Predicting Carriers of Ongoing Selective Sweeps Without Knowledge of the Favored Allele

Roy Ronen\textsuperscript{1,*}, Glenn Tesler\textsuperscript{2,*}, Ali Akbari\textsuperscript{3,*}, Shay Zakov\textsuperscript{4}, Noah A. Rosenberg\textsuperscript{5}, Vineet Bafna\textsuperscript{4,**}

\textsuperscript{1} Bioinformatics Graduate Program, University of California, San Diego, La Jolla, CA, USA
\textsuperscript{2} Department of Mathematics, University of California, San Diego, La Jolla, CA, USA
\textsuperscript{3} Department of Electrical & Computer Engineering, University of California, San Diego, La Jolla, CA, USA
\textsuperscript{4} Department of Computer Science & Engineering, University of California, San Diego, La Jolla, CA, USA
\textsuperscript{5} Department of Biology, Stanford University, Stanford, CA, USA

* These authors contributed equally to this work.
** vbafna@ucsd.edu
Submitted April 6, 2015.

Abstract

Methods for detecting the genomic signatures of natural selection have been heavily studied, and they have been successful in identifying many selective sweeps. For most of these sweeps, the favored allele remains unknown, making it difficult to distinguish carriers of the sweep from non-carriers. In an ongoing selective sweep, carriers of the favored allele are likely to contain a future most recent common ancestor. Therefore, identifying them may prove useful in predicting the evolutionary trajectory — for example, in contexts involving drug-resistant pathogen strains or cancer subclones. The main contribution of this paper is the development and analysis of a new statistic, the Haplotype Allele Frequency (HAF) score. The HAF score, assigned to individual haplotypes in a sample, naturally captures many of the properties shared by haplotypes carrying a favored allele. We provide a theoretical framework for computing expected HAF scores under different evolutionary scenarios, and we validate the theoretical predictions with simulations. As an application of HAF score computations, we develop an algorithm (PreCIOSS: Predicting Carriers of Ongoing Selective Sweeps) to identify carriers of the favored allele in selective sweeps, and we demonstrate its power on simulations of both hard and soft sweeps, as well as on data from well-known sweeps in human populations.

Author summary

Methods for detecting the genomic signatures of natural selection have been heavily studied, and they have been successful in identifying genomic regions under positive selection. However, methods that detect positive selective sweeps do not typically identify the favored allele, or even the haplotypes carrying the favored allele. The main contribution of this paper is the development and analysis of a new statistic (the
HAF score), assigned to individual haplotypes. Using both theoretical analyses and simulations, we describe how the HAF scores differ for carriers and non-carriers of the favored allele, and how they change dynamically during a selective sweep. We also develop an algorithm, PreCIOSS, for separating carriers and non-carriers. Our tool has broad applicability as carriers of the favored allele are likely to contain a future most recent common ancestor. Therefore, identifying them may prove useful in predicting the evolutionary trajectory — for example, in contexts involving drug-resistant pathogen strains or cancer subclones.

**Introduction**

With genome sequencing, we now have an opportunity to more completely sample genetic diversity in human populations, and probe deeper for signatures of adaptive evolution [1–3]. Genetic data from diverse human populations in recent years have revealed a multitude of genomic regions believed to be evolving under recent positive selection [4–16].

Methods for detecting selective sweeps from DNA sequences have examined a variety of signatures, including patterns represented in variant allele frequencies as well as in haplotype structure. Initially, the problem of detecting selective sweeps was approached primarily by considering variant allele frequencies, exploiting the shift in frequency at ‘hitchhiking’ sites linked to a favored allele relative to non-hitchhiking sites [17, 18]. The site frequency spectrum (SFS) within and across populations is often used as a basis for such inference [4, 6, 19–25]. More recently, methods based on haplotype structure have been developed using a variety of approaches, including the frequency of the most common haplotype [26], the number and diversity of distinct haplotypes [27], the haplotype frequency spectrum [28], and the popular approach of long-range haplotype homozygosity [29–32].

In general, haplotype-based methods seek to characterize the population with summary statistics that capture the frequency and length of different haplotypes. However, the haplotypes are related through a genealogy, and relationships among them are inherently lost in such analyses. In addition, data on the site frequency spectrum can be lost or hidden in analyses focused on haplotype spectra. In this paper, we connect related measures of haplotype frequencies and the site frequency spectrum by merging information describing haplotype relationships with variant allele frequencies. Our main contribution is a statistic that we term the haplotype allele frequency (HAF) score, which captures many of the properties shared by haplotypes carrying a favored allele.

Consider a sample of haplotypes in a genomic region. We assume that all sites are biallelic, and at each site, we denote ancestral alleles by 0 and derived alleles by 1. We also assume that all sites are polymorphic in the sample. The HAF vector of a haplotype $h$, denoted $c$, is obtained by taking the binary haplotype vector and replacing non-zero entries (derived alleles carried by the haplotype) with their respective frequencies in the sample (Fig. 1A). For parameter $\ell$, we define the $\ell$-HAF score of $c$ as:

$$\ell\text{-HAF}(c) = \sum_j c_j^\ell$$

where the sum proceeds over all segregating sites $j$ in the genomic region. The 1-HAF score of a haplotype amounts to the sum of frequencies of all derived alleles carried by the haplotype. The $\ell$-HAF score is equivalent to the $\ell$-norm of $c$ raised to the $\ell$th power, or $\lVert c \rVert_\ell^\ell$. We will show that during a selective sweep, the HAF score of a haplotype serves as a proxy to its relative fitness.
Figure 1. The HAF score. Genealogies of three samples (n = 6) progressing through a selective sweep, from left to right. Neutral mutations are shown as red circles, and are numbered in red; the favored allele is shown as a red star. The 1-HAF score of each haplotype is shown below its corresponding leaf, in black. For the rightmost haplotype in (A), the binary haplotype vector $h$ is shown along with its HAF-vector $c$, and 1-HAF and 2-HAF scores. Vector $w^{\text{all}}$ lists the frequencies of all mutations. (A) The favored allele appears on a single haplotype. At this point in time, both the genealogy and the HAF score distribution are largely neutral. Coalescence times ($T_2, \ldots, T_6$) are shown on the left, where $T_k$ spans the epoch with exactly $k$ lineages. (B) Carriers of the favored allele are distinguished by high HAF scores (in large part due to the long branch of high-frequency hitchhiking variation); non-carriers have low HAF scores. (C) After fixation, there is a sharp loss of diversity causing low HAF scores across the sample.
Selective sweeps. The classical model for selection, and the one that has received most attention, is the “hard sweep” model, in which a single mutation conveys higher fitness immediately upon occurrence, and rapidly rises in frequency, eventually reaching fixation. Under this model, we can partition the haplotypes into carriers of the favored allele, and non-carriers. In the absence of recombination, the favored haplotypes form a single clade in the genealogy. As a sweep progresses, HAF scores in the favored clade will rise due to the increasing frequencies of alleles hitchhiking along with the favored allele. HAF scores of non-carrier haplotypes will decrease, as many of the derived alleles they carry become rare (Fig. 1B). After fixation of the favored and hitchhiking alleles, HAF scores will decline sharply (Fig. 1C), as the selected site and other linked sites are no longer polymorphic. Thus, this reduction in the HAF score results from the sudden loss of many high-frequency derived alleles from the pool of segregating sites. Finally, as the site-frequency spectrum recovers to its neutral state due to new mutations and drift, so will the HAF scores.

Recombination is a source of ‘noise’ for the properties of the HAF score, predicted under the assumption of a hard sweep and no recombination, as it allows haplotypes to cross into and out of the favored clade. Recombination can lead to (i) haplotypes that carry the favored allele but little of the hitchhiking variation, thus having relatively low HAF scores despite their high fitness, or (ii) haplotypes that do not carry the favored allele but do carry much of the hitchhiking variation, thus having relatively high HAF scores despite their low fitness. By the same logic, recombination adds ‘noise’ after fixation by making the otherwise sharp decline in HAF scores more subtle and gradual. This more gradual decline occurs due to recombination weakening the linkage between the favored allele and hitchhiking variants.

Recently, “soft sweeps” have generated significant interest. A soft sweep occurs when multiple sets of hitchhiking alleles in a given region increase in frequency, rather than a single favored haplotype. Soft sweeps may take place by one or more of the following mechanisms: (i) selection from standing variation: a neutral segregating mutation, which exists on several haplotypic backgrounds, becomes favored due to a change in the environment; (ii) recurrent mutation: the favored mutation arises several times on different haplotypic backgrounds; or, (iii) multiple adaptations: multiple favored mutations occur on multiple haplotypic backgrounds. Several methods have been developed for detecting soft sweeps, as well as for distinguishing between soft and hard sweeps. In soft sweeps, multiple sets of hitchhiking alleles rise to intermediate frequencies as the favored allele fixes. This could cause the pre-fixation peak and post-fixation trough in HAF scores to be less pronounced and to occur more gradually compared to a hard sweep.

We find (see Results) that the properties of the HAF score remain robust to many soft sweep scenarios. Moreover, the HAF score could potentially be used to detect soft sweeps. However, in this paper, we focus on the foundations, developing theoretical analysis and empirical work that predicts the dynamics of the HAF score. We also develop a single application. Recall that most existing methods for characterizing selective sweeps focus on identifying regions under selection. Here, given a region already identified to be undergoing a selective sweep, we ask if we can accurately predict which haplotypes carry the favored allele, without knowledge of the favored site. Successfully doing so may provide insight into the future evolutionary trajectory of a population. Haplotypes in future generations are more likely to be descended from, and therefore to resemble, extant carriers of a favored allele. This predictive perspective is of particular importance when a sweep is undesirable and measures may be taken to prevent it. For instance, consider a set of tumor haplotypes isolated from single cells, some of which are drug-resistant and therefore favored under drug exposure. Given a genetic sample of the tumor haplotypes, the HAF statistic may be applied to identify the resistant haplotypes — carriers of a
favored allele — before they clonally expand and metastasize.

Below, we start with a theoretical explanation of the behavior of the HAF score under different evolutionary scenarios, validating our results using simulation. We then develop an algorithm (PreCIOSS: Predicting Carriers of Ongoing Selective Sweeps) to detect carriers of selective sweeps based on the HAF score. We demonstrate the power of PreCIOSS on simulations of both hard and soft sweeps, as well as on real genetic data from well-known sweeps in human populations. While our theoretical derivations make use of coalescent theory, and explicitly use tree-like genealogies, we note that HAF scores can be computed for any haplotype matrix including those with recombination events. Our results on simulated and real data imply that the utility of the HAF score extends to cases with recombination as well as other evolutionary scenarios.

Results

Theoretical and empirical modeling of HAF scores

We consider a sample of \( n \) haploid individuals chosen at random from a larger population of size \( N \). Let \( \mu \) denote the mutation rate per generation per nucleotide, and let \( \theta = 2N\mu L \) denote the population-scaled mutation rate in a region of length \( L \) bp. We consider both constant-sized and exponentially growing populations. For exponentially growing populations, let \( N_0 \) denote the final population size, let \( r \) denote the growth rate per generation, and let \( \alpha = 2N_0 r \) the population-scaled growth rate. Let \( \rho \) denote the population-scaled recombination rate. In our theoretical calculations, we assume no recombination \( (\rho = 0) \), and we derive expressions for the general \( \ell \)-HAF score. We use simulations to demonstrate the concordance of theoretical and empirical values of the \( \ell \)-HAF score, and show that the values are robust to the presence of recombination (see ‘Simulations’ in Methods for parameter choices). Although some of our theoretical calculations below derive expressions for the general \( \ell \)-HAF score, we primarily use 1-HAF in the applied sections. Applications of \( \ell \)-HAF with \( \ell > 1 \) will be explored in future work.

Expected \( \ell \)-HAF score under neutrality, constant population size. First, we assume that the genomic region of interest is evolving neutrally, the population size remains constant at \( N \), and that the ancestral states are known or can be derived. In a sample of size \( n \), let \( c(v) \) denote the HAF vector \( c \) for the \( v \)th haplotype \( (v \in \{1, \ldots, n\}) \). Let \( \xi_w \) be the number of sites with derived allele frequency \( w \). We only consider polymorphic sites in the sample, so the frequency is in the range \( w \in \{1, \ldots, n-1\} \); a mutation present in all or none of the haplotypes in the sample would not be detectable. Each of the \( \xi_w \) sites of frequency \( w \) contributes \( w^\ell \) to the \( \ell \)-HAF score of each of the \( w \) haplotypes with the mutation, and contributes \( 0^\ell = 0 \) for each of the other \( n-w \) haplotypes. The mean of the \( \ell \)-HAF scores of all \( n \) haplotypes in the sample is

\[
\frac{1}{n} \sum_{v=1}^{n} \ell \text{-HAF}(c(v)) = \frac{1}{n} \sum_{w=1}^{n-1} \xi_w \cdot w^\ell \cdot w = \frac{1}{n} \sum_{w=1}^{n-1} \xi_w \cdot w^{\ell+1}. \tag{2}
\]

Under the coalescent model, \([42, \text{Eq. (22)}]\) shows that \( E[\xi_w] = \theta/w \) for all \( 1 \leq w \leq n-1 \). By averaging over all haplotypes in all genealogies, the expected \( \ell \)-HAF score is computed as

\[
E[\ell \text{-HAF}] = \frac{1}{n} \sum_{w=1}^{n-1} E[\xi_w] \cdot w^\ell \cdot w = \frac{\theta}{n} \sum_{w=1}^{n-1} w^\ell. \tag{3}
\]
The first two cases ($\ell = 1, 2$) yield

$$\mathbb{E}[1\text{-HAF}] = \frac{\theta(n - 1)}{2}, \quad \mathbb{E}[2\text{-HAF}] = \frac{\theta(n - 1)(2n - 1)}{6}. \quad (4)$$

**Expected $\ell$-HAF score, variable population size.** Our derivation of expected $\ell$-HAF scores for constant, neutrally evolving populations does not immediately extend to other demographic scenarios. We describe a second approach that separates coalescence times from the genealogy, and we apply it to compute the expected $\ell$-HAF in an exponentially growing population.

