dominated by soft sweeps. *PLoS Genetics*, 10:e1004182, 2014.

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

N. H. Barton. Understanding adaptation in large populations. *PLoS Genetics*, 6:e1000987, 2010.

J. J. Berg and G. Coop. Population genetic signal of polygenic adaptation. *PLoS Genetics*, 10:e1004412, 2014.

J. J. Berg and G. Coop. A coalescent model for a sweep of a unique standing variant. *Genetics*, 201:707–725, 2015.

T. Bersaglieri, P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hischhorn. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.*, 74:1111–1120, 2004.

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

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

H. Chen, J. Hey, and K. Chen. Inferring very recent population growth rate from population-scale sequencing data: Using a large-sample coalescent estimator. *Mol. Biol. Evol.*, 32:2996–3011, 2015.

J. Coffin and R. Swanstrom. HIV pathogenesis: dynamics and genetics of viral populations and infected cells. *Cold Spring Harbor perspectives in medicine*, 3(1):a012526, 2013.

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

G. Coop, J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li, D. Absher, R. M. Myers, L. L. Cavalli-Sforza, M. W. Feldman, and J. K. Pritchard. The role of geography in human adaptation. *PLoS Genetics*, 5:e1000500, 2009.

M. M. Desai and D. S. Fisher. Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics*, 176:1759–1798, 2007.

E. Elyashiv, S. Sattath, T. T. Hu, A. Strutsovsky, G. McVicker, P. Andolfatto, G. Coop, and G. Sella. A genomic map of the effects of linked selection in *Drosophila*. *PLoS Genetics*, 12:e1006130., 2016.

A. A. Escalante, D. L. Smith, and Y. Kim. The dynamics of mutations associated with anti-malarial drug resistance in *Plasmodium falciparum*. *Trends Parasit.*, 25(12):557–563, 2009.

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

W. J. Ewens. *Mathematical Population Genetics*. Springer, 2 edition, 2004.

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

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

A. F. Feder, S.-Y. Rhee, S. P. Holmes, R. W. Shafer, D. A. Petrov, and P. S. Pennings. More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. *eLife*, 5:e10670, 2016.

A. Ferrer-Admetlla, M. Liang, T. Korneliussen, and R. Nielsen. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.*, 31:1275–1291, 2014.

N. R. Garud and D. A. Petrov. Elevated linkage disequilibrium and signatures of soft sweeps are common in *Drosophila melanogaster*. *Genetics*, 203:863–880, 2016.

N. R. Garud, P. W. Messer, E. O. Buzbas, and D. A. Petrov. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics*, 11:e1005004, 2015.

P. J. Gerrish and R. E. Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102:127–144, 1998.

J. Gonzáles, K. Lenkov, M. Lipatov, J. M. Macpherson, and D. A. Petrov. High rate of recent transposable element–induced adaptation in *Drosophila melanogaster*. *PLoS Biology*, 6:e251, 2008.

M. Gralka, F. Stiewe, F. Farrell, W. Möbius, B. Waclaw, and O. Hallatschek. Allele surfing promotes microbial adaptation from standing variation. *Ecology Letters*, 19:889–898, 2016.

L. Guio, M. G. Barrón, and J. González. The transposable element *Bari-Jheh* mediates oxidative stress response in *Drosophila*. *Mol. Ecol.*, 23: 2020–2030, 2014.

A. T. Haase. The role of active and covert infections in lentivirus pathogenesis. *Ann. NY Acad. Sci.*, 724(1):75–86, 1994.

M. T. Hamblin and A. Di Rienzo. Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. *Am. J. Hum. Genet.*, 66:1669–1679, 2000.

P. R. Harrigan, R. S. Hogg, W. W. Dong, B. Yip, B. Wynhoven, J. Woodward, C. J. Brumme, Z. L. Brumme, T. Mo, C. S. Alexander, et al. Predictors of HIV drug-resistance mutations in a large antiretroviral-naive cohort initiating triple antiretroviral therapy. *J. Infect. Dis.*, 191 (3):339–347, 2005.

