File S1: Supporting Information

I. NEUTRAL HAPLOTYPE SPECTRUM

The haplotype spectrum expected in a haploid neutral Fisher-Wright model without recombination can be calculated from the Ewens sampling formula (Ewens, 1972). Ewens showed that the probability of a sample of size n is

$$P_n(a_1, a_2, \ldots, a_n) = \frac{n!}{\Theta_{(n)}} \prod_{m=1}^{n} \left(\frac{\Theta}{m}\right)^{a_m} \frac{1}{a_m!} \ , \tag{1}$$

where a_j is the number of allele classes that are sampled j times and $\Theta_{(k)} = \Theta(\Theta+1)\cdots(\Theta+k-1)$ with $\Theta = 2Nu$. The expectation of a_k is therefore given by

$$\begin{aligned}
\langle a_k \rangle &= \sum_{\mathbf{a}} \frac{\Theta}{k} \left(\frac{\Theta}{k}\right)^{a_k-1} \frac{1}{(a_k-1)!} \frac{n!}{\Theta_{(n)}} \prod_{m\neq k} \left(\frac{\Theta}{m}\right)^{a_m} \frac{1}{a_m!} \\
&= \sum_{\mathbf{a}} \frac{\Theta\Theta_{(n-k)} n!}{k\Theta_{(n)}(n-k)!} P_{n-k}(a_1, a_2, \ldots, a_n) = \frac{\Theta\Theta_{(n-k)} n!}{k\Theta_{(n)}(n-k)!} \\
&= \frac{\Theta}{k} \prod_{m=n-k+1}^{n} \frac{m}{\Theta+m-1} = \frac{\Theta}{k} \prod_{m=n-k+1}^{n} \left(1 + \frac{\Theta-1}{m}\right)^{-1} \approx \frac{\Theta}{k} e^{-(\Theta-1)k/n} \approx \frac{\Theta}{k}
\end{aligned} \tag{2}$$

where the last two approximate inequalities are accurate if $k \ll n$ and $k\Theta \ll n$, respectively. Hence the expected number i_c of allele classes with more than n_c members is roughly $\Theta \sum_{k=n_c}^{n\Theta^{-1}} k^{-1} \approx \Theta(\log n - \log \Theta n_c)$, where cutting off the sum at $k = n\Theta^{-1}$ approximately accounts for the exponential. With this approximation, the i_cth abundant allele class is expected to contain

$$n_c \approx \frac{n}{\Theta} \exp(-i_c/\Theta) \tag{3}$$

copies of the allele. A more accurate expression of the spectrum is obtained by determining the n_c such that $i_c = \sum_{k>n_c} \langle a_k \rangle$, using the exact expression given above. This numerical solution for the haplotype spectrum is plotted in Figure 2B of the main text.

II. THE DISTRIBUTION OF HAPLOTYPE FREQUENCIES

In the main text, we calculated the distribution of the establishment time of the ith haplotype and the frequency of the corresponding haplotype. Here, we show how the joint distribution of all seeding times and the resulting frequency spectrum can be calculated assuming that the novel haplotypes are rare and evolve independently, which is justified if they constitute a small share of the total population, i.e., if $u/s \ll 1$. In this case, the probability that k haplotypes $i = 1, \ldots, k$ are present in frequencies x_i is given by

$$P(x_1, \ldots, x_k|t) = \int_0^t \prod_i dt_i \prod_i P(x_i|t_i, t) P(t_1 \ldots, t_k|t) \ , \tag{4}$$

where $P(x_i|t_i, t)$ is the probability that a haplotype has frequency x_i at time t given it became established at time t_i. The distribution of establishment times $P(t_1 \ldots, t_k|t)$ is given by

$$P(t_1 \ldots, t_k|t) = \frac{1}{k!} e^{-\int_0^t dt' \alpha(t')} \prod_i \alpha(t_i) \ , \tag{5}$$