For a sample of size $n$, partition the time spanning from the present back to the sample MRCA into $n - 1$ epochs. Let epoch $k \in \{2, \ldots, n\}$ be the span of time during which the genealogy contains exactly $k$ lineages (Fig. 1). Note that mutations on a given lineage in a given epoch share the same frequency, as they appear in exactly the same leaves. For example, mutations 2 and 3 in Fig. 1A occur on the same lineage in epoch 2, and they share the frequency 3. Consider the path leading from a randomly chosen sample MRCA into $n$ leaves. For example, mutations 2 and 3 in Fig. 1A occur on the same lineage in epoch 2, and they share the frequency 3. Consider the path leading from a randomly chosen haplotype back to the sample MRCA. We can write the $\ell$-HAF score of the haplotype as

$$\ell\text{-HAF}(c) = \sum_{k=2}^{n} m_k w_k^\ell, \quad (5)$$

where $m_k$ is the number of mutations that occurred on the path in the $k^{th}$ epoch, and $w_k$ is the frequency or weight of those mutations. For a given genealogy with haplotypes $v \in \{1, \ldots, n\}$, let $c(v)$ denote $c$ (the HAF vector) for the $v^{th}$ haplotype. Similarly, let $m_k(v)$ and $w_k(v)$ denote the number of mutations and their frequency in the $k^{th}$ epoch for the $v^{th}$ haplotype. Epoch $k$ splits the haplotypes into $k$ equivalence classes, which we call $k$-clades. Let $m_{k,i}$ and $w_{k,i}$ denote the corresponding values on the $i^{th}$ lineage of the $k^{th}$ epoch. We compute the expected value by summing over all haplotypes and genealogies and dividing by $n$. The sum is

$$\sum_{v=1}^{n} \ell\text{-HAF}(c(v)) = \sum_{k=2}^{n} \sum_{i=1}^{n} m_k(v) w_k(v) \ell^i = \sum_{k=2}^{n} \sum_{i=1}^{k} m_{k,i} w_{k,i} \ell^i = \sum_{k=2}^{n} \sum_{i=1}^{k} m_{k,i} w_{k,i} \ell^{i+1}. \quad (6)$$

Let $M_{k,i}$ and $W_{k,i}$ be random variables denoting the number of mutations, and their frequency respectively, on the $i^{th}$ lineage of the $k^{th}$ epoch. As the genealogy of a neutrally evolving sample is independent of branch lengths $[43]$, $M_{k,i}$ and $W_{k,i}$ are independent random variables. Thus, we can compute the expected $\ell$-HAF score of a randomly chosen haplotype as

$$\mathbb{E}[\ell\text{-HAF}] = \frac{1}{n} \sum_{k=2}^{n} \sum_{i=1}^{k} \mathbb{E}[M_{k,i} W_{k,i}^{\ell+1}] = \frac{1}{n} \sum_{k=2}^{n} \sum_{i=1}^{k} \mathbb{E}[M_{k,i}] \mathbb{E}[W_{k,i}^{\ell+1}]. \quad (7)$$

To compute $\mathbb{E}[W_{k,i}^{\ell}]$, we start with a related quantity. For positive integer $\ell$, denote the rising factorial

$$w^{(\ell)} = w(w + 1)(w + 2) \cdots (w + \ell - 1), \quad (8)$$

and set $w^{(0)} = 1$. We show in the Appendix that

$$\mathbb{E}[(W_{k,i})^{(\ell)}] = \ell! \cdot \frac{n^{(\ell)}}{k^{(\ell)}}. \quad (9)$$
We have \( w^{(1)} = w \) and \( w^{(2)} = w(w + 1) = w^2 + w \), so \( w^2 = w^{(2)} - w^{(1)} \), which leads to:
\[
\mathbb{E}[(W_{k,i})^2] = \mathbb{E}[(W_{k,i})^{(2)} - (W_{k,i})^{(1)}] = \frac{2n(n+1)}{k(k+1)} - \frac{n}{k} = \frac{n(2n-k+1)}{k(k+1)}.
\]

(10)

In the Appendix, we generalize this equation to compute \( \mathbb{E}[(W_{k,i})^\ell] \). In addition, we show that for a constant-sized population, the general form in Eq. (7) produces the same result as Eq. (3).

**Exponential population growth.** Eq. (7) can potentially be used to obtain \( \mathbb{E}[\ell\text{-HAF}] \) under arbitrarily complex demographics. Consider a population of current size \( N_0 \) that has been growing exponentially at a rate \( r \). The population size at time \( t \) in the past is given by \( N(t) = N_0 e^{-rt} \). Exponential population growth is of particular interest, as it has been used to analyze the state of a population under a selective sweep shortly after fixation. This is a low point (or trough) of observed \( \ell\text{-HAF} \) scores, as early hitchhiking sites have fixed by this time point, and the (relatively recent) sample MRCA is a carrier of the favored mutation. Immediately after fixation, the population — all of which are carriers of the favored allele — has been growing for the duration of the sweep at a rate that is approximately exponential (with growth rate related to the selection coefficient \( s \)). In addition, all extant and ancestral haplotypes since the sample MRCA are carriers and therefore equally favored, implying that the independence between \( W_{k,i} \) and \( M_{k,i} \) is kept. While the branch lengths and distribution of \( M_{k,i} \) values change under exponential growth, the distribution for \( W_{k,i} \) remains unchanged as described in Eq. (10). This key insight allows us to use Eq. (7) to estimate the expected HAF scores under exponential population growth.

In order to use \( \mathbb{E}[M_{k,i}] = \mu \mathbb{E}[T_k] \) under exponential growth, we implement two numerical methods to compute \( \mathbb{E}[T_k] \): a ‘cumulative time’ method that uses an approximate distribution of \( T_k \) given in [44, p. 559], and a ‘conditional expectation’ method (see Appendix for details). In the conditional expectation method, we compute the expected value of \( T_k \) conditioned on \( T_{k+1}, \ldots, T_n \), as follows (in the order \( k = n, n-1, \ldots, 2 \)):
\[
t_k = \mathbb{E}[T_k \mid T_{k+1} = t_{k+1}, \ldots, T_n = t_n]
\]
\[
= \frac{1}{r} \int_0^1 \ln \left( 1 - \frac{\alpha}{k(k-1)} e^{-r\tau_k} \right) \frac{\ln(u)}{u} du
\]
\[
= \frac{1}{r} \exp \left( \frac{k(k-1)}{\alpha} e^{r\tau_k} \right) E_1 \left( \frac{k(k-1)}{\alpha} e^{r\tau_k} \right),
\]

(11)

where \( \alpha = 2 N_0 r \) is the scaled growth rate, \( \tau_k = t_{k+1} + \cdots + t_n \) (with \( t_n = 0 \)), and \( E_1(x) \) is the exponential integral \( E_1(x) = \int_1^\infty e^{-x t} dt \).

We then use \( \mathbb{E}[(W_{k,i})^\ell] \) (evaluated in Eq. (S21)) to evaluate Eq. (7), yielding \( \mathbb{E}[\ell\text{-HAF}] \) for exponential population growth as
\[
\mathbb{E}[\ell\text{-HAF}] \approx \frac{\theta}{\alpha} \sum_{k=2}^n \left( r \cdot t_k \right) \sum_{q=0}^{\ell} (-1)^{\ell-q} S(\ell, q) \left( (q+1)! \left( \frac{(n+1)(q)}{(k+1)(q-1)} - q \cdot q! \left( \frac{(n+1)(q-1)}{(k+1)(q-1)} \right) \right) \right),
\]

(12)

where \( S(\ell, q) \) denotes the Stirling number of the second kind [45, Ch. 6.1]. We describe these procedures fully in the Appendix.

**Empirical validation of expected HAF score computation.** We tested our theoretical calculations against empirical observations of HAF scores using simulations for neutral evolution with constant population size \( N = 2000 \) (lowered for subsequent
forward simulations; see ‘Simulations’ in Methods). For example, for \( \theta = 48 \) and \( n = 200 \), the expected 1-HAF is exactly 4776.0 (Eqs. (3),(7)), whereas the empirically observed mean 1-HAF score of 20,000 simulated samples is 4777.1 (Supplementary Fig. S1).

We also modeled exponential growth in population size using the scaled growth rate \( \alpha \), using the conditional expectation method in Eq. (12) (see Appendix). As expected, the HAF score is much lower than for constant population size. For \( \alpha = 80 \), \( \theta = 48 \), \( n = 200 \), the theoretical mean 1-HAF score is 126.9, whereas the empirical mean of 20,000 simulations is 128.1 (Fig. S2).

We compared the simulations with theoretical expected \( \ell \)-HAF scores for multiple values of \( \ell \in \{1, 2, 3, 4\} \), and different choices of the population-genetic parameters \((n \in \{100, 200, 300, 400\}, \text{scaled mutation rate } \theta \in \{24, 48\}, \text{scaled recombination rate } \rho \in \{0, 7.568, 15\}, \text{and scaled growth rate } \alpha \in \{0, 30, 60, 80\}\). Note that theoretical expected values were computed assuming scaled recombination rate \( \rho = 0 \), while simulations included \( \rho \in \{0, 7.568, 15\} \). See Methods for a detailed description of the choice of simulation parameters. Fig. S3 shows the concordance between theoretical and empirical means for each choice of parameters. The concordance improves slightly for increasing values of \( n, \ell \). Surprisingly, the values are robust to choice of \( \rho \), and the variance even reduces slightly for higher \( \rho \) (Supplementary Fig. S3).

**HAF score dynamics in selective sweeps**

We now consider the dynamics of HAF scores in a population undergoing a selective sweep. To do this, we use data simulated under several scenarios including both hard and soft sweeps. Though soft sweeps can arise under different circumstances, in what follows we restrict our attention to soft sweeps from the standing variation. Fig. 2 illustrates the HAF score dynamics in a single simulated population undergoing a soft sweep, due to standing variation, with selection coefficient \( s = 0.05 \) and favored allele frequency \( v_0 = 0.3 \) when selection begins. See ‘Simulations’ in Methods for a detailed description of the simulation parameters. Initially (leftmost, time 0) the HAF scores of carriers and non-carriers of the favored allele are similar. As the sweep progresses (times 50–200), carrier HAF scores increase to a peak value (HAF-peak). Soon after fixation (time \( \sim 400 \)), we observe a sharp decline in HAF scores (HAF-trough), followed by slow and steady recovery due to new mutation and drift (times 500–4000).

Below, we provide a theoretical description of these dynamics, as well as empirical validation using simulations. This allows us to predict HAF scores in (a) the post-fixation trough; (b) the pre-fixation peak; and (c) the rate of growth of HAF scores from pre-sweep to peak value.

**Empirical validation of the post-fixation HAF-trough.** We showed using simulations that the HAF score computations for an exponentially growing population (Eq. (12)) also approximate a population evolving under a selective sweep *shortly after fixation*. This enables prediction of the HAF-trough value.

The HAF-trough of a sweep is the value of 1-HAF at fixation. We took the mean of the HAF-trough values over 200 populations simulated under selective sweeps with coefficients \( s \in \{0.02, 0.08\} \) (see ‘Simulations’ in Methods), and compared it to 1-HAF values in simulated neutral populations growing exponentially at rates \( \alpha \in \{32, 101\} \). Fig. 3A shows a close similarity between the 1-HAF values under exponential growth (blue) and the selective sweep trough (red). The resemblance is closer for higher values of \( s \) and \( \alpha \). To some extent, this may stem from the increased simulation variance we see for lower values of \( s \). On a deeper level, selective sweeps with fixed population size demonstrate a logistic curve [46 Sec. 6.1.3]. However, for higher values of \( s \) and \( \alpha \), the concordance between selective sweeps and exponential growth is increased.
Figure 2. HAF scores in a soft sweep. We consider HAF scores in 50 kb segments, examining \( n = 200 \) haplotypes sampled from a constant-sized \( (N = 2000) \) population, evolving with population-scaled mutation rate \( \theta = 48 \) and selection coefficient \( s = 0.05 \). The soft sweep is modeled by the frequency of the favored allele at the onset of selection, which we set to \( \nu_0 = 0.3 \). We do forward simulations, with time \( t = 0 \) at the onset of selection and \( t \) increasing towards the present time. Snapshots of generations are shown at specific times indicated at tick marks on the \( x \)-axis. Note that these times are increasing but neither consecutive nor regularly spaced. Each selected generation is depicted as a tall thin rectangle. A few rectangles are shown for each phase of a simulated population undergoing a selective sweep. Each point within a rectangle represents the 1-HAF score of a randomly chosen haplotype. Red points represent carriers of the favored allele and blue points represent non-carriers. Points are scattered randomly on the \( x \)-axis within each rectangle, but all points within the same rectangle represent the same generation at the time indicated by the tick mark on the \( x \)-axis, regardless of their horizontal position within the rectangle. (A) Simulation of a non-recombining segment. (B) Simulation with population-scaled recombination rate \( \rho = 7.568 \); see Eq. \( \{17\} \).
Figure 3. HAF scores in a selective sweep ‘peak’ and ‘trough’. (A) Observed values (red) of the mean ‘trough’ 1-HAF scores in simulated selective sweeps with coefficients \( s \in [0.02, 0.08] \) and theoretical values (blue) of expected 1-HAF scores under exponential population growth with population-scaled rates \( \alpha \in [32, 101] \) given by Eq. \([12]\). Simulated 1-HAF scores (red) represent the mean of 200 simulated population samples for each value of \( s \), with \( \theta = 48, n = 200 \). (B) Observed mean 1-HAF peak, trough, and difference (peak−trough) for selective sweeps with coefficients \( s \in [0.02, 0.08] \). The dashed line represents the approximate value of the peak 1-HAF score given by Eq. \([15]\). (C) Dynamics of the expected value of 1-HAF\textsuperscript{car} (1-HAF score of haplotypes carrying the favored allele) plotted as a function of the fraction of carriers (\( \nu \)) in the sample during a selective sweep. For each \((\theta, n, \nu)\) with \( \theta \in \{24, 48\} \), \( n \in \{100, 200, 300, 400\} \), \( \nu \in \left\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\right\} \), \( s = 0.08 \), and \( N = 2000 \), we plotted the mean value of \( (1-HAF^{\text{car}})/(\theta n) \) over 1500 trials, and compared against the theoretical values (Eq. \([13]\)).
The pre-fixation 1-HAF-peak. As the selective sweep progresses, the value of the HAF score of haplotypes carrying the favored allele increases, eventually reaching a peak value. Consider $n$ haplotypes sampled from a fixed population of $N$ haploid individuals under a selective sweep. Let $\mu$ denote the mutation rate per base per generation in the genomic region of interest, and assume that there is no recombination. The scaled mutation rate is given by $\theta = 2N\mu$.

