PLOS | GENETICS

RESEARCH ARTICLE

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

Roy Ronen[1☯], Glenn Tesler[2☯], Ali Akbari[3☯], Shay Zakov[4], Noah A. Rosenberg[5], Vineet Bafna[4]*

1 Bioinformatics Graduate Program, University of California, San Diego, La Jolla, California, United States of America, **2** Department of Mathematics, University of California, San Diego, La Jolla, California, United States of America, **3** Department of Electrical & Computer Engineering, University of California, San Diego, La Jolla, California, United States of America, **4** Department of Computer Science & Engineering, University of California, San Diego, La Jolla, California, United States of America, **5** Department of Biology, Stanford University, Stanford, California, United States of America

☯ These authors contributed equally to this work.
* vbafna@ucsd.edu

OPEN ACCESS

Citation: Ronen R, Tesler G, Akbari A, Zakov S, Rosenberg NA, Bafna V (2015) Predicting Carriers of Ongoing Selective Sweeps without Knowledge of the Favored Allele. PLoS Genet 11(9): e1005527. doi:10.1371/journal.pgen.1005527

Editor: Graham Coop, University of California Davis, UNITED STATES

Received: April 6, 2015

Accepted: August 24, 2015

Published: September 24, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported in part by National Science Foundation grants CCF-1115206, IIS-1318386, DBI-1458557, and DBI-1458059. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

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 $\mathbf{c}$, 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 ℓ, we define the ℓ-HAF *score* of $\mathbf{c}$ as:

$$\ell\text{–HAF}(\mathbf{c}) = \sum_j \mathbf{c}_j^{\ell} \tag{1}$$

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 ℓ-HAF score is equivalent to the ℓ-norm of $\mathbf{c}$ raised to the ℓ^{th} power, or $(\|\mathbf{c}\|_\ell)^\ell$. We will show that during a selective sweep, the HAF score of a haplotype serves as a proxy to its relative fitness.

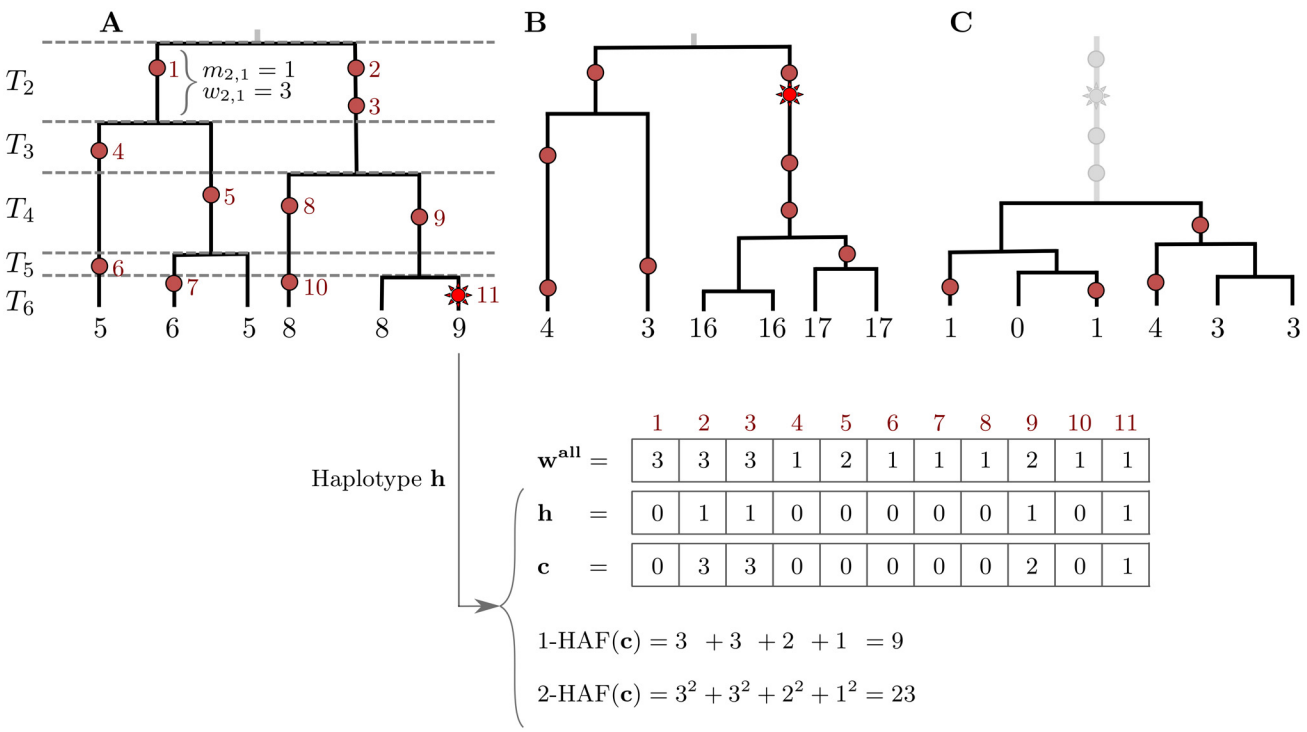

Fig 1. The HAF score. Genealogies of three samples ($n = 6$) progressing through a selective sweep, from left to right. Neutral mutations are shown as red circles, and are numbered in red; the favored allele is shown as a red star. The 1-HAF score of each haplotype is shown below its corresponding leaf, 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 $\mathbf{w}^{all}$ 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.

doi:10.1371/journal.pgen.1005527.g001

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 [17, 33]. 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 [18, 20, 24]. Finally, as the site-frequency spectrum recovers to its neutral state due to new mutations and drift [23], 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 [34–36]. 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 [37, 38], as well as for distinguishing between soft and hard sweeps [39–41]. 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 haploid population of size N. Let μ 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 $\alpha = 2$ N_0 r the population-scaled growth rate. Let ρ denote the population-scaled recombination rate. In our theoretical calculations, we assume no recombination ($\rho = 0$), and we derive expressions for the general ℓ-HAF score. We use simulations to demonstrate the concordance of theoretical and empirical values of the ℓ-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 ℓ-HAF score, we primarily use 1-HAF in the applied sections. Applications of ℓ-HAF with $\ell > 1$ will be explored in future work.