where $\alpha(t') = 2suNx(t')$ is the rate of establishing novel adaptive haplotypes (main text Equation (1) and below). Note that the t_i defined in Equation (5) are not ordered. They are distributed according to a Poisson point process with density $\alpha(t')$. Assuming that established novel haplotypes increase in frequency logistically according to Equation (5) of the main text, we have

$$P(x_i|t_i, t) = \delta\left(x_i - \frac{e^{(s-u)(t-t_i)}}{2Ns + e^{st}}\right) \ , \tag{6}$$

where $\delta(x)$ is the Dirac δ-function (the stochastic analog is calculated below, see also (Desai and Fisher, 2007)). Substituting $P(x_i|t_i,t)$ into Equation (4) and integrating over t_i, we obtain

$$P(x_1,\ldots,x_k|t) = \frac{1}{k!}e^{-\int_0^t dt'\alpha(t')}\prod_i \frac{1}{(s-u)x_i}\alpha(t_i) \tag{7}$$

with $t_i = t - (s-u)^{-1}\log(x_i(2Ns - e^{st}))$. Haplotypes that are common after the sweep are most likely seeded early during the sweep. Furthermore, we showed in Equation (5) of the main text that their relative frequencies stay approximately constant during the amplification phase. Hence we can determine the joint distribution of frequencies at early times $t \ll s^{-1}\log 2Ns$ while $\alpha(t) \approx 2use^{st}$ is still exponential. After substituting the t_i and simplifying, we find

$$P(x_1,\ldots,x_k|t) \approx \frac{e^{-\frac{u}{s}e^{st}}}{k!}\prod_i \frac{2use^{st_i}}{(s-u)x_i} \sim \prod_i x_i^{-2-\frac{u}{s}} \ , \tag{8}$$

where we dropped factors independent of x_i which ensure normalization. A very similar result was found in (Desai and Fisher, 2007). At large t, the form of the prefactor $e^{-\frac{u}{s}e^{st}}$ changes due to the saturation of the allele frequency at 1, but the distribution of the frequencies of the haplotypes that were seeded early during the sweep remains of this form until the spectrum is eroded by genetic drift.

The haplotype spectrum therefore decays with a power $2+u/s$, which is consistent with the power $1-u/s$ obtained for the cumulative or rank spectrum (integrating $x_i^{-2-u/s}$ yields $x_i^{-1-u/s}$). More importantly, this result tells us that the distribution of haplotype frequencies conditional on the number of haplotypes observed is approximately independent of u/s if $u \ll s$. Hence, given that a sweep occurred, all information about the strength of the sweep is contained in the number of haplotypes and the precise values of their frequencies do not contain any additional information if $u \ll s$. However, whenever there are deviations from the assumptions made here, the haplotype frequencies will contain additional information.

III. STOCHASTIC DERIVATION OF THE HAPLOTYPE SPECTRUM

The dynamics of rare haplotypes are strongly influenced by random genetic drift and we have to ascertain the deterministic arguments made in the main text by a more careful stochastic calculation. While hard in general, an approximate analytic calculation of the frequency spectrum of rare haplotypes is feasible in our case for the following reasons: (i) The dynamics of a beneficial allele are essentially deterministic since it is much more frequent than haplotypes that arise through secondary mutations. (ii) The dynamics of rare haplotypes can be described by a linear branching process since they are always a small fraction of the total population.

As already done in Equation (4), we decompose the distribution of haplotype frequencies into the distribution $P(t_1,\ldots,t_k|t)$ of times when the novel haplotypes arise and probability $P(n,t|t_0)$ that a haplotype is present in n copies at time t, given it arose at time t_0. We will derive $P(n,t|t_0)$ first and consider the spectrum due to the superposition of several independent seeding events below.