We let $\nu$ denote the fraction of carrier haplotypes in the sample. When $\nu \leq 1/n$ (i.e., 0 or 1 carriers), there is no selection going back in time, and the time to MRCA can be computed using the neutral Wright-Fisher model [47]. The expected 1-HAF scores for carriers and non-carriers are identical (Eq. (3)). At the time when $\nu$ first equals 1, there are no non-carriers, and the HAF-scores are given by the exponential growth model. In the Appendix, we model the 1-HAF scores for all intermediate values of $\nu$.

Let $1\text{-HAF}^\text{car}$ (respectively, $1\text{-HAF}^\text{non}$) denote the 1-HAF score of a random haplotype carrying the favored allele (respectively, a non-carrier) when a fraction $\nu$ of the $n$ sampled haplotypes carry the favored allele. In the Appendix, we show that under strong selection ($Ns \gg 1$) and no recombination ($\rho = 0$),

$$
E[1\text{-HAF}^\text{car}] \approx \theta n \left( \frac{\nu + 1}{2} - \frac{1}{(1 - \nu)n + 1} \right),
$$

\hspace{1cm} (13)

$$
E[1\text{-HAF}^\text{non}] \approx \theta n \left( \frac{1}{2} + \frac{1}{2n} - \frac{1}{(1 - \nu)n + 1} \right).
$$

\hspace{1cm} (14)

For any sample of size $n$, the carrier haplotypes reach a peak value of $1\text{-HAF}^\text{car}$ as $\nu$ varies along its trajectory. We do not compute the expected value of this peak ($E[\max_\nu(1\text{-HAF}^\text{car}(\nu, n))]$) directly. Instead, we compute the peak value of $E[1\text{-HAF}^\text{car}(\nu, n)]$ (maximizing over all $\nu \in [0, 1]$) as

$$
\max_\nu E[1\text{-HAF}^\text{car}(\nu, n)] = \theta n \left( 1 - \frac{1}{\sqrt{2n}} \right)^2 \approx \theta n.
$$

\hspace{1cm} (15)

Note that under strong selection, this peak does not depend on $s$ (see Fig. 3B).

The trough for each trajectory is computed as the 1-HAF score at fixation (when $\nu = 1$ is first reached).

Empirical validation. Simulated data under selective sweeps with coefficients $s \in [0.02, 0.08]$ show that (i) the pre-fixation HAF peak scores appear to be independent of the selection coefficient (Fig. 3B), and (ii) as predicted by Eq. 15, the mean value of the HAF peak score is approximately $\theta n$ (Fig. 10 in Appendix). We also simulated $(1\text{-HAF}^\text{car})/(n\theta)$ as a function of $\nu$ (Fig. 3C and Fig. S9). The results show a tight correspondence between theory and empirical observations.

HAF score Application: Characterizing carriers and non-carriers.

Our understanding of the dynamics of HAF scores of a haplotype during a selective sweep has many potential applications. For example, we could compare the dynamics of hard and soft sweeps to distinguish between the two events. Second, conditioning on known or deduced selective sweeps in a population sample, we can predict the state (carrier/non-carrier) of the favored allele in its haplotypes. Below we explore the latter, leaving the former to future work.

In Fig. 4A, we show the distributions of haplotype 1-HAF scores aggregated from 500 simulated populations undergoing a hard selective sweep (see ‘Simulations’ in Methods for detailed parameter choices). Scores were computed for random samples of $n = 200$ haplotypes taken at regular time intervals. They are stratified by the frequency of the
favored allele at the time of sampling. Further, scores are stratified into carrier and non-carrier classes (of the favored allele). As with a single population, HAF scores of carriers and non-carriers diverge as the sweep progresses in frequency. We note, however, that even close to fixation (frequencies 80–100%) the distributions of HAF scores between carriers and non-carriers maintain considerable overlap. The high variance in HAF scores makes them only weakly informative of sweep carrier status when comparing across population samples (or genomic regions within a single population). Within a single population sample, however, the HAF scores are highly informative of the carrier status. This is illustrated in Fig. 4B, showing the distributions of HAF score percentile rank within their respective samples. We observe that the rank distributions have minimal overlap for carriers and non-carriers of the favored allele. Any remaining overlap in the percentile rank distributions in the final stages of a sweep (favored allele frequency ≥70%) stems mostly from recombination, which allows the favored allele to recombine onto haplotypes outside the selected clade (creating low HAF score carriers) and vice-versa (creating high HAF score non-carriers). The overall strong separation between carriers and non-carriers is further illustrated by the highly significant \( P \)-values of Wilcoxon rank sum tests rejecting the null hypothesis of identically distributed HAF scores among carriers and non-carriers within each population sample (Fig. 4C).

Fig. 4 does not show how HAF scores are distributed following fixation of the sweep. Starting at fixation, we see a strong decline in HAF scores owing to the loss of many high frequency derived alleles from the pool of segregating sites. However, crossover events may unlink hitchhiking alleles from the favored allele, and they may remain segregating in the population even after fixation of the favored allele. Therefore, the decline in HAF scores may be abrupt or gradual, depending on the linkage between the favored and hitchhiking alleles. Finally, after reaching a trough, HAF scores gradually recover to their neutral levels over time. The post-fixation dynamics of HAF scores are shown in Fig. S5.

PreCIOSS: Predicting Carriers of Ongoing Selective Sweeps. Our simulations suggest that, in a region undergoing a selective sweep, we could use HAF scores to predict whether a haplotype is carrying the favored allele. We implemented a simple algorithm (PreCIOSS) to carry out this prediction by clustering HAF scores in a sample. PreCIOSS takes as input a set of binary haplotypes sampled from a population undergoing a selective sweep. For each haplotype, the \( \ell \)-HAF score is computed (\( \ell = 1 \) by default). We then fit a Gaussian Mixture Model (GMM) with exactly two Gaussians to the haplotype HAF scores. The fit is performed using Expectation Maximization (EM). Finally, we apply the fitted model to assign a label to each haplotype according to the Gaussian component to which it is assigned. Haplotypes whose HAF score is higher are denoted as ‘carriers’.

We apply PreCIOSS to data from simulated populations undergoing hard and soft sweeps (see ‘Simulations’ in Methods). The haplotypes predicted as carriers (True Positives, TP) or non-carriers (False Positives, FP). Similarly, the haplotypes predicted as non-carriers could be True Negatives (TN) or False Negatives (FN). We measure the balanced accuracy

\[
\frac{1}{2} \left( \frac{TP}{TP + FN} + \frac{TN}{FP + TN} \right),
\]

which is more appropriate to use than Rand accuracy \((TP + TN)/(total\ predictions)\) when the positive and negative classes appear at different proportions in the sample [48].

The balanced accuracy of PreCIOSS on hard sweeps is shown in Fig. 5A for a specific choice of parameters (500 samples with \( n = 200, \theta = 48, \rho = 7.568, s = 0.05 \)). Once the sweep reaches frequencies above 30%, the balanced accuracy increases (median \( \sim 70\% \))
Figure 4. HAF score dynamics in ongoing selective sweeps. HAF scores were computed from 500 simulated population samples ($n = 200$) undergoing a hard sweep ($\theta = 48$, $\rho = 7.568$, $s = 0.05$). (A) Each violin shows the Gaussian kernel density estimation (KDE) of 1-HAF scores in carriers (blue) and non-carriers (red) of the favored allele, as the sweep progresses in frequency. A standard box plot is overlaid on each violin (outlier points omitted), and the means are indicated by stars. The horizontal dotted line represents the neutral expectation, as derived. (B) Corresponding violins showing the in-sample percentile rank of 1-HAF scores. (C) $-\log_2(P)$ values for Wilcoxon rank sum tests rejecting the null hypothesis of identically distributed 1-HAF scores among carriers and non-carriers within each population sample. The number above each bin indicates the fraction of significant tests (where $P < 0.05$, shown by the dotted line).
Figure 5. Predicting carriers of hard and soft sweeps. Balanced accuracy (Eq. (16)) of PreCIOSS in populations undergoing hard and soft sweeps. A total of 500 population samples were simulated ($n = 200$, $\theta = 48$, $\rho = 7.568$) undergoing (A) a hard sweep ($s = 0.05$, $\nu_0 = 1/2000$), and (B) soft sweep ($s = 0.05$, $\nu_0 = 0.3$). We split each sweep into intervals as $\nu$ progresses ($[0, 0.1]$, $[0.1, 0.2]$, etc.). Within each interval for $\nu$, we show the distribution of balanced accuracies observed using standard box plots.

and remains high (median $\sim 90\%$) for the remainder of the sweep. At the very beginning and end of the sweep, the balanced accuracy, despite being asymptotically unbiased, suffers from high variance due to the severe class imbalance (few carriers in the beginning, few non-carriers at the end). The accuracy is high for soft sweeps (Fig. 5B, run with similar parameters, and $\nu_0 = 0.3$), where the higher initial carrier frequency makes for a faster increase in accuracy. As expected under soft sweeps, the overall balanced accuracy is somewhat reduced relative to hard sweeps.

We tested PreCIOSS under a wide range of population-genetic and sweep parameters (Table S1), and observed consistently high balanced accuracy in carrier-state prediction as the sweeps progress (see Fig. S6). Generally, PreCIOSS has slightly better balanced accuracy on hard sweeps than on soft sweeps. A higher recombination rate has only a limited impact (see Fig. S6A,E), suggesting that PreCIOSS will be robust to different recombination scenarios.
Applying PreCIOSS to human selective sweeps

To evaluate the effectiveness of PreCIOSS in distinguishing carriers of a selective sweep from non-carriers, we applied it to several genomic regions (e.g., [39]) where (i) there is strong evidence of a selective sweep, and (ii) the favored allele has been characterized. In applying PreCIOSS to the datasets, we assumed that the region was known, but did not supply the favored allele to PreCIOSS. In each case, we tested if PreCIOSS could separate the haplotypes that carried the favored allele. We use phased haplotypes from the HapMap project [49], setting the ancestral allele to that observed in orthologous Chimpanzee sequence [50].

LCT. We consider the well-known sweep in the lactase (LCT) gene region in Northern Europeans. The best characterized variant is C/T-13910 (rs4988235), for which the T allele was found to be 100% associated with lactase persistence in the Finnish population [21]. T-13910 was further shown to be causal by in-vitro analysis, where it was found to increase enhancer activity [52, 53]. We considered haplotypes from the CEU population, applying PreCIOSS to a 50 kb window centered at C/T-13910 (Fig. 6A). This yielded 100% accuracy in classifying carriers from non-carriers. Increasing the window size above 50 kb, the balanced classification accuracy reduced to ~90% (Fig. 6B). LCT shows the highest and most prolonged (with increasing distance from the causal site) statistical significance in separating carrier and non-carrier HAF scores, remaining highly significant for haplotypes of 1.5 Mb (Fig. 6B, blue line). Despite this highly significant separation, the classification accuracy is initially unstable, alternating between ~100% (perfect classification) and 90%. This is due to the pattern of HAF scores observed in the LCT region, where carriers form a tight cluster with the highest scores, but several non-carriers cluster closer to carriers than to the majority of other non-carriers (Fig. 6A). These haplotypes are therefore sometimes included (90% accuracy) and sometimes excluded (100% accuracy) from the reported ‘carriers’ class.

TRPV6. Transient Receptor Potential Cation Channel, Subfamily V, Member 6 (TRPV6) is a membrane calcium channel thought to mediate the rate-limiting step of dietary calcium absorption. It is reportedly under strong positive selection in several non-African populations [54, 55]. Following Peter et al. (2012) [39], we focus on the CEU population and set rs4987682 as the favored allele. This site, of the three non-synonymous SNPs with highest allele frequency differentiation among human populations, is the only one located in the the N-terminal region of TRPV6, thought to be the target of selection [56]. Applying PreCIOSS to a 50 kb window centered at this site, we obtain ~99% balanced classification accuracy (Fig. 6C), which gradually decays to ~80% when considering a 1.5 Mb window (Fig. 6D). As with LCT, the separation in HAF scores between carriers and non-carriers of the allele is highly statistically significant ($P < 10^{-7}$; see Fig. 6D). Unlike LCT, accuracy decays stably with distance from the favored site. This appears to be due to the less complex clustering pattern of non-carrier HAF scores in the region (Fig. 6C).