Expected ℓ-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 $\mathbf{c}(v)$ denote the HAF vector $\mathbf{c}$ for the v^{th} haplotype ($v \in \{1, \ldots, n\}$). Let ξ_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 ξ_w sites of frequency w contributes w^ℓ to the ℓ-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 ℓ-HAF scores of all n haplotypes in the sample is

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

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

$$\mathbb{E}[\ell\text{−HAF}] = \frac{1}{n}\sum_{w=1}^{n-1} \mathbb{E}[\xi_w] \cdot w^\ell \cdot w = \frac{\theta}{n}\sum_{w=1}^{n-1} w^\ell\,. \tag{3}$$

The first two cases (ℓ = 1,2) yield

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

Expected ℓ-HAF score, variable population size. Our derivation of expected ℓ-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 ℓ-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 haplotype back to the sample MRCA. We can write the ℓ-HAF score of the haplotype as

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

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 $\mathbf{c}(v)$ denote $\mathbf{c}$ (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

$$\begin{aligned}\sum_{v=1}^{n} \ell\text{−HAF}(\mathbf{c}(v)) &= \sum_{k=2}^{n}\sum_{v=1}^{n} m_k(v)(w_k(v))^{\ell} \\ &= \sum_{k=2}^{n}\sum_{i=1}^{k}\sum_{j=1}^{w_{k,i}} m_{k,i}(w_{k,i})^{\ell} = \sum_{k=2}^{n}\sum_{i=1}^{k} m_{k,i}(w_{k,i})^{\ell+1}.\end{aligned} \tag{6}$$

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 ℓ-HAF score of a randomly chosen haplotype as

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

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

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

and set $w^{(0)} = 1$. We show in S1 Text that

$$\mathbb{E}[(W_{k,i})^{(\ell)}] = \ell!\cdot\frac{n^{(\ell)}}{k^{(\ell)}}\,. \tag{9}$$

We have $w^{(1)} = w$ and $w^{(2)} = w(w+1) = w^2+w$, so $w^2 = w^{(2)} - w^{(1)}$, which leads to:

$$\mathbb{E}[(W_{k,i})^2] = \mathbb{E}[(W_{k,i})^{(2)} - (W_{k,i})^{(1)}] = \frac{2n(n+1)}{k(k+1)} - \frac{n}{k} = \frac{n(2n-k+1)}{k(k+1)}\,. \tag{10}$$

In S1 Text, 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 ℓ-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 S1 Text for details). In the conditional expectation method, we compute the expected value of T_k conditioned on T_{k+1}, . . ., T_n, as follows (in the order $k = n, n-1, \ldots, 2$):

$$\begin{aligned} t_k &= \mathbb{E}[T_k \mid T_{k+1} = t_{k+1}, \ldots, T_n = t_n] \\ &= \frac{1}{r}\int_0^1 \ln\left(1 - \frac{\alpha}{k(k-1)}\, e^{-r\cdot\tau_k}\, \ln(u)\right) du \\ &= \frac{1}{r}\exp\left(\frac{k(k-1)}{\alpha}\, e^{r\cdot\tau_k}\right) E_1\left(\frac{k(k-1)}{\alpha}\, e^{r\cdot\tau_k}\right), \end{aligned} \tag{11}$$

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

We then use $\mathbb{E}[(W_{k,i})^\ell]$ (evaluated in Eq. (S21) in S1 Text) to evaluate Eq (7), yielding $\mathbb{E}[\ell$-HAF] for exponential population growth as

$$\mathbb{E}[\ell\text{−HAF}] \approx \frac{\theta}{\alpha}\sum_{k=2}^{n}(r\cdot t_k)\sum_{q=0}^{\ell}(-1)^{\ell-q}S(\ell,q)\left((q+1)!\,\frac{(n+1)^{(q)}}{(k+1)^{(q)}} - q\cdot q!\cdot\frac{(n+1)^{(q-1)}}{(k+1)^{(q-1)}}\right), \tag{12}$$

where $S(\ell, q)$ denotes the Stirling number of the second kind [45, Ch. 6.1]. We describe these procedures fully in S1 Text.

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 $N = 20000$ (see 'Simulations' in Methods). For example, for $\theta = 48$ and $n = 200$, the expected 1-HAF is exactly 4776.0 (Eqs (3) and (7)), whereas the empirically observed mean 1-HAF score of 20000 simulated samples is 4786 ± 3956 (sample mean ± sample standard deviation). Interestingly, the estimates improve when simulating with recombination, with an observed mean of 4780 ± 1684 (S1 Fig).

We also modeled exponential growth in population size using the scaled growth rate α, using the conditional expectation method in Eq (12) (see S1 Text). 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 20000 simulations is 128 ± 131.1 for $\rho = 0$, and 127.6 ± 127.4 for $\rho = 25$ (S2 Fig).

We compared the simulations with theoretical expected ℓ-HAF scores for multiple values of $\ell \in \{1, 2, 3, 4\}$ and different choices of the population-genetic parameters: scaled mutation rate $\theta \in \{24, 48\}$, scaled growth rate $\alpha \in \{0, 30, 60, 80\}$), and scaled recombination rate $\rho \in \{0, 25, 50\}$ (see 'Simulations' in Methods). While theoretical expected values were computed assuming $\rho = 0$, S3 Fig shows the concordance between theoretical and empirical means for each choice of parameters. The concordance improves slightly for increasing values of n, ℓ. In each case, the values are robust to choice of ρ, and the variance even reduces slightly for higher ρ (S3D Fig).

As ℓ increases, the normalized HAF score (ℓ-HAF$^{1/\ell}$) distribution (S4 Fig) becomes more left-skewed and has generally smaller values (upper bound of range approaching $n - 1$), with reduced variance. Increasing ℓ increases the relative weight of ancient mutations. As an extreme example, the normalized ∞-HAF score is simply the weight of the highest frequency mutation on the haplotype, and not very informative. However, very recent mutations, including those that appear post-selection among the carriers of the favored allele add 'noise' to the