A. Distribution of rare variants arising in a logistic sweep

To model the stochastic dynamics of rare haplotypes, we use a continuous time branching process in which individuals produce identical copies of themselves with rate $1+g(t)$ and die with rate 1, i.e., the unit of time is chosen to be the generation time. The average number of offspring of a given individual in this model is $1+g(t)$. Hence, $g(t)$ is the growth rate of the haplotype carrying the beneficial allele. In the case of a sweep, we have

$$g(t) = s(1-x(t)) - u \ , \tag{9}$$

where the first term accounts for selection ($x(t)$ is the frequency of the beneficial allele) and the second term accounts for mutations that change the state of the haplotype. The dynamics of $P(n,t|t_0)$ are described by the forward Master equation

$$\partial_t P(n,t|t_0) = (1+g(t))(n-1)P(n-1,t|t_0) + (n+1)P(n+1,t|t_0) - (2+g(t))nP(n,t|t_0) \ , \tag{10}$$

which accounts for replication (first term) and death (second term). To solve for $P(n,t|t_0)$, it is useful to consider the generating function $G(\lambda,t|t_0) = \sum \lambda^n P(n,t|t_0)$, which obeys the equation

$$\partial_t G(\lambda,t|t_0) = \left[-(2+g(t))\lambda + (1+g(t))\lambda^2 + 1\right]\partial_\lambda G(\lambda,t|t_0) \tag{11}$$

with initial condition $G(\lambda, t_0|t_0) = \lambda$. This equation can be solved via the method of characteristics, with the result

$$G(\lambda, t|t_0) \approx 1 - \frac{1-\lambda}{e^{-\int_{t_0}^{t} dt' g(t')} + (1-\lambda)\int_{t_0}^{t} dt' e^{-\int_{t_0}^{t'} dt'' g(t'')}} \ , \tag{12}$$

where we have used $1+g(t) \approx 1$ along the way. The latter is a good approximation if selection is weak in one generation and amounts to neglecting terms of order s^2. We will now substitute the explicit expression for $g(t)$, where it will be convenient to parametrize the frequency of the beneficial allele as $x(t) = (1+e^{s(\tau-t)})^{-1}$ with $\tau = s^{-1}\log 2Ns$. Using this form of $g(t)$, we find for the generating function

$$G(\lambda, t|t_0) = 1 - \frac{\tilde{s}(1-\lambda)(1+e^{-s(t_0-\tau)})e^{-u(t-t_0)}}{\tilde{s} + \tilde{s}e^{-s(t-\tau)} + (1-\lambda)\left[e^{-s(t_0-\tau)-u(t-t_0)} - e^{-s(t-\tau)} + \tilde{s}u^{-1}(1-e^{-u(t-t_0)})\right]} \tag{13}$$

where $\tilde{s} = s - u$. Any haplotype that is abundant enough to be sampled with high probability most likely originated in the early phase of the sweep ($t_0 \ll \tau$), which allows for the approximation $1+e^{-s(t_0-\tau)} \approx e^{-s(t_0-\tau)}(1+\mathcal{O}(n/(sN)))$ where n is the sample size. Furthermore, we will typically observe the spectrum at times $t \gg \tau$ when the sweep is almost complete. Hence we can approximate $1+e^{-s(t-\tau)} \approx 1+x_{WT} \approx 1$ where x_{WT} is the frequency of the deleterious wild type allele at the time of sampling. Using these simplifications, we obtain

$$G(\lambda, t|t_0) \approx 1 - \frac{(1-\lambda)e^{-s(t_0-\tau)-u(t-t_0)}}{1+(1-\lambda)(\tilde{s}^{-1}e^{-s(t_0-\tau)-u(t-t_0)} + u^{-1}(1-e^{-u(t-t_0)}))} \tag{14}$$