PSCA. Prostate Stem Cell Antigen (PSCA) has been proposed to be under selection in a global analysis of allele frequency differentiation [57]. The putative causal site is a non-synonymous SNP (rs2294008) known to be involved in several cancer types [58, 59]. Interestingly, the derived allele is observed in all human populations, but at vastly different frequencies. It is most common in West Africa and East Asia, where it segregates at ~75% frequency. We consider haplotypes from the YRI population and apply PreCIOSS to a 50 kb window centered at rs2294008, yielding balanced accuracy of 97% (Fig. 6E). Unlike LCT and TRPV6, accuracy decays more noticeably with
Figure 6. Predicting carriers of well-known selective sweeps. (Left): Haplotype 1-HAF scores in a 50 kb window centered at known favored sites indicated by the gene name, and the SNP identifier. (A) LCT/rs4988235, (C) TRPV/rs4987682, (E) PSCA/rs2294008, (G) ADH1B/rs1229984, and (I) EDAR/rs3827760. Points represent haplotype 1-HAF scores, red indicating a carrier of the favored allele and blue indicating a non-carrier. At the top of each panel, the number of haplotypes, n, is shown, with the number of carriers in parenthesis. Areas shaded in gray indicate haplotypes designated as ‘carrier’ by PreCIOSS. (Right) classification Balanced Accuracy (black) and $-\log_2(P)$ values (blue) as function of window size around the favored allele in (B) LCT, (D) TRPV6, (F) PSCA, (H) ADH1B, and (J) EDAR. P-values are for Wilcoxon rank sum tests rejecting the null hypothesis of identically distributed 1-HAF scores among carriers and non-carriers. Red circles indicate the 50 kb windows shown on the left.
distance, reaching 63% at 1.25 Mb (Fig. 6F). This sharper decay in accuracy is even more pronounced when considering the sweep in the CHB population (Fig. S4). Such decay is consistent with a (soft) sweep from the standing variation, which would allow more time for recombination to break the linkage between the favored allele and hitchhiking variation. Indeed, the sweep in PSCA was proposed to be from the standing variation by Bhatia et al. (2011) [57], and further substantiated as such by Peter et al. (2012) [39].

ADH1B. ADH1B encodes one of four subunits of Alcohol dehydrogenase (ADH1), which plays a key role in alcohol degradation. ADH1 genes (including ADH1B) have been studied extensively on both a functional and a population-genetic level, as they are thought to be one of the major drivers of alcoholism risk [60]. These genes have also been suggested to cause the “alcohol flush” phenotype common in Asian populations [61]. A specific non-synonymous mutation in ADH1B (Arg47His, rs1229984) has been proposed to be the target of selection. This is because (i) the derived allele has been shown to cause increased enzymatic activity [62,63], and (ii) the estimated age of the allele coincides with rice domestication [61,64] and the availability of fermented beverages [65]. Computing HAF scores for phased haplotypes from East Asian populations (CHB+JPT) and applying PreCIOSS, we obtained balanced classification accuracy of 92% using a 50 kb window centered at rs1229984 (Fig. 6G). Both accuracy and statistical significance (of class separation) gradually decay with increasing window size (Fig. 6H). As before, statistical significance decays more stably than classification accuracy.

EDAR. EDAR encodes a cell-surface receptor and has been associated with development of distinct hair and teeth morphologies [66,67]. Specifically, a non-synonymous SNP (rs3827760, V370A) has been associated with these phenotypes [68]. The SNP is located within a DEATH-domain, which is highly conserved in mammals [69], and has been experimentally confirmed (in vitro) to increase EDAR activity [68]. It is found at very high frequencies in East Asian and American populations, while being completely absent from Europeans and Africans [68]. The EDAR gene has been found to be under selection in multiple studies [30,54,70], showing one of the strongest signatures of selection genome wide among the 1000 Genomes populations [39]. Applying PreCIOSS to phased CHB+JPT haplotypes in a 50 kb region centered at rs3827760, we obtained 98% balanced accuracy in predicting carriers vs. non-carriers of the allele.

In each case, PreCIOSS was applied to a 50 kb window centered at the favored allele, and separated the carriers and non-carriers with high accuracy of 97-100% (Fig. 6A–F). The accuracy decayed with increasing window size, but in many cases stayed high even for windows of 1.5 Mbp.

Discussion

This paper introduces a new perspective on the genetic signatures of selective sweeps. From identifying and characterizing sweeps in a population sample — the topic of typical studies of selective sweeps — we progress to considering the role of individual haplotypes within an ongoing sweep. Using both simulated and real data, we show that the HAF score is well-correlated with the relative fitness of individual haplotypes, and that our algorithm (PreCIOSS) is highly effective at predicting carriers of selective sweeps.

The HAF framework has many natural extensions and potential applications. On the theoretical side, we have obtained the expected HAF score in both constant-sized and exponentially growing populations evolving neutrally (Eqs. [12]). However, we do not yet know the variance. This quantity would provide a better understanding of the respective distributions, and a means to statistically test for deviations from
neutrality. Moreover, although we have observed in simulation and in practice that our theoretical argument is robust to recombination (genealogies violating a tree structure), a theoretical argument supporting these observations would be valuable.

In terms of application, several additional directions are worth investigating. The HAF framework is potentially useful in distinguishing hard from soft sweeps. Intuitively, hard sweep genealogies will likely have a single hitchhiking branch dominating the HAF scores, and leading to near-uniform scores in favored haplotypes. However, soft sweep genealogies may have several hitchhiking branches, potentially leading to distinct HAF score peaks. Even if the different favored clades happen to have similar scores, the haplotypes within them will not form a highly-related group as expected in hard sweeps.

Our results on known selective sweeps in humans illustrates this idea already (Fig. 6). A recent study by Peter et al. (2012) \[39\] assigned posterior probabilities to hard vs. soft sweeps occurring in the same genes. Peter et al. assigned the highest likelihood of a hard sweep to LCT (0.99), followed by EDAR (0.89), ADH1B (0.78), TRPV (0.45), and finally PSCA (0.24). This is in striking concordance with the spread in HAF scores in Fig. 6. The clusters capturing the carriers in LCT and EDAR have tightly distributed HAF scores (Fig. 6A, I). The cluster for ADH1B (Fig. 6G) has more variance by comparison, and the variance increases for TRPV6 (Fig. 6C) and PSCA (Fig. 6E), with PSCA showing the highest variance of HAF scores in carriers.

Finally, perhaps the highest potential impact of the HAF score could be in predicting the ‘MRCA of the future’. We know that future haplotypes are more likely similar to favored individuals than to unfavored ones, and that HAF scores correlate well with relative fitness in ongoing selective sweeps. Therefore, high HAF haplotypes are more likely to be similar to future generations. This relationship is particularly valuable when action may be taken based on such predictions. For instance, rapid influenza viral evolution is known to change the strain composition from year to year. The mutations are a mix of favored and deleterious mutations. The fitness and frequency of the current year’s strain have been used to predict the next year’s dominant strain \[71\]. The HAF score may allow for a careful look at the dynamics of the current strain and possibly offer better insight into the problem. As a second example, tumor cells show great heterogeneity and much variation occurs at the single cell level. This intra-tumor variation allows sub-population of cells to resist therapy and proliferate \[72\]. Once again, HAF scores of haplotypes in cells undergoing treatment can potentially distinguish between carriers and non-carriers of drug resistance mutations, and thereby improve our insight into mechanisms of drug resistance.

Methods

Simulations. We simulated data for various evolutionary scenarios. Neutral samples were generated using the coalescent (backwards) simulator ms \[73\]; sweep samples were generated using the Wright-Fisher (forwards) simulator mpop \[5\]. All simulations generated samples of \(n \in \{100, 200, 300, 400\}\) haplotypes from a larger effective population of \(N = 2000\) haplotypes, each of length 50 kb. A mutation rate of approximately \(\mu = 2.4 \cdot 10^{-8}\) mutations per bp per generation was used \[74, 75\]. For simulating populations from a region of length \(L = 50\) kb, and assuming effective population size of 10,000 individuals (20,000 haplotypes), we obtain \(\theta = 2N\mu L = 24\). For our simulations, we choose \(\theta \in \{24, 48\}\). However, we worked with a smaller value of \(N = 2,000\) haplotypes to reduce the computational requirements of forward simulation, and compensated by appropriately scaling the mutation rate to \(2.4 \cdot 10^{-7}\) to maintain the scaled mutation rate, \(\theta\).

The simulations either had no recombination, or used \(3.784 \cdot 10^{-8}\) crossovers per
generation, per base (e.g. [76]). For a 50 kb region, we used the population-scaled rate
\[
\rho = 2N \cdot (3.784 \cdot 10^{-5}) \cdot (5 \cdot 10^4) = 7.568 .
\] (17)
We also worked with a higher population-scaled rate of \( \rho = 15 \).

For exponential growth, we used \( N = 2000 \) as the size of the final (current) population. Let \( r \) denote the growth rate per generation, so that at \( t \) generations prior to the current generation, the population size was \( N(t) = Ne^{-rt} \). Define the scaled growth rate \( \alpha = 2N\!\!r \). We set \( \alpha \) to a range of values in \([30, 100]\) (corresponding to \( r \in [0.0075, 0.025] \)).

For selective sweeps, we used forward simulations run assuming a diploid population with recombination and mutation parameters as described above. While diploid populations were simulated to incorporate recombination, we used phased haplotypes for our analysis. We assumed a single favored allele under selection coefficient \( s \in [0.02, 0.8] \) and heterozygosity 0.5 (haploid carriers get half the fitness advantage of diploid carriers). When \( s \) is ‘low’ \((0.001 \leq s \leq 0.01)\), none of the available tests can detect a selective sweep with reasonable power [21, 23]. Selection with \( s \geq 0.1 \) is considered ‘high’ (e.g., see [21]). For high values of \( s \), the carrier haplotypes are identical or very similar in simulations, making the problem of detecting carriers easy. Therefore, we chose intermediate values \((s \in [0.02, 0.08])\) in our simulations.

Soft sweeps can arise either due to standing variation or due to multiple favored alleles. Here, we focus on the former, where the favored allele is present in at least one carrier in the population \((\nu_0 \in [1/N, 1])\), and drifting at the onset of selection. In our simulations, we set \( \nu_0 \) at the beginning of the sweep to \( \nu_0 = 1/N \) for hard sweeps and \( \nu_0 = 0.3 \) for soft sweeps.

**Data Preprocessing.** We downloaded pre-phased haplotype data from the HapMap [49] project website. Both HapMap 3 [49] and HapMap 2 [77] project data were used depending on whether the causal allele was sampled or not. For LCT (rs4988235), we used 88 CEU individuals haplotypes from HapMap 3; for PSCA (rs2294008), we used 60 YRI individuals from HapMap 2; for TRPV6, 88 CEU individuals from HapMap 3; for ADH1B (rs1229984), 90 CHB+JPT individuals from HapMap 2; and, for EDAR (rs3827760), we chose 90 CHB+JPT individuals from HapMap 2. The number of phased haplotypes was twice the number of individuals in each case.

We downloaded Chimpanzee genome alignments [50] to identify the ancestral allele. A total of \( \sim 93\% \) of the sites analyzed had were covered by the Chimpanzee data. For these sites, we set the ancestral allele to the Chimpanzee allele, and we discarded sites that were not covered.

**Software**

The PreCIOSS software is available from the website http://bix.ucsd.edu/projects/precioss/ .

**Acknowledgments**

This work was supported in part by National Science Foundation grants CCF-1115206 and IIS-1318386.

**References**


common variation in EDAR is a genetic determinant of shovel-shaped incisors.

Positive selection in East Asians for an EDAR allele that enhances NF-kappaB

Genome-wide detection and characterization of positive selection in human populations.

CD. Simultaneous inference of selection and population growth from patterns of
variation in the human genome. Proc Natl Acad Sci USA. 2005 May;102(22):7882–
7887.


analyses of transcriptional heterogeneity during drug tolerance transition in cancer

73. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic

74. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in


76. Fledel-Alon A, Leffler EM, Guan Y, Stephens M, Coop G, Przeworski M. Vari-

77. Frazer KA, et al. A second generation human haplotype map of over 3.1 million

78. Kingman JFC. On the genealogy of large populations. Journal of Applied

79. Campbell RB. Coalescent size versus coalescent time with strong selection. Bull
Appendix

Figure S1. HAF scores in neutrally evolving constant-sized populations. The distribution of $4 \times 10^6$ 1-HAF scores aggregated from 20,000 population samples (each of $n = 200$ haplotypes) simulated under a standard coalescent model without recombination. Plugging in the simulation parameters $\theta = 48$, $n = 200$, Eqs. (3),(7) evaluate to 4776. The observed mean 1-HAF score is 4777.1.
Figure S2. HAF scores in neutrally evolving exponentially growing populations. The distribution of $4 \times 10^9$ 1-HAF scores aggregated from 20,000 population samples (each of $n = 200$ haplotypes) simulated under a coalescent model of exponential growth without recombination. Computing the conditional expectation as described in Eq. (12) with the simulation parameters ($\theta = 48$, $n = 200$, $\alpha = 80$) gives 126.9. The observed mean 1-HAF score is 128.1.
Figure S3. HAF scores for a range of simulation parameters. Each empirical test is the average of 1,000 trials. (A) Empirical mean and theoretical expected ℓ-HAF scores for a fixed size population (ℓ ∈ {1, 2, 3, 4}, θ ∈ {24, 48}, ρ = 0). (B) Empirical mean and theoretical expected 1-HAF scores for an exponentially growing population (α ∈ {0, 30, 60, 80}, θ ∈ {24, 48}, ρ = 0). (C) Theoretical expected 1-HAF scores (computed assuming ρ = 0) compared against empirical means of 1-HAF scores from samples with different recombination rates (ρ ∈ {0, 7.568, 15}, θ ∈ {24, 48}). (D) Interestingly, higher recombination rates reduce the variance in 1-HAF estimates. In the three green curves for θ = 24 (and in the three red curves for θ = 48), the variation from the expected value (blue) decreases as ρ increases. Rate ρ = 0 (dotted) has the most variation; ρ = 7.568 (dashed) has less; and ρ = 15 (solid) has the least. The theoretical values are based on (A) Eqs. 3, S22, (B) Eq. 12, and (C,D) Eq. 4.
Figure S4. Predicting carriers of the PSCA sweep in CHB. (A) Haplotype 1-HAF scores in a 50 kb window centered at the favored site. (B) Balanced classification accuracy (black) and $-\log_2(P)$ values (blue) as function of window size around the favored allele. $P$-values are for Wilcoxon rank sum tests rejecting the null hypothesis of identically distributed 1-HAF scores among carriers and non-carriers. The red circle indicates the balanced accuracy obtained for the 50 kb window shown on the left. As with the YRI population, we achieve high classification accuracy when considering $\sim 100$ kb window centered at the favored allele. But unlike in YRI, we see a sharp decline in both accuracy and $-\log_2(P)$ values beginning at larger distances from the favored allele. See Fig. 6 for further details on the conventions used.