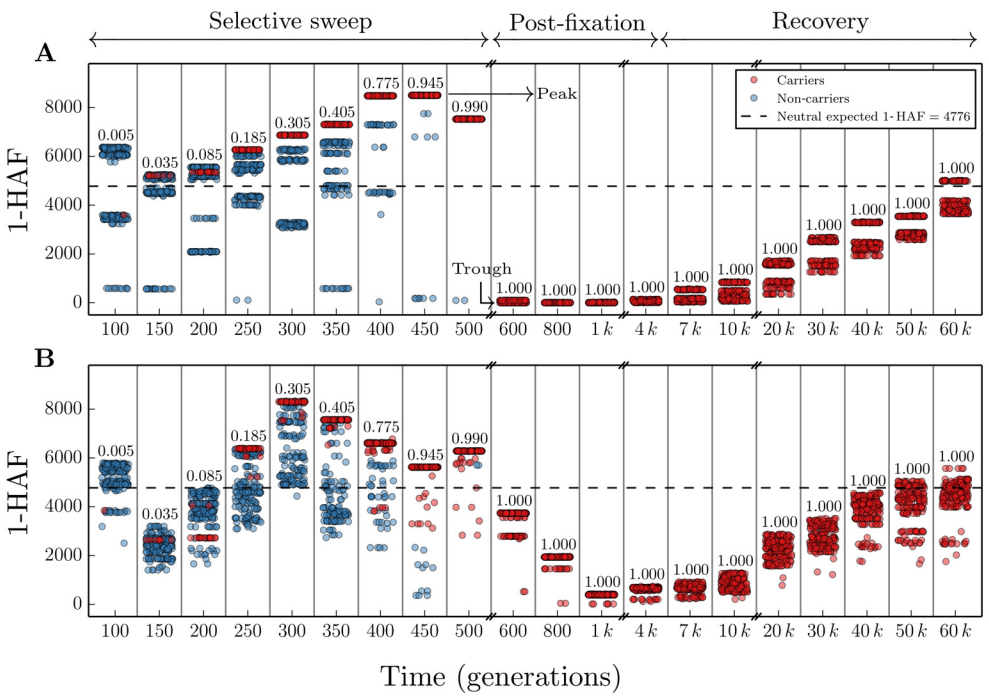

Fig 2. Schematic of HAF score dynamics. We consider HAF scores in 50 kb segments, examining $n = 200$ haplotypes sampled from a constant-sized ($N = 20000$ haploids) population, evolving with population-scaled mutation rate $\theta = 48$ and selection coefficient $s = 0.05$. We do forward simulations, 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. The number in each rectangle is the frequency of the favored allele (carriers). 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. Darker shades of red or blue indicate a higher density of points at that level. The dotted line represents the expected 1-HAF score in the neutral population. (A) Simulation of a non-recombining segment. (B) Simulation with population-scaled recombination rate $\rho = 25$ (see Methods).

doi:10.1371/journal.pgen.1005527.g002

HAF-score, and an appropriate choice of $\ell > 1$ may perform better for some applications. We will explore this in future work.

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. Fig 2 illustrates the HAF score dynamics in a single simulated population undergoing a hard sweep, with selection coefficient $s = 0.05$. 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 100–450), carrier HAF scores increase to a peak value (HAF-peak). Soon after fixation (time $\sim$450), we observe a sharp decline in HAF scores (HAF-trough), followed by slow and steady recovery due to new mutation and drift (times 500–50000). We observe similar behavior for the HAF score dynamics in an exponentially growing population, and soft sweep scenarios (S5 and S6 Figs). Though soft sweeps can arise under different circumstances, we restrict our attention to soft sweeps arising from standing

variation. While the behavior is similar, we note that during a soft sweep, the HAF scores do not have as sharp a decline as in the hard sweep scenarios.

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 $s \in [0.005, 0.040]$ (see 'Simulations' in Methods), and compared it to 1-HAF values in simulated neutral populations growing exponentially at rates $\alpha \in [100, 600]$. Fig 3A shows a close similarity between the 1-HAF values under exponential growth (blue) and the selective sweep trough (red).

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 μ 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 v denote the fraction of carrier haplotypes in the sample. When $v \leq 1/n$ (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 [46]. The expected 1-HAF scores for carriers and non-carriers are identical (Eq (3)). At the time when v first equals 1, there are no non-carriers, and the HAF-scores are given by the exponential growth model. In S1 Text, we model the 1-HAF scores for all intermediate values of v.

Let 1-HAF^{car} (respectively, 1-HAF^{non}) denote the 1-HAF score of a random haplotype carrying the favored allele (respectively, a non-carrier) when a fraction v of the n sampled haplotypes carry the favored allele. In S1 Text, we show that under strong selection ($Ns \gg 1$) and no recombination ($\rho = 0$),

$$\mathbb{E}[1-\text{HAF}^{\,\text{car}}] \approx \theta n\left(\frac{v+1}{2} - \frac{1}{(1-v)n+1}\right), \tag{13}$$

$$\mathbb{E}[1-\text{HAF}^{\,\text{non}}] \approx \theta n\left(\frac{1}{2} + \frac{1}{2n} - \frac{1}{(1-v)n+1}\right). \tag{14}$$

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

$$\max_v \mathbb{E}[1-\text{HAF}^{\,\text{car}}(v,n)] = \theta n\left(1 - \frac{1}{\sqrt{2n}}\right)^2 \approx \theta n\,. \tag{15}$$

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 $v = 1$ is first reached).

Empirical validation. Simulated data under selective sweeps with coefficients $s \in [0.005, 0.040]$ show that for strong selection ($Ns \gg 1$) (i) the pre-fixation HAF peak scores appear to