This expression is straightforwardly expanded into a geometric series in λ whose coefficients are $P(n, t|t_0)$. For large n, one finds

$$P(n, t|t_0) \approx \frac{e^{-s(t_0-\tau)-u(t-t_0)}}{\hat{n}^2} e^{-n/\hat{n}} \quad \text{where} \quad \hat{n} = \frac{e^{-s(t_0-\tau)-u(t-t_0)}}{s-u} + \frac{1-e^{-u(t-t_0)}}{u} \ , \tag{15}$$

with relative corrections being on the order of $\hat{n}^{-1}$. The quantity $\hat{n}$ is the mean copy number of the haplotype conditional on non-extinction, and the two terms contributing to $\hat{n}$ have a straightforward interpretation: The first term is the contribution of selection, which amplifies the haplotype before the fixation of the beneficial allele. The second term is the contribution of random genetic drift, which evaluates simply to $t-t_0$ in the limit of $u(t-t_0) \ll 1$. The latter is the analog of the well known fact that a non-extinct neutral allele in a neutral Moran process is on average present in n copies after n generations. The expression for $\hat{n}$ exhibits a crossover from an early regime where selection dominates $\hat{n}$ to a random drift dominated regime at large t. In the limit $u(t-t_0) \ll 1$, we have

$$\hat{n}(t) \approx \begin{cases} (s-u)^{-1}e^{s(\tau-t_0)} & s(t-t_0) \ll e^{s(\tau-t_0)} \\ (t-t_0) & s(t-t_0) \gg e^{s(\tau-t_0)} \end{cases} \tag{16}$$

This crossover will inform us below about how long the contribution of random drift can be neglected when applying our estimator of the strength of the selective sweep.

B. The haplotype frequency spectrum

Having calculated the copy number distribution of a haplotype that originated at time t_0, we now have to determine the distribution of seeding times and calculate the resulting spectrum of haplotype frequencies. New haplotypes that contain the beneficial allele are produced at rate

$$\gamma(t) = Nux(t) = \frac{Nu}{1+e^{-s(t-\tau)}} \ , \tag{17}$$

where, as before, $x(t)$ is the frequency of the sweeping allele. Note that this differs from the rate of establishment of novel variants by a factor s, which will reemerge from the stochastic calculation. The deterministic approximation for $\gamma(t)$ is valid if it is unlikely that new variants are seeded before establishment of the founding variant, which requires $s \gg u$ (see Desai and Fisher (2007)). Since novel haplotypes are seeded and evolve independently to a good approximation, the number of haplotypes present in n copies at time t is Poisson distributed with mean

$$Q(n, t) = \int_0^t dt_0 \ \gamma(t_0) P(n, t|t_0) \ , \tag{18}$$

Due to the exponential nature of $P(n,t|t_0)$, it is convenient to sum $Q(n,t)$ over $n > n_c$ and calculate the expected number of haplotypes with copy numbers greater than n_c. The sum is well approximated by the integral $W(n_c,t) = \int_{n_c} dn Q(n,t)$, and we have

$$\begin{aligned} W(n_c,t) &= Nu \int_{n_c}^{\infty} dn \int_0^t dt_0 x(t_0) P(n,t|t_0) \\ &= Nu \int_0^t dt_0 \frac{e^{-s(t_0-\tau)-u(t-t_0)}}{\hat{n}(1+e^{-s(t_0-\tau)})} e^{-n_c/\hat{n}} \approx Nu \int_0^t dt_0 \frac{e^{-u(t-t_0)}}{\hat{n}} e^{-n_c/\hat{n}} \ , \end{aligned} \tag{19}$$

where the last approximation assumes that novel haplotypes are seeded while the beneficial allele is still expanding exponentially, i.e., $e^{-s(t_0-\tau)} \gg 1$.

It does not seem possible to evaluate the integral over t_0 in Equation (19) analytically. However, the integral is dominated by contributions from a well defined time interval and can be evaluated perturbatively. For a very large population where drift is negligible, and for $s \gg u$, this integral simplifies to

$$W(n_c,t) \approx Nu \int_0^t dt_0 e^{s(t_0-\tau)-sn_c e^{s(t_0-\tau)}} \ , \tag{20}$$

which is very sharply cutoff for $s(t_0-\tau) \gg -\log sn_c$. Hence we can send the upper integration boundary to infinity without loss of accuracy and evaluate this integral exactly. One finds $W(n_c,t) \approx Nu/(sn_c)$. The contribution to this integral come from a narrow peak of width s^{-1} and height Nue^{-1}/n_c. Genetic drift and mutation predominantly change only this height and width, leaving the shape of the integral approximately invariant. Hence we can evaluate this integral by calculating where the integral peaks and how wide this peak is (Laplace's method).