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

E. Huerta-Sanchez, X. Jin, Asan, Z. Biaba, B. M. Peter, N. Vinckenbosch, Y. Liang, X. Yi, M. He, M. Somel, P. Ni, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512: 194–197, 2014.

C. J. E. Ingram, C. A. Mulcare, Y. Itan, M. G. Thomas, and D. M. Swallow. Lactose digestion and the evolutionary genetics of lactase persistence. *Human Genetics*, 124:579–591, 2009a.

C. J. E. Ingram, T. O. Raga, A. Tarekegn, S. L. Browning, M. F. Elamin, E. Bekele, T. M. G, M. E. Weale, N. Bradman, and D. M. Swallow.

Multiple rare variants as a cause of a common phenotype: Several different lactase persistence associated alleles in a single ethnic group. *Mol. Biol. Evol.*, 69:579–588, 2009b.

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

C. B. Jabara, C. D. Jones, J. Roach, J. A. Anderson, and R. Swanstrom. Accurate sampling and deep sequencing of the HIV-1 protease gene using a primer ID. *Proc. Nat. Acad. Sci.*, 108(50):20166–20171, 2011.

N. G. Jablonski and G. Chaplin. Human skin pigmentation, migration and disease susceptibility. *Phil. Trans. R. Soc. B*, 367:785–792, 2012.

J. D. Jensen. On the unfounded enthusiasm for soft selective sweeps. *Nature Commun.*, 5:5281, 2014.

S. Jeong, M. Rebeiz, P. Andolfatto, T. Werner, J. True, and C. S. B. The evolution of gene regulation underlies a morphological difference between two *Drosophila* sister species. *Cell*, 132:783–793, 2008.

B. L. Jones, T. O. Raga, A. Liebert, P. Zmarz, E. Bekele, E. T. Danielsen, A. Krüger Olsen, N. Bradman, J. T. Troelsen, and D. M. Swallow. Diversity of lactase persistence alleles in Ethiopia: Signature of a soft selective sweep. *Am. J. Hum. Genet.*, 93:538–544, 2013.

D. A. Jones and J. Wakeley. The influence of gene conversion on linkage disequilibrium around a selective sweep. *Genetics*, 180:1251–1259, 2008.

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

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

Y. Kim and R. Nielsen. Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167(3):1513–1524, 2004.

H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475:493–496, 2011.

M. M. Magwire, F. Bayer, C. L. Webster, C. Cao, and F. M. Jiggins. Successive increases in the resistance of *Drosophila* to viral infection through a transposon insertion followed by a duplication. *PLoS Genetics*, 7:e1002337, 2011.

S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.

I. Mathieson and G. McVean. Demography and the age of rare variants. *PLoS Genetics*, 8:e1004528, 2014.

J. Maynard Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genet. Res. Camb.*, 23:23–35, 1974.

R. K. Mehlotra, H. Fujioka, P. D. Roepe, O. Janneh, L. M. Ursos, V. Jacobs-Lorena, D. T. McNamara, M. J. Bockarie, J. W. Kazura, D. E. Kyle, et al. Evolution of a unique *Plasmodium falciparum* chloroquine-resistance phenotype in association with pfcrt polymorphism in Papua New Guinea and South America. *Proc. Nat. Acad. Sci.*, 98(22):12689–12694, 2001.

P. Menozzi, M. A. Shi, A. Lougarre, Z. H. Tang, and D. Fournier. Mutations of acetylcholinesterase which confer insecticide resistance in *Drosophila melanogaster* populations. *BMC Evol. Biol.*, 4:4, 2004.

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

P. W. Messer, S. P. Ellner, and N. G. Hairston Jr. Can population genetics adapt to rapid evolution? *Trends Genet.*, 32:408–418, 2016.