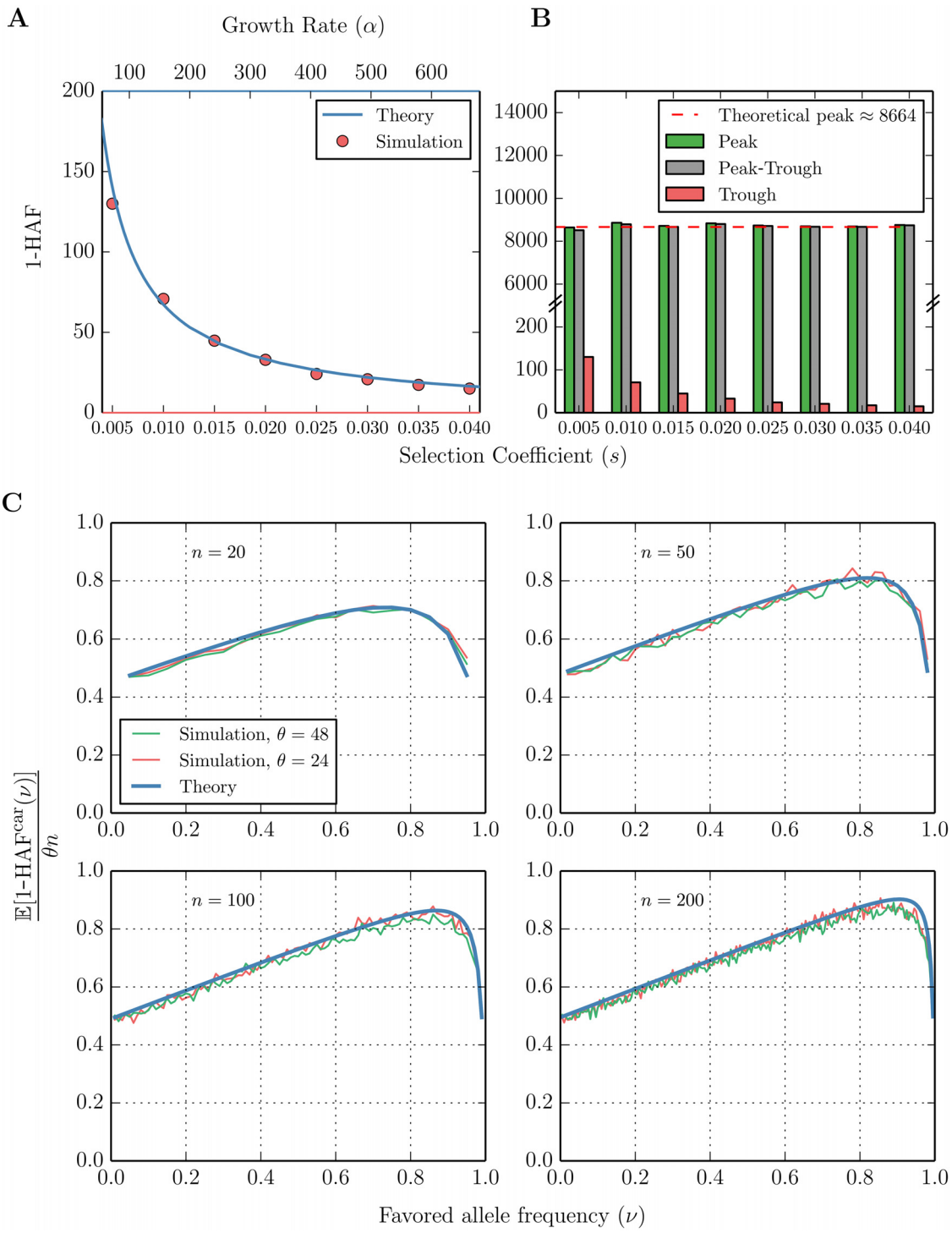

Fig 3. HAF scores in a selective sweep 'peak' and 'trough'. (A) Observed values (red) of the mean 'trough' 1-HAF scores in simulated selective sweeps with coefficients $s \in [0.005, 0.040]$. Theoretical values (blue) of expected 1-HAF scores under exponential population growth with population-scaled rates $\alpha \in [100, 600]$ given by Eq (12). Simulated 1-HAF scores (red) represent the mean of 2000 simulated population samples for each value of s, with $\theta = 48$, $n = 200$. (B) Observed mean 1-HAF peak, trough, and difference (peak minus trough) for selective sweeps with coefficients $s \in [0.005, 0.040]$. 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$^{\text{car}}$ (1-HAF score of haplotypes carrying the favored allele) plotted as a function of the fraction of carriers (ν) in the sample during a selective sweep. For each (θ, n, ν) with $\theta \in \{24, 48\}$, $n \in \{20, 50, 100, 200\}$, $\nu \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\}$, $s = 0.01$, and $N = 20000$, we plotted the mean value of (1-HAF$^{\text{car}}$)/(θn) over 1000 trials, and compared against the theoretical values (Eq (13)).

doi:10.1371/journal.pgen.1005527.g003

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 θn. We also simulated $(1\text{-HAF}^{car})/(n\theta)$ as a function of v (Fig 3C and S15 Fig). 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, HAF scores of haplotypes in a region under selection might help predict the future MRCA of a population. Finally, by 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 last application, leaving the first two 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 $\geq 70\%$) 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 S7 Fig.

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 ℓ-HAF score is computed ($\ell = 1$ by default). We then fit a Gaussian Mixture Model

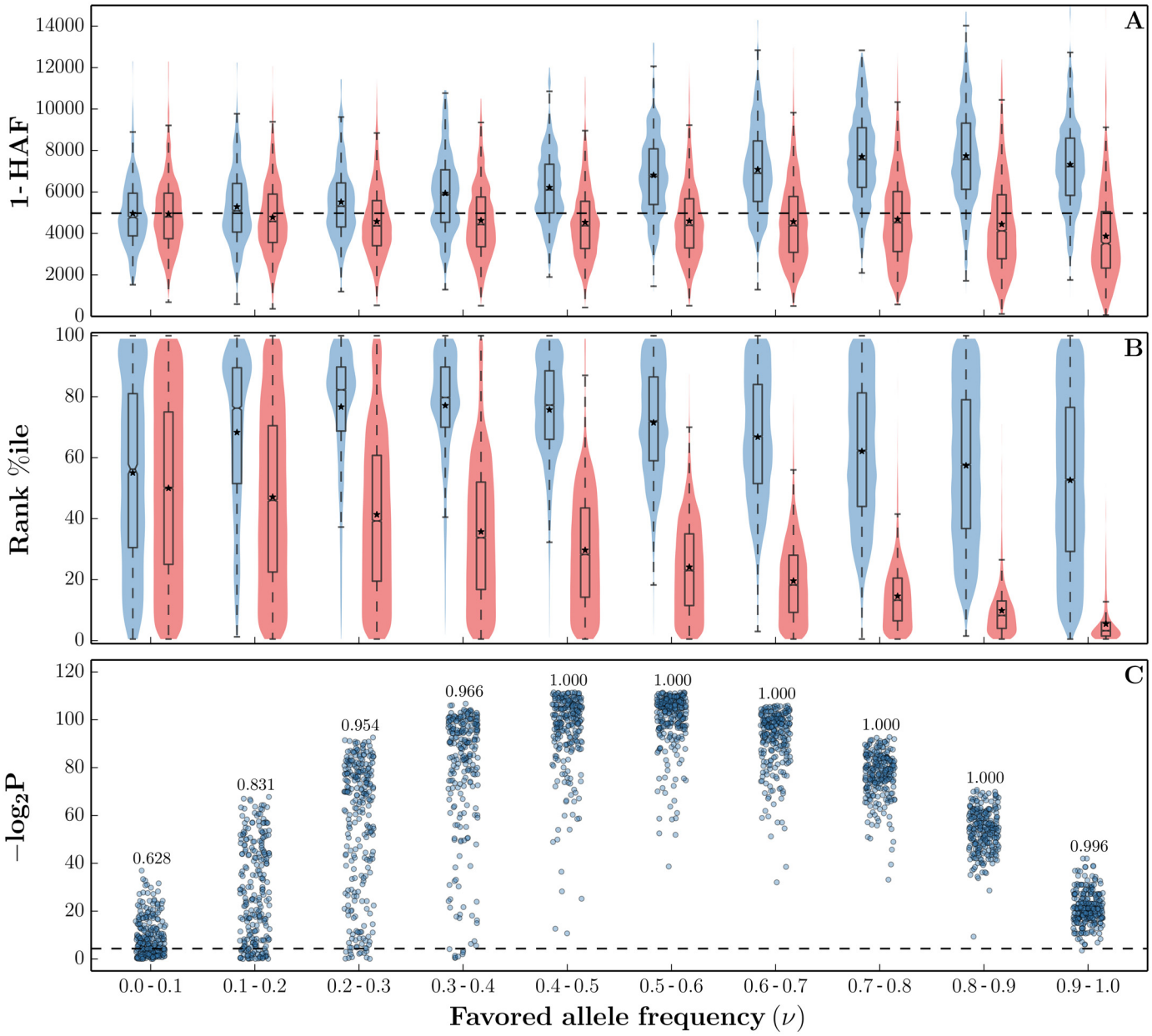

Fig 4. HAF score dynamics in ongoing selective sweeps. HAF scores were computed from 250 simulated population samples ($n = 200$) undergoing a hard sweep ($\theta = 48$, $\rho = 25$, $s = 0.01$), using the simulation software *msms* [47]. (A) Each violin shows the Gaussian kernel density estimation (KDE) 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 to mark the 25th, 50th, and 75th percentiles, with means indicated by asterisks. The horizontal dashed line represents the expected 1-HAF scores under neutrality (Eq (4)). (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 dashed line).