Including the correction due to drift and finite s/u term corrections, the integrand peaks when $n_c \approx \hat{n}$, which translates into $\tilde{s}(\tau - t^*) \approx \log(s(n_c - t + t^*))$ as opposed to $s(\tau - t^*) \approx \log sn_c$ without the corrections. The second derivative of the logarithm of the integrand is approximately given approximately by

$$\frac{1}{\hat{n}^2}\left(\frac{d\hat{n}}{dt_0}\right)^2 \approx s^2 \frac{(\hat{n}-\Delta t)^2}{\hat{n}^2} \ , \tag{21}$$

where $\Delta t = t - t^*$ is the age of the haplotypes. Hence the peak dominating the integral becomes wider by a factor $\frac{\hat{n}}{\hat{n}-\Delta t}$. With these corrections to the "height" and the "width" of the integrand, we obtain the approximate expression

$$W(n_c,t) \approx \frac{Nue^{-ut}}{s(n_c-\Delta t)} \left(\frac{N}{n_c}\right)^{u/s} \tag{22}$$

for the integral. This expression is accurate as long as $sn_c \gg s(t-\tau) - \log(sn_c)$ and $ut \ll 1$. The age of the haplotype evaluates approximately to $\Delta t = t - \tau + s^{-1}\log(sn_c)$, which is of order τ. The additional factor $(N/n_c)^{u/s}$ accounts for the additional time the older haplotypes have been degraded by mutations, while the $n_c - \Delta t$ accounts for the contribution of drift to the frequency of the haplotypes.

After the sweep, the frequency of the most common haplotype is $x_0 = e^{-ut}$ and the expected number of haplotypes above frequency x_c is given by

$$W(x_c,t) \approx \frac{u}{s}\left[\frac{x_0}{(x_c - \Delta t/N)}\right]^{1+u/s} \ . \tag{23}$$

This result tells us that the mean number of haplotypes with frequencies greater than x_c is approximately linear in u/s and decreases with x_c approximately as x_c^{-1}. Furthermore, the expected number of haplotypes above x_c is increasing with time, since rare haplotypes increase in frequency due to random drift.

Given that we observe i_c haplotypes at frequencies higher then x_c, we can use Equation (23) to estimate s/u:

$$\frac{s}{u} \approx \frac{1}{i_c}\left[\frac{x_0}{(x_c - \Delta t/N)}\right]^{1+i_c x_c/x_0} \ , \tag{24}$$

This equation differs from Equation (9) of the main text by a reduction of x_c due to random drift, which has been ignored in the simple deterministic derivation given in the main text. This reduction allows us to correct for the effects of genetic drift as long as drift is not to strong. Obviously, the correction fails as Δt approaches Nx_c.

Equation (24) informs us about the regimes where the proposed methods to estimate the selection coefficient is likely to work. Random genetic drift will degrade the signature of the sweep for haplotype frequencies smaller than $\Delta t/N$. Since the time needed for completion of the sweep is on the order of $(\log Ns)/s$, we require $Nsx_c > \log Ns$. Since $x_c \approx u/(si_c)$ is itself small, we need $Ns \gg i_c \log Ns$ for the method to work. The breakdown of the method is clearly seen in Figure 3 of the main text once Ns falls below 100.

IV. PRUNING RECOMBINANT HAPLOTYPES

Haplotypes that arise by mutation from the founding haplotype differ at exactly one position from the founding haplotype, while haplotypes that result from a recombination event with a member of the diverse ancestral population typically differ at several positions. Furthermore, haplotypes that are mutants of mutants will differ at two positions from the founding haplotype. In the main text, we argued that one can restrict the haplotype spectrum to those haplotypes that differ only at a single site from the founding haplotype, thereby removing most recombinant haplotypes and mutants of mutants. Here, we show that frequency spectrum of such a restricted set of haplotypes differs slightly from that of all haplotypes.