Figure S5. Recovery of HAF scores after a selective sweep. Each violin shows the Gaussian kernel density estimation (KDE) of 1-HAF scores in populations sampled at regular time intervals following the fixation of a selective sweep. All individuals at this stage are carriers of the favored allele. A standard boxplot is overlaid on each violin (outlier points not shown). The horizontal dotted line represents the neutral expected value. HAF scores were computed from 500 simulated populations undergoing a hard sweep, sampled at regular time intervals.
Empirical validation of PreCIOSS

In Fig. 5A, we showed the balanced accuracy of PreCIOSS for predicting carriers of ongoing hard sweeps. For soft sweeps, the balanced accuracy stays robust but is slightly reduced (Fig. 5B). In Fig. S6 we apply PreCIOSS to data from 500 trials of a simulated population for each sweep type (hard or soft), and for different population genetics parameters provided in Table S1. For soft sweeps, we set $\nu_0 = 0.3$ (frequency of the favored allele at the onset of selection).

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$s$</th>
<th>$\theta$</th>
<th>$n$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.05</td>
<td>48</td>
<td>200</td>
<td>7.568</td>
</tr>
<tr>
<td>B</td>
<td>0.02</td>
<td>48</td>
<td>200</td>
<td>7.568</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>24</td>
<td>200</td>
<td>7.568</td>
</tr>
<tr>
<td>D</td>
<td>0.05</td>
<td>48</td>
<td>100</td>
<td>7.568</td>
</tr>
<tr>
<td>E</td>
<td>0.05</td>
<td>48</td>
<td>200</td>
<td>15</td>
</tr>
</tbody>
</table>

Table S1. Simulation parameter sets used for generating Fig. S6. In simulations B through E, we changed one parameter (in boldface) at a time vs. simulation A.
Figure S6. Predicting carriers of hard and soft sweeps. Balanced accuracy of PreCIOSS in populations undergoing hard (left figures) and soft (right figures) sweeps. Balanced accuracy is shown for each allele frequency bin as a standard box plot computed over 500 simulated populations (for each sweep type and each parameter set in Table S1) sampled at regular time intervals during sweep progression.
Derivation of the expected $ℓ$-HAF score.

We compute the expected value by averaging over all haplotypes in a genealogy (sum over the haplotypes and divide by $n$) and then averaging over all genealogies. Recall from the main text that $M_{k,i}$ and $W_{k,i}$ refer to random variables denoting the number of mutations and the frequency, respectively on the $i^{th}$ lineage of the $k^{th}$ epoch. As the genealogy of a neutrally evolving sample is independent of branch lengths [43], $M_{k,i}$ and $W_{k,i}$ are independent random variables. Recall Eq. (7):

$$
\mathbb{E}[\ell\text{-HAF}] = \frac{1}{n} \sum_{k=2}^{n} \sum_{i=1}^{k} \mathbb{E}[\ell i M_{k,i} W_{k,i}^{\ell+1}] = \frac{1}{n} \sum_{k=2}^{n} \sum_{i=1}^{k} \mathbb{E}[M_{k,i}] \mathbb{E}[W_{k,i}^{\ell+1}].
$$

(S1)

Expected value of $M_{k,i}$, constant population size. Let $T_k$ denote the duration of epoch $k$. For a population of constant size $N$, the duration $T_k$ is exponentially distributed with rate ${k \choose 2}/N$ (see [78]):

$$
\mathbb{E}[T_k] = \frac{N}{k(k-1)}.
$$

(S2)

and the number of mutations on a lineage $i$ in epoch $k$, denoted $M_{k,i}$, is Poisson distributed with rate $\mu T_k$. For a constant-sized population, this implies

$$
\mathbb{E}[M_{k,i}] = \frac{N \mu}{k(k-1)} = \frac{\theta}{k(k-1)}.
$$

(S3)

Later, we will explore $\mathbb{E}[T_k]$ and $\mathbb{E}[M_{k,i}]$ for exponentially growing populations.

Rising factorials and moment calculations. The probability distribution of $W_{k,i}$, or the number of descendants of the $i^{th}$ lineage in epoch $k$, is given by [42, Eq. (14)]:

$$
\Pr(W_{k,i} = w) = \frac{\binom{n-w-1}{k-2} \binom{n-1}{k-1}}{n},
$$

(S4)

where $1 \leq i \leq k$, $2 \leq k \leq n$, and $1 \leq w \leq n - k + 1$. We wish to evaluate Eq. (7) (or (S1)) for $\ell = 1$, and generally for any $\ell$. This requires $\mathbb{E}[M_{k,i}]$, which we already have for constant populations, and $\mathbb{E}[W_{k,i}^{\ell}]$, which we derive next. For a positive integer $\ell$, denote the rising factorial

$$
w^{(\ell)} = w(w+1)(w+2)\cdots(w+\ell-1)
$$

(S5)

and set $w^{(0)} = 1$. Then $W_{k,i}^{(\ell)}$ denotes the $\ell^{th}$ rising factorial of random variable $W_{k,i}$, while $W_{k,i}^{(\ell)}$ denotes the ordinary $\ell^{th}$ power. We will compute $\mathbb{E}[W_{k,i}^{(\ell)}]$, and use it to compute $\mathbb{E}[W_{k,i}^{\ell}]$.

Deriving $\mathbb{E}[W_{k,i}^{(\ell)}]$

Let $\ell$ be a nonnegative integer. We will prove that

$$
\mathbb{E} \left[ \frac{W_{k,i}^{(\ell)}}{\ell!} \right] = \frac{n^{(\ell)}}{k^{(\ell)}},
$$

(S6)

which is equivalent to Eq. (9) in the text. Using Eq. (S4), the expected value is the sum

$$
\mathbb{E} \left[ \frac{W_{k,i}^{(\ell)}}{\ell!} \right] = \sum_{w=1}^{n-k+1} \frac{n-w-1}{(k-2)} \cdot \frac{n-w-1}{(k-1)} \cdot \frac{1}{\ell!} = \frac{1}{\ell!} \sum_{w=1}^{n-k+1} \frac{n-w-1}{k-2} \cdot \left( \frac{n-w-1}{k-1} \right) \cdot \left( \frac{w+\ell-1}{\ell} \right).
$$

(S7)
Here we use that rising factorials and binomial coefficients are related via \( \frac{n^{(l)}}{l!} = (w+\ell-1)! \).

The rightmost sum may be evaluated using a combinatorial method: we will count all \((k + \ell - 1)\)-element subsets of \(\{1, 2, \ldots, n + \ell - 1\}\) by two methods. First, it is the binomial coefficient \(\binom{n+\ell-1}{k+\ell-1}\). Second, we will show that it also equals the rightmost sum in \(S7\).

To show this, we may order any such subset and partition it into three parts, \((A, b, C)\): \(A\) is the set of its smallest \(\ell\) elements; \(b\) is the \((\ell + 1)\)th smallest element; and \(C\) is the set of the largest \(k - 2\) elements.

Element \(b\) must be in the range \(1 \leq b \leq n - k + \ell + 1\) to allow for \(\ell\) smaller elements and \(k - 2\) larger elements. We rewrite \(b\) as \(b = \ell + w\), where \(w\) is in the range \(1 \leq w \leq n - k + 1\).

Given \(b\), we may choose \(A\) in one of \(\binom{b-1}{\ell}\) ways and choose \(C\) in one of \(\binom{n+\ell-1-b}{k-2}\) ways. Multiply these counts for a given \(b\), and sum over all \(b\) (by summing over all \(w\)) to obtain that the total number of subsets is the rightmost sum in \(S7\). But we also showed the number of such subsets is \(\binom{n+\ell-1}{k+\ell-1}\). Thus, \(S7\) evaluates to

\[
E \left[ \frac{W_{k,1}^{(l)}}{l!} \right] = \frac{(n+\ell-1)!}{(k + \ell - 1)! (n-k)!} / \frac{(n-1)!}{(k-1)! (n-k)!} \left[ \frac{(n-1)!}{(k+\ell-1)! (n-1)!} \right] = \frac{(n+\ell-1)!}{(k + \ell - 1)! (n-k)!} \frac{(n-1)!}{(k-1)! (n-k)!} = \binom{n+\ell-1}{k+\ell-1}.
\]

**Deriving \(E[\ell\text{-HAF}]\) in a constant-sized population**

Our derivation of \(E[\ell\text{-HAF}]\) consists of three main steps. We start by defining a more general form of the HAF score for an arbitrary function \(f(w)\), which we denote HAF\(_f\). We then set \(f(w)\) to the rising factorial function and derive \(E[HAF_w(w)]\). Finally, we use Stirling Numbers to convert between rising factorials and powers, and obtain \(E[\ell\text{-HAF}]\).

**Step 1: The generalized form \(E[HAF_f]\)**

Consider an arbitrary polynomial \(f(w)\) (or more generally, an arbitrary function \(f : \mathbb{Z} \to \mathbb{Z}\)). Define the HAF\(_f\) score of a HAF vector \(c\) as

\[
HAF_f(c) = \sum_j f(c_j) = \sum_{k=2}^n m_k f(w_k). \tag{S8}
\]

We now generalize the derivation of Eq. \((7)\) to HAF\(_f\). Sum the above equation over all haplotypes in a genealogy:

\[
\sum_{v=1}^n HAF_f(c(v)) = \sum_{k=2}^n \sum_{v=1}^n m_k(v) f(w_k(v)) = \sum_{k=2}^n \sum_{i=1}^k m_{k,i} f(w_{k,i}) = \sum_{k=2}^n \sum_{i=1}^k m_{k,i} w_{k,i} \cdot f(w_{k,i}). \tag{S9}
\]

Note that in the sum over \(j\), all summands are independent of \(j\), so that sum is replaced by multiplying by the number of terms, \(w_{k,i}\). We divide the above sum by \(n\) and then average over all genealogies to obtain

\[
E[HAF_f] = \frac{1}{n} \sum_{k=2}^n \sum_{i=1}^k E[M_{k,i} W_{k,i} f(W_{k,i})] = \frac{1}{n} \sum_{k=2}^n \sum_{i=1}^k E[M_{k,i}] E[W_{k,i}] f(W_{k,i}). \tag{S10}
\]
Further, since the random variables $W_{k,i}$ for $i = 1, \ldots, k$ are identically distributed, all terms in the inner sum are the same, and may be consolidated into $k$ times the value of the term on one branch:

$$
E[HAF_f] = \frac{1}{n} \sum_{k=2}^{n} k \cdot E[M_{k,1}] E[W_{k,1} \cdot f(W_{k,1})].
$$  \hfill (S11)

In a constant-sized population, Eq. (S3) states that $E[M_{k,i}] = \theta/(k(k-1))$. Plug this in:

$$
E[HAF_f] = \frac{\theta}{n} \sum_{k=2}^{n} \frac{1}{k-1} E[W_{k,1} \cdot f(W_{k,1})].
$$  \hfill (S12)

**Step 2: The rising factorial form $E[HAF_{w(\ell)}]$**

We will show that for $f(w) = w^{(\ell)}$, for any nonnegative integer $\ell$:

$$
E[HAF_{w(\ell)}] = \frac{\theta}{n(\ell + 1)} \cdot (n - 1)^{\ell+1}.
$$  \hfill (S13)

For this proof only, abbreviate $W = W_{k,1}$. Note that

$$
W \cdot W^{(\ell)} = (W + \ell - \ell)W^{(\ell)} = (W + \ell)W^{(\ell)} - \ell W^{(\ell)} = W^{(\ell+1)} - \ell W^{(\ell)}.
$$  \hfill (S14)

Using this and Eq. (S6) gives

$$
E[HAF_{w(\ell)}] = \frac{\theta}{n} \sum_{k=2}^{n} \frac{1}{k-1} E[W \cdot W^{(\ell)}]
$$

$$
= \frac{\theta}{n} \sum_{k=2}^{n} \frac{1}{k-1} E[W^{(\ell+1)} - \ell W^{(\ell)}]
$$

$$
= \frac{\theta}{n} \sum_{k=2}^{n} \frac{1}{k-1} \left( (\ell + 1)! \cdot \frac{n^{(\ell+1)}}{k^{(\ell+1)}} - \ell ! \cdot \frac{n^{(\ell)}}{k^{(\ell)}} \right)
$$

$$
= \frac{\theta \cdot \ell! \cdot n^{(\ell)}}{n} \sum_{k=2}^{n} \left( \frac{(\ell + 1)(n + \ell)}{(k-1)^{(\ell+2)}} - \frac{\ell}{(k-1)^{(\ell+1)}} \right).
$$  \hfill (S15)

We will split this sum into two telescoping sums. One may show that

$$
\frac{1}{k^{(\ell)}} = \frac{1}{\ell - 1} \left( \frac{k + \ell - 1}{k^{(\ell)}} - \frac{k + \ell}{(k + 1)^{(\ell)}} \right).
$$  \hfill (S17)

Using this with shifted $k'$s and $\ell'$s, part of the sum in Eq. (S16) can be written

$$
\sum_{k=2}^{n} \frac{1}{(k-1)^{\ell+2}} = \sum_{k=2}^{n} \frac{1}{\ell + 1} \left( \frac{k + \ell}{(k-1)^{(\ell+2)}} - \frac{k + \ell + 1}{k^{(\ell+2)}} \right),
$$

which is a telescoping sum that evaluates to

$$
= \frac{1}{\ell + 1} \left( \frac{2 + \ell}{1^{(\ell+2)}} - \frac{n + \ell + 1}{n^{(\ell+2)}} \right).
$$

Note that $1^{(\ell+2)} = (\ell + 2)!$, simplifying this to

$$
= \frac{1}{\ell + 1} \left( \frac{1}{(\ell + 1)!} - \frac{n + \ell + 1}{n^{(\ell+2)}} \right).
$$  \hfill (S18)
Similarly,
\[
\sum_{k=2}^{n} \frac{1}{(k-1)(\ell+1)^{n-k+1}} = \frac{1}{\ell} \left( \frac{1}{\ell!} - \frac{n+\ell}{n^{\ell+1+1}} \right).
\] (S19)

Plugging Eqs. (S18), (S19) into Eq. (S16), we obtain:
\[
\mathbb{E}[\text{HAF}_{w(\ell)}] = \frac{\theta \cdot \ell! \cdot n^{(\ell)}}{n} \left( \frac{(\ell+1)(n+\ell)}{\ell+1} \right) \left( \frac{1}{(\ell+1)!} - \frac{n+\ell+1}{n^{(\ell+2)}} - \frac{\ell}{\ell!} \left( \frac{1}{\ell!} - \frac{n+\ell}{n^{(\ell+1)}} \right) \right)
\]
\[
= \frac{\theta \cdot n^{(\ell)}}{n} \left( \frac{n+\ell+1}{\ell+1} - \frac{1}{n^{(\ell+1)}} \right) - \theta \cdot \frac{n+\ell}{n^{(\ell+1)}}
\]
\[
= \frac{\theta \cdot n^{(\ell)}}{n} \left( \frac{n+\ell+1}{\ell+1} - 1 \right) = \frac{\theta \cdot n^{(\ell)}}{n} \cdot \frac{n-1}{\ell+1} = \frac{\theta \cdot (n-1)^{\ell+1}}{n^{(\ell+1)}}.
\]

This proves Eq. (S13).