doi:10.1371/journal.pgen.1005527.g004

(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 Guassian 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 might in fact be 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*

$$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{FP+TN}\right), \tag{16}$$

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

While there are no tools currently available that directly predict the carrier state of a haplotype, some approaches are relevant. For example, Grossman et al. [49] developed a 'composite of multiple signals' (CMS) statistic to reduce the number of candidates for the favored mutation, but CMS cannot directly be used to identify carriers of the favored mutation. Similarly, the iHS statistic uses the dominant haplotype frequency decay in a window centered around each locus, as a test for recent positive selection [30]. As a comparison, we used iHS to distinguish carriers from non-carriers based on segregating alleles at the locus with peak iHS score. The balanced accuracy of PreCIOSS on hard sweeps is shown in Fig 5A for a specific choice of parameters (200 samples with $n = 200$, $\theta = 48$, $\rho = 25$, $s = 0.01$). Once the sweep reaches frequencies above 30%, the balanced accuracy increases (median $\sim$70%) and remains high (median $\sim$90%) for the remainder of the sweep. At the beginning 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 reduced for soft sweeps (Fig 5B, run with similar parameters), as increasing the carrier haplotype frequency leads to higher variance in 1-HAF scores.

We tested PreCIOSS under a wide range of population-genetic parameters (S1 Table), and observed consistently high balanced accuracy in carrier-state prediction as the sweeps progressed (S8 Fig). Specifically, PreCIOSS is quite robust to changes in sample size (S8A–S8D Fig). A higher recombination rate has only a limited impact (S8A and S8H Fig), while setting $\rho = 0$ shows reduced performance at an early stage of the sweep (S8A and S8G Fig). This is consistent with selection acting more efficiently in the presence of recombination.

We tested the effect of the position of the carrier mutation (unknown to PreCIOSS) on the performance of PreCIOSS. We considered different 50 kb windows, with the carrier mutation located at one end (0 kb), and moving towards the middle (25 kb). For each location of the carrier, we simulated 200 samples with $n = 200$, $\theta = 48$, $\rho = 25$, $s = 0.01$ (S9 Fig), but did not observe a marked change in accuracy. However, when the favored allele is in the middle of the window, the median balanced accuracy is generally higher and has lower variance (S10 Fig).

Finally, we tested PreCIOSS on a popular model of European demography [50]. The model (S11 Fig) suggests an Out-of-Africa migration 51 kya (51 thousand years ago), followed by a European and East Asian split 23 kya. It also suggests bottlenecks that reduced the effective population sizes of the European ($N_{\text{Eu0}} = 1032$), and East-Asian ($N_{\text{As0}} = 550$) populations, and exponential growth in the populations following the bottleneck events. We simulated populations based on this model, as well as selection events (hard sweep) at different times after the Out-of-Africa migration, and partitioned all samples into two categories depending on whether the selection event happened before or after the bottleneck. These scenarios are challenging for most tests of adaptation (see, e.g., [23]). However, there are still significant differences in the 1-HAF scores of carriers and non-carriers. The balanced accuracy of PreCIOSS is shown in Fig 6A for ancient selection and Fig 6B for recent (after bottleneck) selection. The performance is

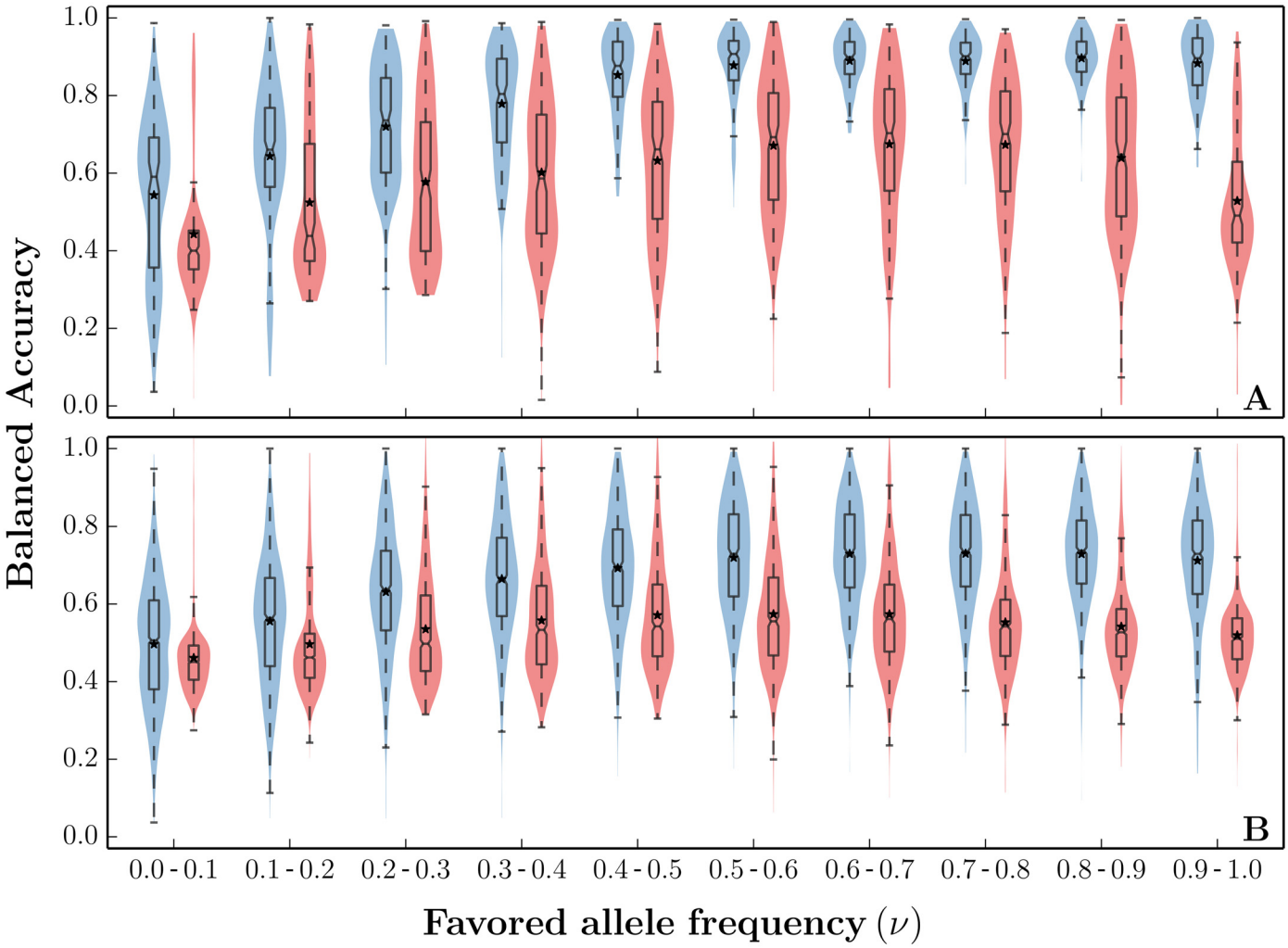

Fig 5. Predicting carriers of hard and soft sweeps. Balanced accuracy (Eq (16)) of PreCIOSS in populations undergoing hard and soft sweeps. For each frequency bin, (A) 200 samples were simulated ($n = 200$, $\theta = 48$, $\rho = 25$) undergoing a hard sweep ($s = 0.01$, $v_0 = 1/20000$), and (B) 200 samples were simulated undergoing a soft sweep ($s = 0.01$, $v_0 = 0.02$). We split each sweep into intervals as v progresses ([0.0, 0.1] through [0.9, 1.0]). For each v interval, we show the distribution of balanced accuracy using standard violin plots (blue). For comparison, we also plotted the balanced accuracy of iHS adapted to predicting carrier haplotypes (red).