As before, the frequency of the beneficial allele will typically follow Equation (1) of the main text. The frequency of the founding haplotype, however, will remain below this frequency due to loss through recombination and mutation.

$$x_0(t) = \frac{e^{\tilde{s}t}}{2Ns + e^{st}} \ , \tag{25}$$

where we have abbreviated the initial growth rate of a haplotype by $\tilde{s} = s - u - r$. Mutations on this haplotype establish with rate $\alpha(t) = 2u\tilde{s}Nx_0(t)$. Haplotypes that establish at time t_i then typically follow a frequency trajectory

$$x_i(t) = \frac{e^{\tilde{s}(t-t_i)}}{2N\tilde{s} + e^{st}} \ . \tag{26}$$

The most likely seeding time of the ith haplotype is given by

$$t_i = \frac{1}{\tilde{s}} \log \frac{si}{u} \ . \tag{27}$$

Hence we obtain for the ratios of haplotype frequencies at times when the beneficial allele is near fixation

$$\frac{x_i(t)}{x_0(t)} = e^{-\tilde{s}t_i} = \frac{u}{si} \ . \tag{28}$$

This differs from Equation (7) of the main text in that the ratio is proportional to i^{-1}, rather than $i^{-1+u/s}$. This difference is due to the fact that here haplotypes grow with the same rate as the rate at which they are seeded. In the previous case where all haplotypes are considered, haplotypes grow with rate $s - u - r$, while the seeding rate is proportional to the frequency of the beneficial allele which grows with rate s.

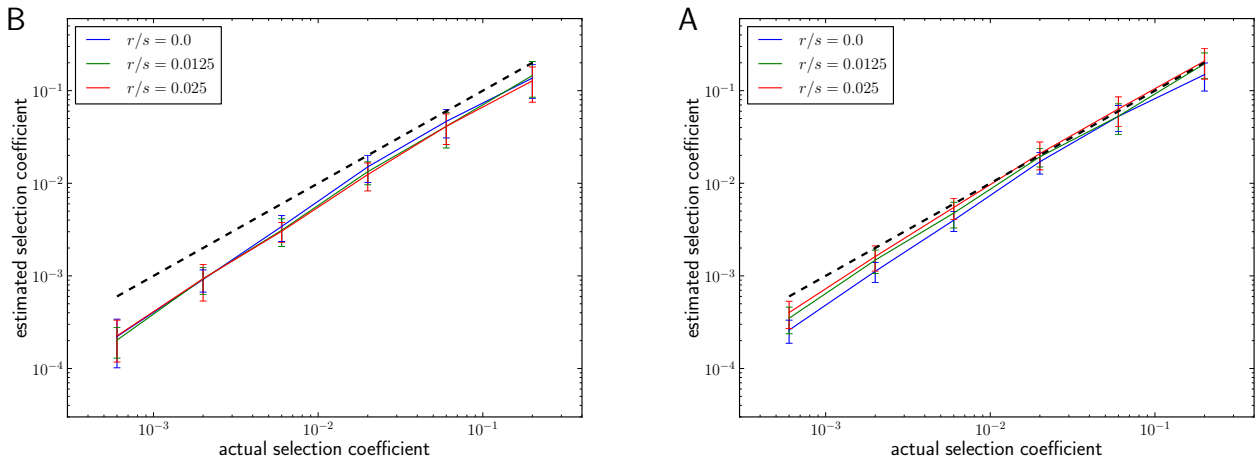

Figure S 1 Estimating selection strength in presence of recombination. This figure is the analog of Fig 4 of the main text with a ten-fold reduced number of segregating sites in the initial samples, i.e., lower ancestral diversity. The estimates are very similar, showing that the ancestral diversity has no impact on the accuracy of the estimation as long as recombination events almost always give rise to unique haplotypes that differ from previous haplotypes at more than 1 site.