**Step 3: The final form** $\mathbb{E}[\ell\text{-HAF}]$

Powers $w^\ell$ and rising factorials $w^{(\ell)}$ may be expressed as linear combinations of each other via Stirling numbers. Let $c(\ell, q)$ denote the Unsigned Stirling Number of the First Kind, and $S(\ell, q)$ denote the Stirling Number of the Second Kind. Then [45, p. 264]
\[
(a) \quad w^\ell = \sum_{q=0}^{\ell} (-1)^{\ell-q} S(\ell, q) w^{(q)}, \quad (b) \quad w^{(\ell)} = \sum_{q=0}^{\ell} c(\ell, q) w^q.
\] (S20)

The expected value of $(W_{k,i})^\ell$ is then:
\[
\mathbb{E}[(W_{k,i})^\ell] = \sum_{q=0}^{\ell} (-1)^{\ell-q} S(\ell, q) \mathbb{E}[(W_{k,i})^{(q)}] = \sum_{q=0}^{\ell} (-1)^{\ell-q} S(\ell, q) q! \cdot \frac{n^{(q)}}{k^{(q)}}.
\] (S21)

We evaluate $\mathbb{E}[\ell\text{-HAF}] = \mathbb{E}[\text{HAF}_{w^\ell}]$ as a linear combination of terms $\mathbb{E}[\text{HAF}_{w^{(q)}}]$, using the same coefficients that express $w^\ell$ as a linear combination of terms of $w^{(q)}$ (Eq. (S20a)):
\[
\mathbb{E}[\ell\text{-HAF}] = \sum_{q=0}^{\ell} (-1)^{\ell-q} S(\ell, q) \mathbb{E}[\text{HAF}_{w^{(q)}}] = \frac{\theta}{n} \sum_{q=0}^{\ell} (-1)^{\ell-q} S(\ell, q) \cdot \frac{(n-1)^{(q+1)}}{q+1}.
\] (S22)

For $\ell = 0$:
\[
\mathbb{E}[0\text{-HAF}] = \frac{\theta}{n} (-1)^{0-0} S(0, 0) \frac{(n-1)^{(1)}}{1} = \frac{\theta}{n} 1 \cdot (n-1) = \frac{\theta(n-1)}{n} = \theta \left( 1 - \frac{1}{n} \right).
\] (S23)

For $\ell = 1$:
\[
\mathbb{E}[1\text{-HAF}] = \frac{\theta}{n} \left( (-1)^{1-0} S(1, 0) \frac{(n-1)^{(1)}}{1} + (-1)^{1-1} S(1, 1) \frac{(n-1)^{(2)}}{2} \right)
\]
\[
= \frac{\theta}{n} \left( -0 \cdot (n-1) + 1 \cdot \frac{(n-1)(n)}{2} \right) = \theta \cdot \frac{n-1}{2}.
\]

For $\ell = 2$:
\[
\mathbb{E}[2\text{-HAF}] = \frac{\theta}{n} \left( (-1)^{2-0} S(2, 0) \frac{(n-1)^{(1)}}{1} + (-1)^{2-1} S(2, 1) \frac{(n-1)^{(2)}}{2} \right.
\]
\[
+ (-1)^{2-2} S(2, 2) \frac{(n-1)^{(3)}}{3} \right)
\]
\[
= \frac{\theta}{n} \left( 0 \cdot \frac{n-1}{1} - 1 \cdot \frac{(n-1)n}{2} + 1 \cdot \frac{(n-1)n(n+1)}{3} \right) = \theta(n-1)(2n-1).
\]
It is straightforward to compute this for any $\ell$ in the same fashion.

$$E[3\text{-HAF}] = \frac{\theta \cdot n(n-1)^2}{4} \quad E[4\text{-HAF}] = \frac{\theta \cdot (n-1)(2n-1)(3n^2 - 3n - 1)}{30} \quad \ldots$$

**Equivalence of Eq. (3) and Eq. (7) for $E[\ell\text{-HAF}]$ in a constant-sized population**

Eq. (3) has simple derivation, but a variable number of terms, $n$, that grows as $n$ grows. Eq. (7) leads to an evaluation (S22) that is a polynomial in $n$ of degree $\ell$. Both represent $E[\ell\text{-HAF}]$ in a constant-sized population; we now show that they are equal as a consequence of the following lemma:

**Lemma 1** In a constant-sized population, for any polynomial $f(w)$, we have

$$E[HAF_f] = \frac{\theta}{n} \sum_{w=1}^{n-1} f(w).$$

Applying this to $f(w) = w^\ell$ gives equivalence of the two formulas (3) and (S22) for $E[\ell\text{-HAF}]$.

**Proof of lemma.** Both sides are linear in $f$, so it suffices to prove it on any basis of polynomials. We use the basis of rising factorials, $f(w) = w^\ell$ for $\ell \geq 0$. Eq. (S13) evaluates $E[HAF_f]$ on this basis. We now evaluate $\sum_{w=1}^{n-1} w^\ell$. Note that

$$w^{(\ell+1)} - (w-1)^{(\ell+1)} = w^\ell \cdot (w + \ell - (w - 1)) = (\ell + 1) \cdot w^\ell,$$

so

$$w^\ell = \frac{w^{(\ell+1)} - (w-1)^{(\ell+1)}}{\ell + 1}.$$

This leads to a telescoping sum:

$$\sum_{w=1}^{n-1} w^\ell = \frac{1}{\ell + 1} \sum_{w=1}^{n-1} \left( w^{(\ell+1)} - (w-1)^{(\ell+1)} \right) = \frac{(n-1)^{(\ell+1)} - 0^{(\ell+1)}}{\ell + 1} = \frac{(n-1)^{(\ell+1)}}{\ell + 1}.$$

Multiplying this sum by $\theta/n$ gives the same value as (S13) gives for $E[HAF_{w^\ell}]$. Thus, the lemma holds on the basis $f(w) = w^\ell$ for $\ell \geq 0$. By linearity, it holds on all polynomials $f(w)$. ■

**Deriving $E[\ell\text{-HAF}]$ in an exponentially growing population**

Let the population size at time $t$ in the past be given by $N(t) = N_0 e^{-rt}$, where $N_0$ is the current population size. $M_{k,i}$ depends on $T_k$, which in turn depends on the size of the population in epoch $k$. Slatkin and Hudson (1991) [44, p. 559] derived an iterative formula to generate a sequence of random times $T_k$ ($k = n, n-1, \ldots, 2$). The actual formula implemented in the simulator **ms** [73] is slightly different; in our notation, it is as follows. Let $\alpha = 2N_0r$ be a scaled growth rate. Generate random values $U_n, \ldots, U_2$ that are uniformly distributed in $(0,1)$, and iteratively compute $T_n, \ldots, T_2$ via

$$T_k = \frac{1}{r} \ln \left[ 1 - \frac{\alpha}{k(k-1)} e^{-r\tau_k} \ln(U_k) \right], \quad \text{where} \quad \tau_k = \sum_{i=k+1}^{n} T_i. \quad \text{(S24)}$$

Then generate a random value for each $M_{k,i}$ using a Poisson distribution with mean $\mu T_k$ ($k = 2, \ldots, n$ and $i = 1, \ldots, k$). This does not lead to a closed form expression for
the expected HAF scores, but we may nonetheless use it to estimate the expected HAF scores computationally, as illustrated in Fig. S2.

The first method, cumulative time, is to generate random values of $T_k$’s and $M_{k,i}$’s as above, and plug them into (S10), along with exact values of $\mathbb{E}[W_{k,i} \cdot f(W_{k,i})]$ computed using (S6) or (S21).

The second method is conditional expectation. For a given $k$, we may compute the expected value of $\mathbb{E}[\text{HAF}]$ given the condition that $T_{k+1}, \ldots, T_n$ are known. This gives estimates $T_k = t_k$ (computed in order $k = n, n-1, \ldots, 2$) as follows:

$$t_k = \mathbb{E}[T_k \mid T_{k+1} = t_{k+1}, \ldots, T_n = t_n]$$

$$= \frac{1}{r} \int_0^1 \ln \left( 1 - \frac{\alpha}{k(k-1)} e^{-r \tau_k} \ln(u) \right) du$$

$$= \frac{1}{r} \exp \left( \frac{k(k-1)}{\alpha} e^{r \tau_n} \right) E_1 \left( \frac{k(k-1)}{\alpha} e^{r \tau_n} \right),$$

where $\tau_k = t_{k+1} + \cdots + t_n$ (with $\tau_n = 0$) and $E_1(x)$ is the exponential integral $E_1(x) = \int_1^\infty \exp(-x t) \, dt$.

In practice, Eq. (S25) is used to generate scaled times,

$$r \cdot t_k = \mathbb{E}[r \cdot T_k \mid r \cdot T_{k+1} = r \cdot t_{k+1}, \ldots, r \cdot T_n = r \cdot t_n].$$

In terms of these times, $\mathbb{E}[M_{k,i}] = \mu \mathbb{E}[T_k] \approx \mu t_k$ (which is approximate since the formula for $t_k$ uses conditional expectation while this equation does not). We then estimate $\mathbb{E}[\text{HAF}]$ via (S11):

$$\mathbb{E}[\text{HAF}] \approx \frac{1}{n} \sum_{k=2}^n k \cdot \mu k \cdot \mathbb{E}[W_{k,1} \cdot f(W_{k,1})] = \frac{\mu}{r} \sum_{k=2}^n (r \cdot t_k) \cdot \frac{k}{n} \cdot \mathbb{E}[W_{k,1} \cdot f(W_{k,1})].$$

(S26)

For $f(w) = w^{(\ell)}$, similarly to (S15), this gives an estimate

$$\mathbb{E}[\text{HAF}^{(\ell)}] \approx \frac{\mu}{r} \sum_{k=2}^n (r \cdot t_k) \cdot \frac{k}{n} \left( (\ell + 1) \frac{n^{(\ell+1)}}{k^{(\ell+1)}} - \ell \cdot \ell! \frac{n^{(\ell)}}{k^{(\ell)}} \right)$$

$$= \frac{\mu}{r} \sum_{k=2}^n (r \cdot t_k) \cdot \left( (\ell + 1) \frac{n^{(\ell+1)}}{k^{(\ell+1)}} - \ell \cdot \ell! \left( \frac{(n+1)^{(\ell)}}{k^{(\ell)}} - \frac{(n+1)^{(\ell-1)}}{k^{(\ell-1)}} \right) \right),$$

(S27)

and for $f(w) = w^q$, we use (S20) to obtain

$$\mathbb{E}[\ell-\text{HAF}] \approx \frac{\mu}{r} \sum_{k=2}^n (r \cdot t_k) \sum_{q=0}^\ell (-1)^{\ell-q} S(\ell, q) \left( (q + 1)! \frac{(n+1)^{(q)}}{(k+1)^{(q)}} - q \cdot q! \frac{(n+1)^{(q-1)}}{(k+1)^{(q-1)}} \right).$$

(S28)

In terms of $\alpha = 2 N_0 r$ and $\theta = 2 N_0 \mu$, the coefficient $\mu/r$ may be replaced by $\theta/\alpha$.

We used Maple (maplesoft.com) to generate the sequence of scaled times $r \cdot t_k$ by iterating Eq. (S25). Maple supports arbitrary precision numerical computations, which is needed due to complications with double precision arithmetic. Let $x = \frac{k(k-1)}{\exp(r \cdot \tau_k)}$. Small rounding errors in $x$ due to limited precision are amplified in $e^x$, which may lead to slightly different results when iterating the recursion using double precision arithmetic vs. Maple’s arbitrary precision arithmetic. Additionally, as $x$ increases, $e^x$ grows rapidly and leads to an overflow in double precision arithmetic at approximately $x \geq 709.7827$, while $E_1(x)$ decays rapidly and leads to an underflow at approximately $x \geq 701.8334$, even
though $e^x E_1(x)$ is usually representable in double precision. For example, $E_1(702) \approx 1.9 \cdot 10^{-308}$ underflows in double precision arithmetic, but $e^{702} E_1(702) \approx 0.0014$ is representable. The largest value of $x$ used in Eq. (S25) is $x = n(n-1)/\alpha$ in iteration $k = n$, leading to an underflow in double precision if $n(n-1)/\alpha > 701.8$.