doi:10.1371/journal.pgen.1005527.g005

quite robust, although somewhat worse in the early stages of the sweep. Once the favored allele frequency reaches 60%, the median accuracy is at 0.9. The accuracy is improved for recent adaptation, compared to ancient adaptation. Even for very recent sweeps, where the carrier frequency is 30–40%, the median balanced accuracy is close to 0.8. We used a lower selection coefficient for ancient selection compared to recent selection to ensure that we have sufficient cases of incomplete sweeps. Not surprisingly, the performance of PreCIOSS is worse for ancient selection compared to recent selection.

Our results suggest that for cases of recent adaptation (e.g., lactase adaptation, shown in Fig 7A, which happened between 2 kya and 20 kya and rapidly spread to high frequencies in the European population), PreCIOSS would show good performance in separating the carriers and non-carriers.

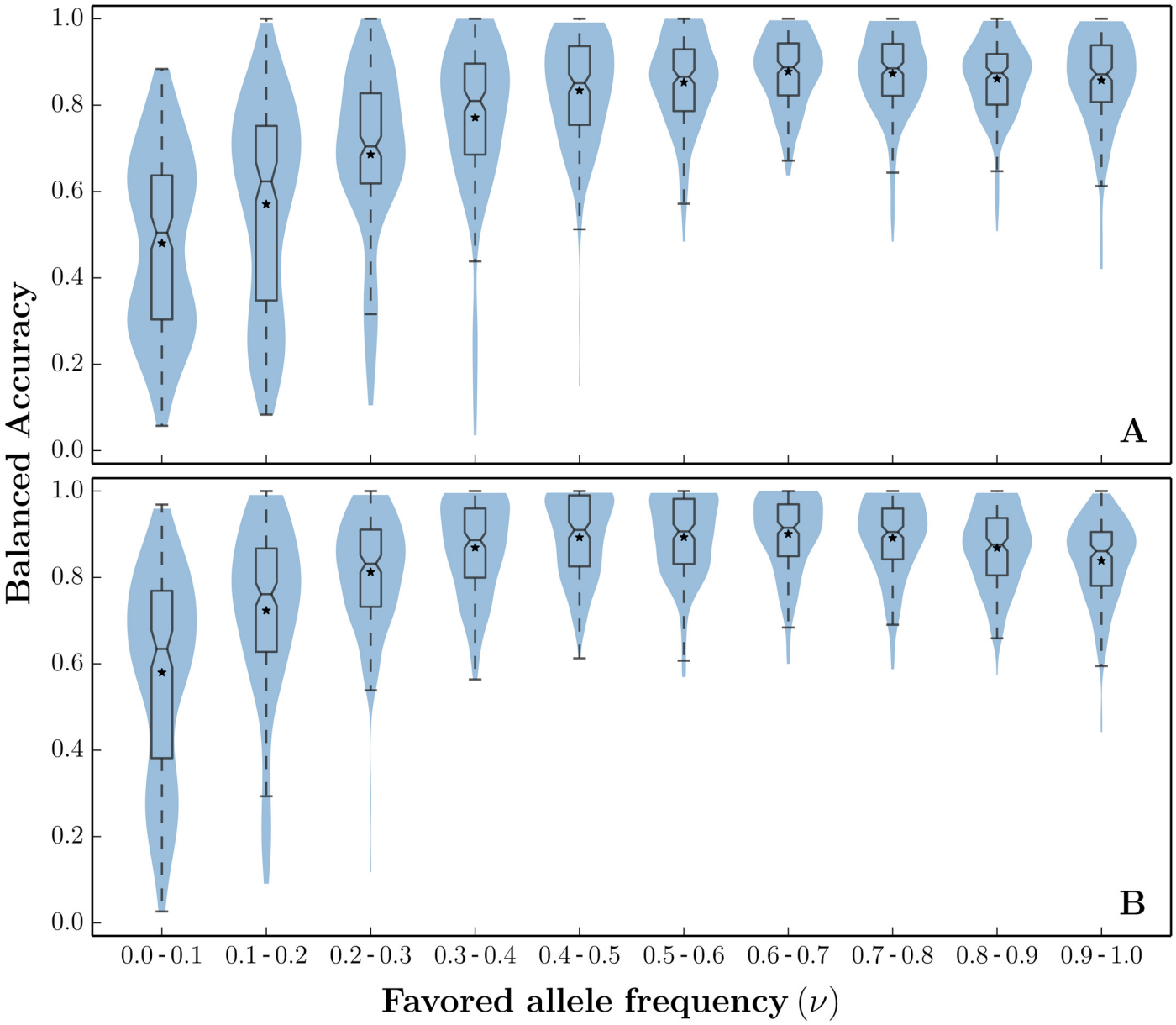

Fig 6. Balanced accuracy of PreCIOSS on a model of European demography. Populations were simulated for a popular model of human demography (S11 Fig and Gravel et al. (2011) [50]). The onset times of selection were separated into (A) pre-bottleneck (51 kya–23 kya) and (B) post-bottleneck (23 kya–current) epochs, with 10000 start times in each bin. All samples were simulated with $n = 200$, $\theta = 48$, $\rho = 25$. Samples were simulated with selection coefficient $s = 0.005$ in the pre-bottleneck epoch and $s = 0.02$ in the post-bottleneck epoch.