S. Nair, J. T. Williams, A. Brockman, L. Paiphun, M. Mayxay, P. N. Newton, J.-P. Guthmann, F. M. Smithuis, T. T. Hien, N. J. White, et al. A

selective sweep driven by pyrimethamine treatment in Southeast Asian malaria parasites. *Mol. Biol. Evol.*, 20(9):1526–1536, 2003.

S. Nair, D. Nash, D. Sudimack, A. Jaidee, M. Barends, A.-C. Uhlemann, S. Krishna, F. Nosten, and T. J. Anderson. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Mol. Biol. Evol.*, 24(2):562–573, 2007.

J. Novembre and E. Han. Human population structure and the adaptive response to pathogen-induced selection pressures. *Phil. Trans. R. Soc. B*, 367:878–886, 2012.

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

S. P. Otto and M. Whitlock. The probability of fixation in populations of changing size. *Genetics*, 146:723–733, 1997.

R. Paredes, C. M. Lalama, H. J. Ribaudo, B. R. Schackman, C. Shikuma, F. Giguel, W. A. Meyer, V. A. Johnson, S. A. Fiscus, R. T. D'Aquila, R. M. Gulick, and D. R. Kuritzkes. Pre-existing minority drug-resistant HIV-1 variants, adherence, and risk of antiretroviral treatment failure. *J. Infect. Dis.*, 201(5):662–671, 2010.

J. Parsch, S. Novozhilov, S. S. Saminadin-Peter, K. M. Wong, and P. Andolfatto. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Mol. Biol. Evol.*, 27:1226–1234, 2010.

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

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

P. S. Pennings, S. Kryazhimskiy, and J. Wakeley. Loss and recovery of genetic diversity in adapting populations of HIV. *PLoS Genetics*, 10 (1):e1004000, 2014.

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

J. K. Pritchard, J. K. Pickrell, and G. Coop. The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.*, 20:R208–R215, 2010.

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

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

P. Ralph and G. Coop. Parallel adaptation: One or many waves of advance of an advantageous allele? *Genetics*, 186(2):647–668, 2010.

P. Ralph and G. Coop. The role of standing variation in geographic convergent adaptation. *Am. Nat.*, 186:S5–S23, 2015a.

P. Ralph and G. Coop. Convergent evolution during local adaptation to patchy landscapes. *PLoS Genetics*, 11:e1005630, 2015b.

A. Ranciaro, M. C. Campbell, J. B. Hirbo, W.-Y. Ko, A. Froment, P. Anagnostou, M. J. Kotze, M. Ibrahim, T. Nyambo, S. A. Omar, and S. A. Tishkoff. Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am. J. Hum. Genet.*, 94:496–510,, 2014.

C. Roper, R. Pearce, S. Nair, B. Sharp, F. Nosten, and T. Anderson. Intercontinental spread of pyrimethamine-resistant malaria. *Science*, 305(5687):1124–1124, 2004.

P. Salgueiro, J. L. Vicente, C. Ferreira, V. Teófilo, A. Galvão, V. E. do Rosário, P. Cravo, and J. Pinto. Tracing the origins and signatures of selection of antifolate resistance in island populations of *Plasmodium falciparum*. *BMC Infect. Dis.*, 10(1):1, 2010.

T. A. Schlenke and D. Begun. Linkage disequilibrium and recent selection at three immunity receptor loci in *Drosophila simulans*. *Genetics*, 169: 2013–2022, 2005.

1319 D. R. Schrider and A. D. Kern. S/HIC: Robust identification of soft
1320 and hard sweeps using machine learning. *PLoS Genetics*, 12:e1005928,
1321 2016a.

1322 D. R. Schrider and A. D. Kern. Soft sweeps are the dominant
1323 mode of adaptation in the human genome. *bioRxiv*, page doi:
1324 https://doi.org/10.1101/090084, 2016b.

1325 D. R. Schrider, F. K. Mendes, M. W. Hahn, and A. D. Kern. Soft shoulders
1326 ahead: Spurious signatures of soft and partial selective sweeps result
1327 from linked hard sweeps. *Genetics*, 200:267–284, 2015.