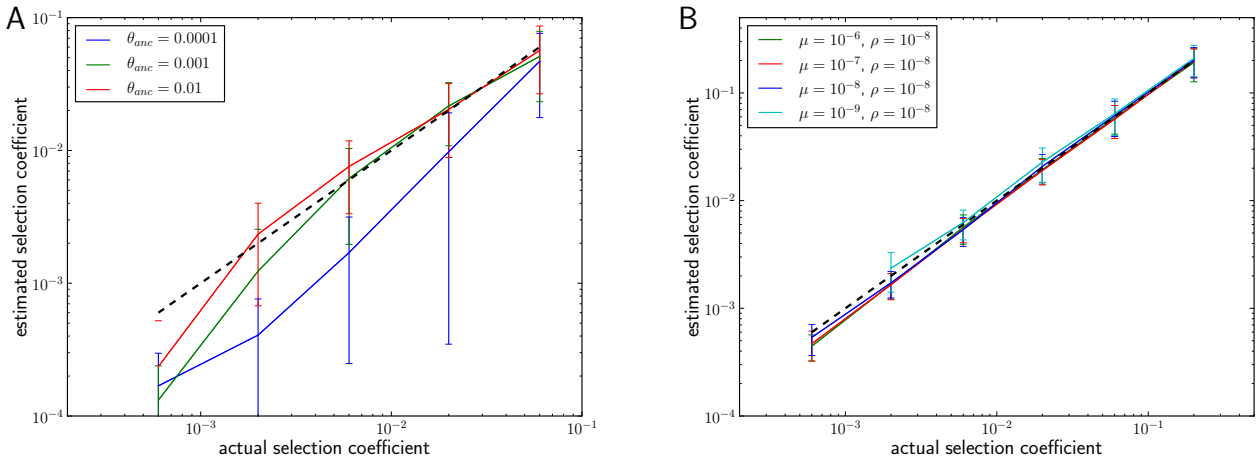

Figure S 2 Comparison of the accuracy of our estimator to the accuracy of estimates obtained from the program `sweepfinder` (Nielsen *et al.*, 2005). Haplotype samples around a selective sweep occurring in the middle of the locus were simulated with the program `msms` (Ewing and Hermisson, 2010). The population size used was $N = 10^5$ and samples were taken when the adaptive allele had reached frequency 0.99 in the population. Panel A: Mean and variance of estimates obtained from the program `sweepfinder` when applied to samples of depth 100 for a locus of size $2s/\rho$ as a function of different levels of ancestral diversity. Note that the length of the simulated locus should provide ample surrounding neutral sequence for `sweepfinder`, given that the dip in diversity is expected to be only of size $0.1s/\rho$. Panel B: Mean and variance of the estimates from our estimator (Equation 9 in the main manuscript) when applied to samples of depth 1000 for a locus of size $0.1s/(\mu + \rho)$ as a function of different mutation rates. Analogously to Figure 3 in the main manuscript we used a cutoff of $i_c = 5$ for the analysis. Recombination rate was always $\rho = 10^{-8}$. `Sweepfinder` performs well only if the ancestral diversity is in the range 0.01 and selection coefficients exceed $s = 0.001$. Our method, in contrast, obtains reliable estimates regardless of the ancestral diversity and also for weaker selection coefficients. Note that the two methods were applied to different data sets (deep population samples of a short locus for our estimator vs. a longer locus at only moderate coverage for `sweepfinder`). The total amount of sequence provided to either method, however, was comparable.

References

Desai, M. M., and D. S. Fisher, 2007 Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics **176**: 1759–98.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theoretical Population Biology **3**: 87–112.

Ewing, G., and J. Hermisson, 2010 Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**: 2064–5.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, *et al.*, 2005 Genomic scans for selective sweeps using snp data. Genome research **15**: 1566–75.