doi:10.1371/journal.pgen.1005527.g006

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

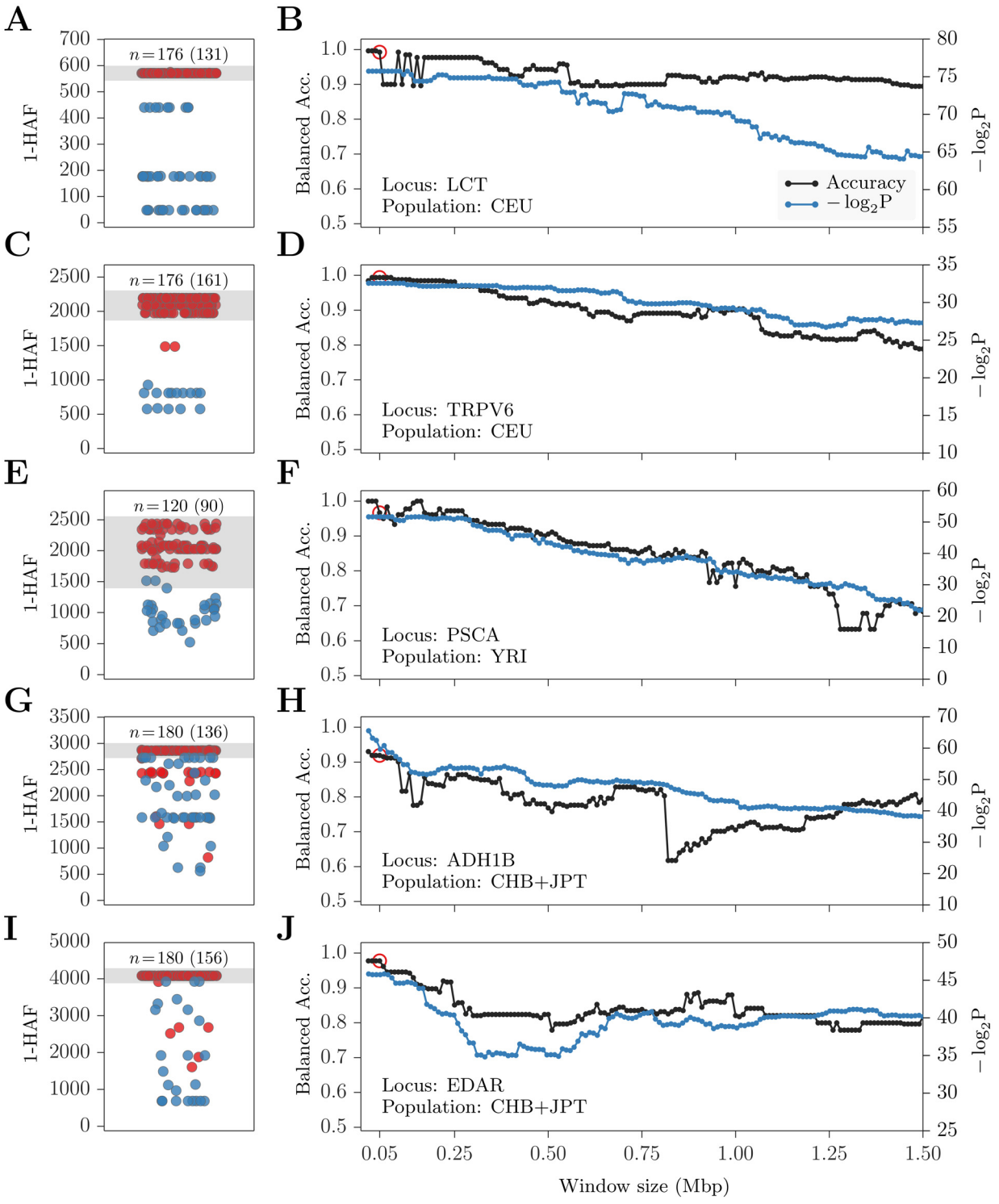

Fig 7. Predicting carriers of well-known selective sweeps. (Left): Haplotype 1-HAF scores in a 50 kb window centered at known favored sites indicated by the gene name, and the SNP identifier. (A) LCT/rs4988235, (C) TRPV/rs4987682, (E) PSCA/rs2294008, (G) ADH1B/rs1229984, and (I) EDAR/rs3827760. Points 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.

doi:10.1371/journal.pgen.1005527.g007

carried the favored allele. We use phased haplotypes from the HapMap project [51], setting the ancestral allele to that observed in orthologous Chimpanzee sequence [52].

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 [53]. T-13910 was further shown to be causal by in-vitro analysis, where it was found to increase enhancer activity [54, 55]. We considered haplotypes from the CEU population, applying PreCIOSS to a 50 kb window centered at C/T-13910 (Fig 7A). This yielded 100% accuracy in classifying carriers from non-carriers. Increasing the window size above 50 kb, the balanced classification accuracy reduced to $\sim$90% (Fig 7B). 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 7B, blue line). Despite this highly significant separation, the classification accuracy is initially unstable, alternating between $\sim$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 7A). 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 [56, 57]. 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 [58]. Applying PreCIOSS to a 50 kb window centered at this site, we obtain $\sim$99% balanced classification accuracy (Fig 7C), which gradually decays to $\sim$80% when considering a 1.5 Mb window (Fig 7D). As with LCT, the separation in HAF scores between carriers and non-carriers of the allele is highly statistically significant ($P < 10^{-7}$; see Fig 7D). 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 7C).

PSCA. Prostate Stem Cell Antigen (PSCA) has been proposed to be under selection in a global analysis of allele frequency differentiation [59]. The putative causal site is a non-synonymous SNP (rs2294008) known to be involved in several cancer types [60, 61]. 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 $\sim$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 7E). Unlike LCT and TRPV6, accuracy decays more noticeably with distance, reaching 63% at 1.25 Mb (Fig 7F). This sharper decay in accuracy is even more pronounced when considering the sweep in the CHB population (S12 Fig). 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) [59], 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 [62]. These genes have also been suggested to cause the "alcohol flush" phenotype common in Asian populations [63]. 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 [64, 65], and (ii) the estimated age of the allele coincides with rice domestication [63, 66] and the availability of fermented beverages [67]. 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 7G). Both accuracy and statistical significance (of class separation) gradually decay with increasing window size (Fig 7H). 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 [68, 69]. Specifically, a non-synonymous SNP (rs3827760, V370A) has been associated with these phenotypes [70]. The SNP is located within a DEATH-domain, which is highly conserved in mammals [71], and has been experimentally confirmed (in vitro) to increase EDAR activity [70]. It is found at very high frequencies in East Asian and American populations, while being completely absent from Europeans and Africans [70]. The EDAR gene has been found to be under selection in multiple studies [30, 56, 72], 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 7A–7F). 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 (3) and (12)). However, we do not yet know the variance. This quantity would provide a better understanding of the respective distributions, and a means to 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 7). 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 7. The clusters capturing the carriers in LCT and EDAR have tightly distributed HAF scores (Fig 7A and 7I). The cluster for ADH1B (Fig 7G) has more variance by comparison, and the variance increases for TRPV6 (Fig 7C) and PSCA (Fig 7E), 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 [73]. 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 [74]. 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 and sweep samples were generated using the simulator *msms* [47]. All simulations generated samples of $n \in [20, 400]$ haplotypes from a larger effective population of $N = 20000$ haplotypes, each of length 50 kb. A mutation rate of approximately $\mu = 2.4 \cdot 10^{-8}$ mutations per bp per generation was used [75, 76]. For our simulations, we choose a population-scaled mutation rate $\theta \in \{24, 48\}$. For human recombination events, a population scaled rate of $\rho = 1.32\theta$ has been proposed (e.g. [77]). We use simulations either with no recombination, or with $\rho \in \{25, 50\}$ in a 50 kb region to approximate human rates.