1328 S. Seixas, N. Ivanova, Z. Ferreira, J. Rocha, and B. L. Victor. Loss and
1329 gain of function in SERPINB11: An example of a gene under selection
1330 on standing variation, with implications for host-pathogen interactions.
1331 *PLoS One*, 7:e32518, 2012.

1332 S. Sheehan and Y. S. Song. Deep learning for population genetic inference.
1333 *PLoS Comp. Biol.*, 12:e1004845, 2016.

1334 P. Sjödin, I. Kaj, S. Krone, M. Lascoux, and M. Nordborg. On the mean-
1335 ing and existence of an effective population size. *Genetics*, 169:1061–
1336 1070, 2005.

1337 W. Stephan, Y. S. Song, and C. H. Langley. The hitchhiking effect on
1338 linkage disequilibrium between linked neutral loci. *Genetics*, 172:2647–
1339 2663, 2006.

1340 F. Tajima. Statistical method for testing the neutral mutation hypothesis.
1341 *Genetics*, 123:585–595, 1989.

1342 N. Takahata. Allelic genealogy and human evolution. *Mol. Biol. Evol.*,
1343 10:2–22, 1993.

1344 S. A. Tishkoff, F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt,
1345 et al. Convergent adaptation of human lactase persistence in Africa
1346 and Europe. *Nature Genetics*, 39:31–40, 2007.

UK Collaborative Group on HIV Drug Resistance. Long term probability of detection of HIV-1 drug resistance after starting antiretroviral therapy in routine clinical practice. *Aids*, 19(5):487–494, 2005.

O. A. van Herwaarden and N. J. van der Wal. Extinction time and age of an allele in a large finite population. *Theor. Pop. Biol.*, 61:311–318, 2002.

J. Wakeley. *Coalescent Theory: An Introduction*. Roberts & Company Publishers, 2008.

A. M. Wensing, V. Calvez, H. F. Günthard, V. A. Johnson, R. Paredes, D. Pillay, R. W. Shafer, and D. D. Richman. 2015 update of the drug resistance mutations in HIV-1. *Topics in Antiviral Medicine*, page 132, 2015.

B. A. Wilson, D. A. Petrov, and P. W. Messer. Soft selective sweeps in complex demographic scenarios. *Genetics*, 198:669–684, 2014.

B. A. Wilson, N. R. Garud, A. F. Feder, Z. J. Assaf, and P. S. Pennings. The population genetics of drug resistance evolution in natural populations of viral, bacterial and eukaryotic pathogens. *Mol. Ecol.*, 25(1): 42–66, 2016.

F. Zanini, V. Puller, J. Brodin, J. Albert, and R. Neher. *In-vivo* mutation rates and fitness landscape of HIV-1. *arXiv preprint arXiv:1603.06634*, 2016.

P. A. Zimmerman, I. Woolley, G. L. Masinde, M. S. M, D. T. McNamara, F. Hazlett, C. S. Mgone, M. P. Alpers, B. Genton, B. A. Boatini, and J. W. Kazura. Emergence of FYA^{null} in a *Plasmodium vivax*-endemic region of Papua New Guinea. *Proc. Nat. Acad. Sci.*, 96:13973–13977, 1999.