### Number of mutations in a genealogy

Let $f(w) = 1/w$ for $w > 0$ and $f(0) = 0$. Note that for an individual HAF vector $c$, some entries may be 0, but in a genealogy, all $w_{k,i} \neq 0$ since every branch of the observed tree has at least one haplotype as its descendant. The sum (S9) of HAF$_f(c)$ over all haplotypes in a genealogy becomes

$$\sum_{v=1}^{n} \text{HAF}_f(c(v)) = \sum_{k=2}^{n} \sum_{i=1}^{k} m_{k,i},$$

(S29)

and thus (S11) becomes

$$\mathbb{E}[\text{HAF}_f] = \frac{1}{n} \sum_{k=2}^{n} k \cdot \mathbb{E}[M_{k,1}].$$

(S30)

In a constant-sized population, Eqs. (S3) and (S29) give that the expected number of mutations in a genealogy is

$$\mathbb{E}[\text{HAF}_f] = \frac{1}{n} \sum_{k=2}^{n} k \cdot \frac{\theta}{k(k-1)} = \frac{\theta}{n} \sum_{k=2}^{n} \frac{1}{k-1} = \frac{\theta}{n} H(n-1),$$

(S31)

where $H(n) = \sum_{k=1}^{n} \frac{1}{k}$ is the Harmonic number.

Under exponential growth, we instead estimate times $t_k$ or scaled times $r \cdot t_k$ via (11).

Then we estimate the expected number of mutations by

$$\mathbb{E}[\text{HAF}_f] \approx \frac{\mu}{n} \sum_{k=2}^{n} k \cdot t_k = \frac{\mu / r}{n} \sum_{k=2}^{n} k \cdot (r \cdot t_k) = \frac{\theta / \alpha}{n} \sum_{k=2}^{n} k \cdot (r \cdot t_k).$$

(S32)

### Computing the mean HAF$_f$ score from a SNP matrix

For a sample of $n$ individuals whose genealogy has $q$ segregating sites, the SNP matrix $A$ consists of $n$ rows (each row representing one individual) and $q$ columns, with entry 0 denoting ancestral alleles and 1 denoting derived alleles.

The $v$th row of $A$ is a haplotype vector as depicted as $h$ in Fig. 1.

Let $w^\text{all}$ denote the sum of the rows of $A$. The frequency of the $j$th allele is $w_j^\text{all}$, the sum of all entries in the $j$th column of $A$.

In $\sum_{v=1}^{n} \text{HAF}_f(c(v))$, the $j$th mutation contributes $f(w_j)$ for each of the $w_j$ haplotypes that have that mutation ($A_{ej} = 1$) and contributes 0 on each of the $n - w_j$ others.

Thus, the average of HAF$_f$ over the genealogy is

$$\frac{1}{n} \sum_{v=1}^{n} \text{HAF}_f(c(v)) = \frac{1}{n} \sum_{j=1}^{q} f(w_j^\text{all}) \cdot w_j^\text{all}.$$  

(S33)

We use this to compute empirical averages of HAF$_f$ in simulations ($ms$ [73] outputs a SNP matrix rather than a tree), as well as in further theoretical developments in this Appendix. In particular, given a haplotype vector $c$ and the corresponding HAF vector $c$, we have

$$1 - \text{HAF}(c) = w^\text{all} \cdot h.$$  

(S34)
HAF score dynamics and peak values

Consider \( n \) haplotypes randomly sampled from a Wright-Fischer (WF) fixed-size population of \( N \) haploid individuals under a hard sweep with selection coefficient \( s \). Let \( \nu \) denote the fraction of individuals that carry the favored allele. We assume the sample has at least one carrier and at least one non-carrier, so that \( \nu \in \left\{ \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n} \right\} \); at fixation (\( \nu = 1 \)), there are no non-carriers remaining. We assume strong selection (\( Ns \gg 1 \)) and no recombination in the region being sampled. We measure time in generations going backwards. See Fig. [S7]. The current time is time 0; let \( T^{\text{car}} \) denote the time when all sampled carriers coalesce to their most recent common ancestor (denoted by MRCA\textsuperscript{car}). Let \( T^{\text{all}} \) denote the time when all sampled individuals coalesce to a common ancestor (denoted by MRCA\textsuperscript{all}). Let \( T(k) \) denote time to MRCA of \( k \) randomly chosen haplotypes in a population of size \( N \). From Nordborg [47], we have

\[
\mathbb{E}[T(k)] = 2N \left( 1 - \frac{1}{k} \right). \tag{S35}
\]

At time \( T^{\text{car}} \), we have exactly one ancestor of the favored allele carrier, and assume we have \( m \) ancestors of non-carriers. In a hard-sweep scenario, the favored mutation arises at the same time as the onset of selection; therefore, the time between \( T^{\text{car}} \) and \( T^{\text{all}} \) is governed by the neutral WF model with population \( N \), and is well approximated by neutral coalescent theory. Applying Eq. (S35) to the remaining sample of \( m + 1 \) individuals at time \( T^{\text{car}} \) gives the expected time to coalesce:

\[
\mathbb{E}[T^{\text{all}} - T^{\text{car}}] = \mathbb{E}[T(m + 1)] = 2N \left( 1 - \frac{1}{m + 1} \right). \tag{S36}
\]

Let \( A \) denote the SNP matrix of a sample of \( n \) individuals, with \( \nu n \) carriers of the favored allele (Fig. [S8]). We order the columns so that all mutations are ordered chronologically from left to right. Similarly, the rows are ordered so that the first \( \nu n \) rows correspond to carrier haplotypes.

The \( i \)\textsuperscript{th} haplotype is represented by the \( i \)\textsuperscript{th} row of \( A \) (denoted by row vector \( h_i \)) and has 1-HAF score denoted by \( 1\text{-HAF}_i \). From Eq. (S34), we have

\[
1\text{-HAF}_i = w^{\text{all}} \cdot h_i. \tag{S37}
\]

We partition the matrix into 3 submatrices: \( A_1 \) consists of the rows corresponding to carriers, and columns corresponding to mutations occurring prior to the onset of selection (times between \( T^{\text{car}} \) and \( T^{\text{all}} \)). \( A_2 \) is the submatrix from all non-carrier rows, and mutations prior to onset of selection. Finally, \( A_3 \) is the submatrix of all columns after the onset of selection (times between 0 and \( T^{\text{car}} \)). Define \( w^{\text{car}} \) as the sum of all carrier haplotypes; similarly, let \( w^{\text{non}} \) be the sum of all non-carrier haplotypes. Thus,

\[
w^{\text{all}} = w^{\text{car}} + w^{\text{non}}
\]

Also, we can partition the equation for computing \( 1\text{-HAF}_i \) by rows as

\[
1\text{-HAF}_i = w^{\text{car}} \cdot h_i + w^{\text{non}} \cdot h_i.
\]

We partition a haplotype vector \( h \) in the matrix by separating SNPs that occurred before (subvector \( h^b \)) and after (subvector \( h^a \)) the onset of mutation. A similar partitioning works for \( w^{\text{all}}, w^{\text{car}}, \) and \( w^{\text{non}} \). Thus, for a random haplotype \( h \),

\[
1\text{-HAF} = (w^{\text{car}})^b \cdot h^b + (w^{\text{non}})^b \cdot h^b + (w^{\text{all}})^a \cdot h^a.
\]

The expected 1-HAF score of a carrier (or non-carrier) haplotype \( h \) can similarly be decomposed as

\[
\mathbb{E}[1\text{-HAF}] = \mathbb{E}[(w^{\text{car}})^b \cdot h^b] + \mathbb{E}[(w^{\text{non}})^b \cdot h^b] + \mathbb{E}[(w^{\text{all}})^a \cdot h^a]. \tag{S38}
\]

To compute the expected 1-HAF score, we bound each of these constituent terms.
Figure S7. The coalescence of a sample of $n$ individuals to their most recent common ancestor MRCA\textsuperscript{all}, during a hard sweep. We assume that the current time has $\nu n$ carriers of the favored allele. These coalesce to MRCA\textsuperscript{car} in $T\textsuperscript{car}$ generations. From that point, the coalescence to MRCA\textsuperscript{all} is governed by neutral coalescent theory. $T(k)$ is time to MRCA of $k$ randomly chosen haplotypes in a neutrally evolving population.

Figure S8. Partitioning the SNP matrix $A$ of a sample of $n$ individuals.
Lemma 2 Consider a sample of \( n \) individuals under a hard sweep with \( \nu n \) carriers \((\frac{1}{n} \leq \nu \leq \frac{2 - m}{n})\). Let \( m \) be the number of non-carrier haplotypes remaining at time \( T^{\text{car}} \). Then for a random carrier haplotype \( h \), we have

\[
\mathbb{E}[(w^{\text{car}})^b \cdot h^b] = 2N\mu\nu n \left(1 - \frac{1}{m+1}\right),
\] (S38)

\[
\mathbb{E}[(w^{\text{non}})^b \cdot h^b] = 2N\mu(1 - \nu)n \left(\frac{1}{2} - \frac{1}{m+1}\right),
\] (S39)

while for a random non-carrier haplotype \( h \), these become

\[
\mathbb{E}[(w^{\text{car}})^b \cdot h^b] = 2N\mu\nu n \left(\frac{1}{2} - \frac{1}{m+1}\right),
\] (S40)

\[
\mathbb{E}[(w^{\text{non}})^b \cdot h^b] = 2N\mu(1 - \nu)n \left(\frac{1}{2} - \frac{m-1}{2(m+1)(1 - \nu)n}\right).
\] (S41)

Proof of Eq. (S38). Mutations that happen on the lineage from MRCA^{all} to MRCA^{car} are shared by all carriers, so all rows of matrix \( A_1 \) are identical. Thus, for all carrier haplotypes \( h_i, h_j \), we have \( h_i^b = h_j^b \). On restricting \( w^{\text{car}} = \sum_{i=1}^{\nu n} h_i \) to the mutations from before the onset of selection, we obtain

\[
\mathbb{E}[(w^{\text{car}})^b \cdot h^b] = \sum_{i=1}^{\nu n} \mathbb{E}[h_i^b \cdot h^b] = \nu n \mathbb{E}[h^b \cdot h^b].
\] (S42)

The term \( \mathbb{E}[h^b \cdot h^b] \) is the expected number of ones in \( h^b \), which is the expected number of mutations on the lineage from MRCA^{all} to MRCA^{car}. Therefore, \( \mathbb{E}[h^b \cdot h^b] = \mu \mathbb{E}[T^{\text{all}} - T^{\text{car}}] \). By Eq. (S36),

\[
\mathbb{E}[h^b \cdot h^b] = \mu \mathbb{E}[T^{\text{all}} - T^{\text{car}}] = 2N\mu \left(1 - \frac{1}{m+1}\right),
\] (S43)

Plugging this into Eq. (S42) gives Eq. (S38).

Proof of Eq. (S39). On restricting \( w^{\text{non}} = \sum_{i=\nu n+1}^{n} h_i \) to the mutations from before the onset of selection, we obtain

\[
\mathbb{E}[(w^{\text{non}})^b \cdot h^b] = \sum_{i=\nu n+1}^{n} \mathbb{E}[h_i^b \cdot h^b].
\] (S44)

Going back from time \( T^{\text{car}} \) to \( T^{\text{all}} \), we observe the coalescence of the ancestors of carrier haplotype \( h \) and non-carrier haplotype \( h_i \) at a time \( T' \), where \( T^{\text{car}} \leq T' \leq T^{\text{all}} \). The term \( \mathbb{E}[h_i^b \cdot h^b] \) is the expected number of ones in common between these two haplotypes among the SNPs occurring before the onset of selection. This equals the expected number of mutations in the lineage from \( T' \) to \( T^{\text{all}} \):

\[
\mathbb{E}[h_i^b \cdot h^b] = \mu \mathbb{E}[T^{\text{all}} - T'] = \mu \mathbb{E}[(T^{\text{all}} - T^{\text{car}}) - (T' - T^{\text{car}})] = \mu \left(\mathbb{E}[T(m+1)] - \mathbb{E}[T(2)]\right).
\] (S45)

Note that \( T' - T^{\text{car}} \) is the time for a sample of 2 individuals (ancestors of \( h \) and \( h_i \) at time \( T^{\text{car}} \)) to coalesce. We evaluate Eq. (S45) using Eq. (S35):

\[
\mathbb{E}[h_i^b \cdot h^b] = 2N\mu \left(\frac{1}{2} - \frac{1}{m+1}\right).
\] (S46)

Plug this into Eq. (S44). All \( n - \nu n = (1 - \nu)n \) terms are the same, so it simplifies to Eq. (S39).
Proof of Eq. (S40). We have
\[ E[(w_{\text{car}})^b \cdot h^b] = \sum_{i=1}^{\nu n} E[h_i^b \cdot h^b]. \] (S47)

Here, \( h \) is a non-carrier while \( h_i \) is a carrier, so they are distinct and they coalesce prior to the onset of selection. Thus, Eq. (S46) applies here. All \( \nu n \) terms of the sum in Eq. (S47) evaluate to Eq. (S46), so it simplifies to Eq. (S40). \( \square \)

Proof of Eq. (S41). We have
\[ E[(w_{\text{non}})^b \cdot h^b] = \sum_{i=\nu n+1}^{n} E[h_i^b \cdot h^b]. \] (S48)

The term \( E[h_i^b \cdot h^b] \) is the expected number of ones in common in haplotypes \( h \) and \( h_i \) among the SNPs occurring before the onset of selection.

We apply Lemma 3 (upcoming) to show that two random non-carrier haplotypes \( h \) and \( h_i \) have the same ancestor at time \( T^{\text{car}} \) with probability
\[ p = 1 - \left(1 + \frac{1}{(1-\nu)n}\right) \left(1 - \frac{2}{m+1}\right) \]
or different ancestors with probability \( 1 - p \). Thus, with probability \( p \), we have \( h_i^b = h^b \), and with probability \( 1 - p \), we have \( h_i^b \neq h^b \). Therefore,
\[ E[h_i^b \cdot h^b] = p \cdot E[h_i^b \cdot h^b \mid h_i^b = h^b] + (1 - p)E[h_i^b \cdot h^b \mid h_i^b \neq h^b]. \] (S49)

Term \( E[h_i^b \cdot h^b \mid h_i^b = h^b] \) is the expected number of ones in \( h^b \), which is \( \mu E[T^{\text{all}} - T^{\text{car}}] \).
Term \( E[h_i^b \cdot h^b \mid h_i^b \neq h^b] \) is the expected number of ones common between \( h_i^b \) and \( h^b \), which is \( \mu E[T^{\text{all}} - T'] \). These expected times were evaluated in Eqs. (S43) and (S45). Then
\[ E[h_i^b \cdot h^b] = p \mu E[T^{\text{all}} - T^{\text{car}}] + (1 - p)\mu E[T^{\text{all}} - T'] \]
\[ = \left(1 - \left(1 + \frac{1}{(1-\nu)n}\right) \left(1 - \frac{2}{m+1}\right)\right) 2N\mu \left(1 - \frac{1}{m+1}\right) \]
\[ + \left(1 + \frac{1}{(1-\nu)n}\right) \left(1 - \frac{2}{m+1}\right) 2N\mu \left(\frac{1}{2} - \frac{1}{m+1}\right) \]
\[ = 2N\mu \left(\frac{1}{2} - \frac{m-1}{2(m+1)(1-\nu)n}\right). \] (S50)

Eq. (S48) is a sum of \((1-\nu)n\) terms, each of form (S50), so Eq. (S48) evaluates to Eq. (S41). \( \square \)

In the preceding proof, the probability \( p \) that two non-carriers share the same ancestor was computed using the following lemma, applied to \( n_0 = (1-\nu)n \) non-carriers in the sample and \( m \) ancestors of non-carriers at time \( T^{\text{car}} \).

Lemma 3 Consider a subsample of \( n_0 \) individuals, which coalesce to \( m \) ancestors at some time \( T \), while the rest of the sample coalesces to one or more ancestors distinct from these \( m \). Draw two individuals (with replacement) from the subsample. The probability that they have the same ancestor at time \( T \) is
\[ \frac{2n_0 - m + 1}{n_0(m+1)} = 1 - \left(1 + \frac{1}{n_0}\right) \left(1 - \frac{2}{m+1}\right). \] (S51)
Proof. Throughout this proof, pairs of individuals \((x, y)\) are drawn from the subsample of size \(n_0\) with replacement. There are \(n_0^2\) such pairs.

Consider the \(m\) lineages at time \(T\). The number of pairs of individuals deriving from the \(i\)th lineage is \((w_{m,i})^2\). Sum over all lineages to obtain the total number of pairs of individuals with the same time \(T\) ancestor:

\[
\sum_{i=1}^{m} (w_{m,i})^2 = m \sum_{w=1}^{n_0-m-1} \left( \frac{(n_0-w-1)}{(m-1)} \right) w^2 = \frac{n_0(2n_0 - m + 1)}{m + 1}.
\]

Note that \(\Pr(W_{m,i} = w)\) is given by Eq. \((S4)\), and holds upon restricting the full genealogy to the subsample. Divide this expected value by \(n_0^2\) pairs and simplify to obtain Eq. \((S51)\).

Lemma 4 Consider a sample of \(n\) individuals under a hard sweep. For a random haplotype \(h\) (carrier or non-carrier), we have

\[
0 \leq E[\langle w^{all} \rangle^a \cdot h^a] \leq \mu n E[T^{car}] .
\]  

Proof. We have

\[
E[\langle w^{all} \rangle^a \cdot h^a] = \sum_{i=1}^{n} E[h_i^a \cdot h^a] .
\]  

The dimension of vectors \(h_i^a\) and \(h\) is the number of mutations in \([0, T^{car}]\). The expected value of this dimension is \(\mu E[T^{car}]\). The dot product of two binary vectors is at least 0, and is bounded above by their dimension, so

\[
0 \leq E[h_i^a \cdot h^a] \leq \mu E[T^{car}] .
\]

Plugging Eq. \((S54)\) into Eq. \((S53)\) gives Eq. \((S52)\).

Lemma 5 Consider a population of size \(N\) with selection coefficient \(s\), under a hard sweep with strong selection \((Ns \gg 1)\). Then

\[
0 \leq E[\langle w^{all} \rangle^a \cdot h^a] \leq \theta n \ln(2Ns) 2Ns .
\]  

Proof. Campbell \[79\] shows that under strong selection, the fixation time is approximately \(\frac{1}{s} \ln(2Ns)\). As \(T^{car}\) is the time for the favored allele to reach frequency \(\nu n\), it is dominated by the fixation time. Therefore,

\[
E[T^{car}] \leq \frac{1}{s} \ln(2Ns) .
\]

Plug Eq. \((S56)\) and the scaled mutation rate \(\theta = 2N\mu\) into Lemma 4 to obtain Eq. \((S55)\).

Lemma 6 Consider a sample of \(n\) individuals under a hard sweep with \(\nu n\) carriers \(\left(\frac{1}{n} \leq \nu \leq \frac{n-1}{n}\right)\). In the time \([0, T^{car}]\), let the number of coalescent events among the \((1-\nu)n\) non-carrier haplotypes be \(\varepsilon_\nu(1-\nu)n\). Then

\[
0 \leq \varepsilon_\nu \leq \varepsilon^{\max}_\nu \quad \text{where} \quad \varepsilon^{\max}_\nu = \frac{1}{1 + \frac{2Ns}{n\ln(2Ns)}} .
\]
Proof. Since we have \((1 - \nu)n\) non-carrier haplotypes in our sample, the expected number of non-carriers in the whole population is \((1 - \nu)N\). Let \(N_0^{\text{non}} = (1 - \nu)N\), and let \(N_u^{\text{non}}\) be the expected number of non-carriers in the whole population when the \(u\)th coalescent event occurs among the non-carrier samples. As we go backwards in time (increasing \(u\)), the expected number of non-carriers increases:

\[
N_u^{\text{non}} \geq N_{u-1}^{\text{non}} \geq N_0^{\text{non}} = (1 - \nu)N \quad \text{for} \quad 1 \leq u \leq \varepsilon \nu (1 - \nu)n. \tag{S58}
\]

Additionally, the chance of coalescing decreases and the expected coalescent time increases. For a fixed size population of size \(N\), the expected time for one coalescent event in a sample of size \(n\) is \(\frac{N}{n^2}\) (see [47]). Hence, the expected time for the \(u\)th coalescent event is at least

\[
\varepsilon \nu (1 - \nu)n \sum_{u=1}^{\varepsilon \nu (1 - \nu)n} \frac{2}{((1 - \nu)n - u + 1)((1 - \nu)n - u)}.
\]

On the other hand, the sum of all coalescent times during the selective sweep is at most \(T_{\text{car}}\). Therefore,

\[
\mathbb{E}[T_{\text{car}}] \geq \frac{2N\varepsilon \nu n}{(1 - \varepsilon \nu)n} \leq \ln(2Ns) s. \tag{S59}
\]

Comparing this lower bound on \(\mathbb{E}[T_{\text{car}}]\) with the upper bound in Eq. (S56) gives

\[
\frac{2N\varepsilon \nu n}{(1 - \varepsilon \nu)n} \leq \ln(2Ns) s. \tag{S60}
\]

This gives the upper bound on \(\varepsilon \nu\) in Eq. (S57). \[\blacksquare\]

**Theorem 7** Consider a sample of \(n\) individuals under a hard sweep with \(\nu n\) carriers \((\frac{1}{n} \leq \nu \leq \frac{n-1}{n})\). Assume strong selection \((Ns \gg 1)\). In the time \([0, T_{\text{car}}]\), let the number of coalescent events among the \((1 - \nu)n\) non-carrier haplotypes be \(\varepsilon \nu (1 - \nu)n\). The expected 1-HAF score of a random carrier haplotype is

\[
\mathbb{E}[1\text{-HAF}_{\text{car}}] \approx \theta n \left(\frac{\nu + 1}{2} - \frac{1}{(1 - \varepsilon \nu)(1 - \nu)n + 1}\right), \tag{S61}
\]

while for a random non-carrier haplotype,

\[
\mathbb{E}[1\text{-HAF}_{\text{non}}] \approx \theta n \left(\frac{1}{2} + \frac{1}{2n} - \frac{1 - \varepsilon \nu (1 - \nu)}{(1 - \varepsilon \nu)(1 - \nu)n + 1}\right). \tag{S62}
\]

Proof. Recall the decomposition of \(\mathbb{E}[1\text{-HAF}]\) given in Eq. (S37):

\[
\mathbb{E}[1\text{-HAF}] = \mathbb{E}[(w_{\text{car}})^b \cdot h^b] + \mathbb{E}[(w_{\text{non}})^b \cdot h^b] + \mathbb{E}[(w_{\text{all}})^b \cdot h^b]. \tag{S63}
\]

The number of ancestors of non-carriers at time \(T_{\text{car}}\) is \(m = (1 - \varepsilon \nu)(1 - \nu)n\).
For a random carrier haplotype $h$, we evaluate the first two terms in Eq. (S63) by using Eqs. (S38) and (S39), and bound the third term by using Lemma 5, to obtain:

$$0 \leq E[1-HAF_{\text{car}}] - \theta n \left( \frac{\nu + 1}{2} - \frac{1}{(1 - \nu)(1 - \nu)n + 1} \right) \leq \theta n \cdot \frac{\ln(2Ns)}{2Ns}.$$ 

For a random non-carrier haplotype $h$, we use Eqs. (S40) and (S41) and Lemma 5:

$$0 \leq E[1-HAF_{\text{non}}] - \theta n \left( \frac{1}{2} + \frac{1}{2n} \right) \leq \theta n \cdot \frac{\ln(2Ns)}{2Ns}.$$ 

Ignoring the small term $\frac{\ln(2Ns)}{2Ns}$, we obtain the desired expressions (S61) and (S62).

We find the maximum of Eq. (S61) over $0 \leq \nu \leq 1$ using differentiation; the result is

$$\max_{\nu} E[1-HAF_{\text{car}}] \approx \theta n \left( 1 - \frac{1}{\sqrt{2(1 - \nu)n}} \right)^2 \approx \theta n.$$ 

We simulated selective sweeps for a variety of parameters and compared their trajectories against these results. In Fig. S9, we compared the trajectories of both carriers and non-carriers in 500 selective sweeps for each pair $(\theta, n)$ with $\theta \in \{24, 48\}$, $n \in \{100, 200, 300, 400\}$, $s = 0.08$, and $N = 2000$. Over the course of a trajectory, the frequency $\nu$ of the favored allele varies from $\nu = 1/n$ (1 carrier) to $\nu = 1$ (fixation), but different trajectories will go through a different sequence of values of $\nu$. We aligned the trajectories by the values of $\nu$. For each $\nu$, we separated carriers and non-carriers and averaged together $(1-HAF)/(n\theta)$ over the generation (if any) with allele frequency $\nu$ in each sweep. We plotted these averaged trajectories in green ($\theta = 24$) and pink ($\theta = 48$). We compared these against the expected value (Theorem 7). The expected value is bounded above in blue ($\nu = 0$) and below in red ($\nu = \nu_{\text{max}}$); see Lemma 6. These are bounds on the expected value, not on simulation means. Simulation means may range over the whole distribution, but tend to vary around the expected value.

In Fig. S10, for each $(n, \theta)$, we compute average simulated values of $(1-HAF)/\theta$ over 500 trajectories, stratified by $\nu$ by the same method. The maximum value of this average over all $\nu$ is determined and plotted, and compared against the expected value (Eq. (S64)).

In this paper, due to the run time for forward simulations, we used a relatively small value for population size ($N = 2000$); in reality, $N$ is a lot larger (10,000), so that $\nu$ becomes smaller (see Lemma 6). Therefore, the upper and lower bounds in the proof of Theorem 7 move closer together, and we approximate $\nu = 0$. To simplify the presentation in the main text, we set $\nu = 0$, giving:

$$E[1-HAF_{\text{car}}] \approx \theta n \left( \frac{\nu + 1}{2} - \frac{1}{(1 - \nu)n + 1} \right),$$

$$E[1-HAF_{\text{non}}] \approx \theta n \left( \frac{1}{2} + \frac{1}{2n} - \frac{1}{(1 - \nu)n + 1} \right),$$

$$\max_{\nu} E[1-HAF_{\text{car}}] \approx \theta n \left( 1 - \frac{1}{\sqrt{2n}} \right)^2 \approx \theta n.$$
Figure S9. Dynamics of expected 1-HAF score during a selective sweep. For each \((\theta, n, \nu)\) with \(\theta \in \{24, 48\}, n \in \{100, 200, 300, 400\}, \nu \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\}, s = 0.08, \) and \(N = 2000,\) we did 1500 trials. We plotted the mean value of \((1 - \text{HAF})/(\theta n)\) as a function of \(\nu,\) for both carriers and non-carriers, and compared against the theoretical expected value. The expected value of \((1 - \text{HAF})/(n \theta)\) lies somewhere between the blue and red curves. The mean values may range over the whole distribution (and are not constrained by the blue and red curves) but tend to vary around the expected value.
Figure S10. HAF-peak as a function of $(\theta, n)$. For each $(\theta, n, \nu)$ with $\theta \in \{24, 48\}$, $n \in \{100, 200, 300, 400\}$, and $\nu \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\}$, we did 1500 trials. We averaged the values of $(1-\text{HAF})/\theta$ for carriers at each $\nu$ across the sweeps and determined the maximum over all $\nu$. Each thin curve shows the results for a fixed $\theta$ as $n$ varies, while the thick curves in red and blue give the theoretical lower and upper bounds on $\mathbb{E}[1-\text{HAF}^{\text{car}}]/\theta$. The solid red spots are the average over all trials over all $\theta$, for each $n$. 