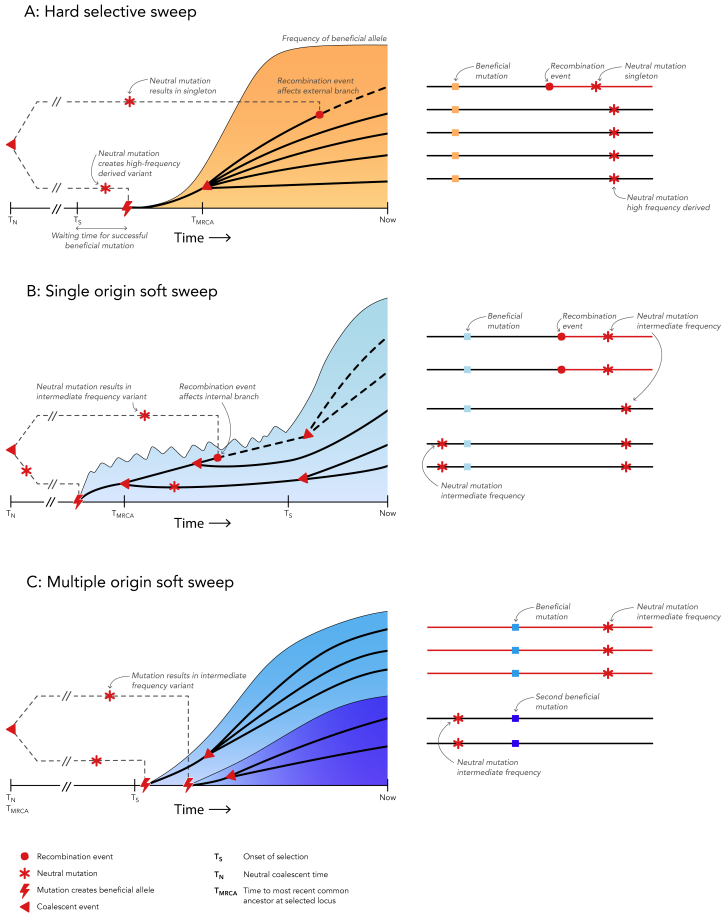

Figure 1: Hard and soft sweep types. Colored regions mark the frequency of those copies of the beneficial allele that still have descendants at the time of sampling. Black and dashed lines show coalescent histories at linked sites. On the right, mutations and recombination events are also shown on haplotypes of the five sampled individuals. (A) For a hard sweep, the time to the most recent common ancestor at the selected site T_{MRCA} is shorter than the time since the onset of selection T_S. All ancestral variation at tightly linked sites is eliminated. Recombination leads to low-frequency and high-frequency derived variants in flanking regions. (B) For a single-origin soft sweep from the SGV, multiple lines of descent of the beneficial allele reach into the "standing phase" before T_S. Early recombination introduces ancestral haplotypes at intermediate frequencies. (C) The beneficial allele traces back to multiple origins. Each origin introduces an ancestral haplotype, typically at intermediate frequency.

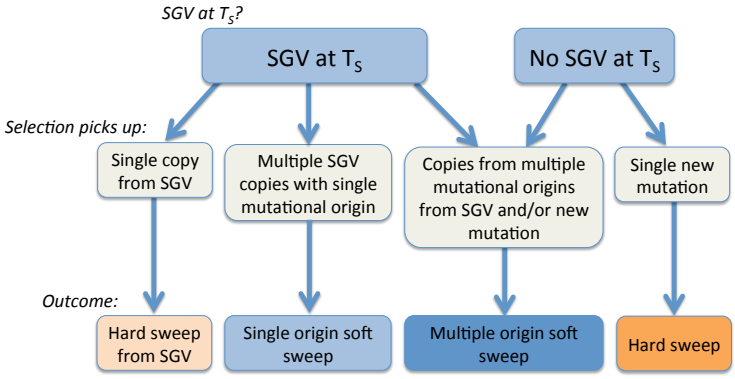

Figure 2: Possible sweeps types depending on whether SGV is available at the onset of selection (T_S), and depending on how many copies or mutational origins of the beneficial allele are picked up by selection.

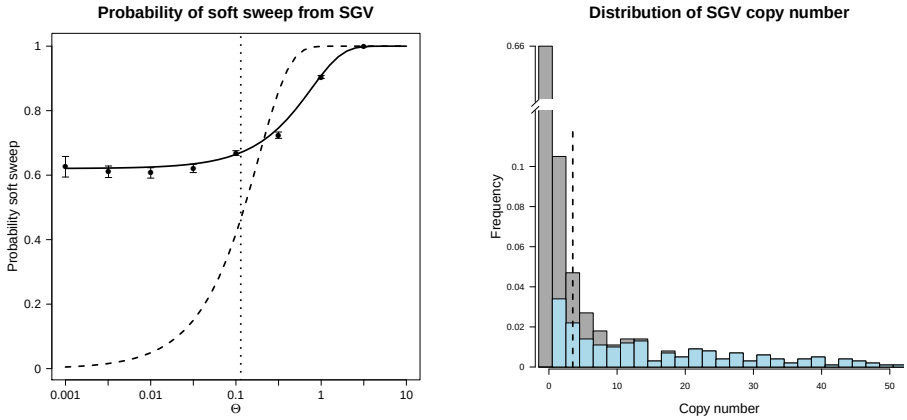

Figure 3: The probability that a sweep from SGV is soft. Panel A compares the deterministic (Eq. 6, dashed) and the stochastic approximation (Eq. 7, solid) to simulation results (dots with standard deviation). The deterministic approximation predicts that hard sweeps dominate for $\Theta < 0.114$ (left of the vertical line). Panel B shows a simulated distribution for the number of copies of the beneficial allele in the SGV (grey) and the fraction that leads to a successful sweep (blue). The vertical dashed line marks the value $N_e u/s_d = \Theta/4s_d$ assumed in the deterministic approximation. Parameters: $s_b = 0.1$, $s_d = 0.01$, $\Theta = 0.1$ (in panel B).

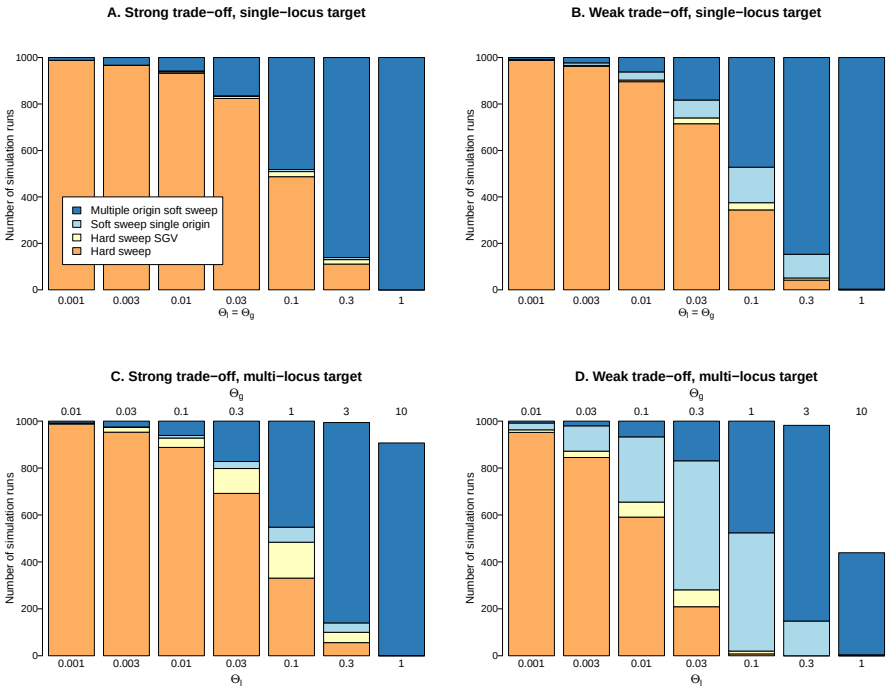

Figure 4: The probability of hard and soft sweeps for beneficial alleles with single-locus mutation target (top row) and for alleles with a target of 10 identical loci (bottom row). $2N_e = 10000$ and $s_b = 0.1$. Panels A and C assume a strong fitness trade-off $s_d = s_b = 0.1$. Panels on the right side assume a weak trade-off $s_d = 0.001$. See the Methods section for further details on the simulations.