For exponential growth, we used $N = 20000$ 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 $\alpha = 2Nr$. We set α to a range of values in [200, 1600].

For selective sweeps, we used forward simulations 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 $s \in [0.005, 0.050]$ and heterozygosity 0.5 (haploid carriers get half the fitness advantage of diploid carriers). When s is 'low' ($0.001 \leq s \leq 0.01$), the available tests do not detect a selective sweep with reasonable power [21, 23]. Selection with $s \geq 0.08$ 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 ($s \in [0.005, 0.050]$) 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 ($v_0 \in [1/N, 1]$), and drifting at the onset of selection. In our simulations, we set v_0 at

the beginning of the sweep to $v_0 = 1/N = 5 \cdot 10^{-5}$ for hard sweeps and $v_0 \in \{0.001, 0.02\}$ for soft sweeps, corresponding to 20–400 carrier haplotypes at the onset of selection.

In comparing the performance of PreCIOSS against iHS, we used the software *selscan* [78] to compute iHS scores.

To investigate the performance of HAF scores on human populations, we used a popular demographic model (S11 Fig) with parameters suggested from Gravel et al. (2011) [50]. Among the different properties, the model assumes an out-of-Africa migration at 51 kya, and a European, East Asian split 23 kya. The European Asian split was accompanied by a bottleneck event that reduced the effective population sizes of the European and Asian populations, and was followed by an exponential growth in these populations. We used *msms* to simulate populations according to this model.

In modeling selection, we partitioned the onset of selection into two epochs: 'pre-bottleneck' events between 51 kya and 23 kya, and 23 kya, and 'post-bottleneck' epoch between 23 kya, and the current generation. For each epoch, we picked 10000 times for onset of selection chosen uniformly from the time interval, and performed forward simulations with a sample size of $n = 200$. Samples were chosen during the sweeps, and partitioned according to carrier allele frequency, with 200 samples randomly chosen for each bin. Samples in the pre-bottleneck epoch were simulated with $s = 0.005$ to reduce the chance of fixation, and $s = 0.02$ in the post-bottleneck epoch. The balanced accuracy measurements were done independently for the two epochs.

Data preprocessing

We downloaded pre-phased haplotype data from the HapMap [51] project website. Both HapMap 3 [51] and HapMap 2 [79] 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 [52] 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/.

Supporting Information

S1 Fig. HAF scores in neutrally evolving constant-sized populations. The distribution of 4×10^6 1-HAF scores aggregated from 20000 population samples (each of $n = 200$ haplotypes) simulated under a standard coalescent model without recombination. Plugging the simulation parameters $\theta = 48$, $n = 200$ into Eqs (3) or (7) give an expected 1-HAF score of 4776. The observed mean 1-HAF score is 4786 ± 3956 with no recombination ($\rho = 0$), and 4780 ± 1684 with $\rho = 25$ (blue line).
(TIF)

S2 Fig. HAF scores in neutrally evolving exponentially growing populations. The distribution of 4×10^6 1-HAF scores aggregated from 20000 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 128.0 with $\rho = 0$ (red line), and 127.4 with $\rho = 25$ (blue line).
(TIF)

S3 Fig. HAF scores for a range of simulation parameters. Each empirical test is the average of 1000 trials. (A) Empirical mean and theoretical expected ℓ-HAF scores for a fixed size population ($\ell \in \{1, 2, 3, 4\}$, $\theta \in \{24, 48\}$, $\rho = 0$). (B) Empirical mean and theoretical expected 1-HAF scores for an exponentially growing population ($\alpha \in \{0, 30, 60, 80\}$, $\theta \in \{24, 48\}$, $\rho = 0$). (C) Theoretical expected 1-HAF scores (computed assuming $\rho = 0$) compared against empirical means of 1-HAF scores from samples with different recombination rates ($\rho \in \{0, 25, 50\}$, $\theta \in \{24, 48\}$). (D) Interestingly, higher recombination rates reduce the variance in 1-HAF estimates. In the three green curves for $\theta = 24$ (and in the three red curves for $\theta = 48$), the variation from the expected value (blue) decreases as ρ increases. Rate $\rho = 0$ (dotted) has the most variation; $\rho = 25$ (dashed) has less; and $\rho = 50$ (solid) has the least. The theoretical values are based on (A) Eqs (3) and (S22), (B) Eq (12), and (C, D) Eq (4).
(TIF)

S4 Fig. Distribution of normalized ℓ-HAF scores (ℓ-HAF$^{1/\ell}$). Results are based on simulated samples of size $n = 200$ drawn from a larger population size of neutrally evolving haploid population with $N = 20000$ ($\theta = 48$, $\rho = 0$, $\alpha = 0$). The green line marks the sample mean of the ℓ^{th} root of ℓ-HAF, while the red dashed line marks the ℓ^{th} root of the sample mean of ℓ-HAF. The latter matches the blue dotted line, which marks the theoretically computed value of $(\mathbb{E}[\ell\text{-HAF}])^{1/\ell}$, using Eq. (S22). As ℓ increases, the high frequency mutations dominate the normalized ℓ-HAF score. The distribution becomes more left-skewed and has generally smaller values (upper bound of range approaching $n - 1$), with reduced variance.
(TIF)

S5 Fig. Schematic of HAF score dynamics in an exponentially growing population with current population size $N = 20000$, population-scaled growth rate $\alpha = 80$, and population-scaled mutation rate $\theta = 48$. The population is under selection with $s = 0{:}05$. See Fig 2 for an explanation of the conventions used.
(TIF)

S6 Fig. Schematic of HAF score dynamics in a population undergoing a soft sweep due to standing variation with $\nu_0 = 0{:}002$. Samples were simulated with $\theta = 48$, $n = 200$, $s = 0{:}05$, and $\rho \in \{0, 25\}$. See Fig 2 for an explanation of the conventions used.
(TIF)

S7 Fig. 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 box plot is overlaid on each violin. The horizontal dotted line represents the neutral expected value. At each time point, HAF scores were computed from 1000 simulations with *msms* [47], each with $n = 200$ haplotypes undergoing a hard sweep, with parameters $N = 20000$, $\theta = 48$, $\rho = 25$, $n = 200$. At each time point, box plots marking 25$^{\text{th}}$, 50$^{\text{th}}$, and 75$^{\text{th}}$ percentiles were computed for the 1000×200 HAF scores, with an asterisk marking the mean.
(TIF)

S8 Fig. Predicting carriers of hard sweeps. Balanced accuracy of PreCIOSS in populations undergoing hard sweeps. Balanced accuracy is shown for each allele frequency bin as a

standard box plot computed over 200 samples for each frequency bin, and each parameter set in S1 Table.
(TIF)

S9 Fig. Balanced accuracy with different positions of favored allele. In each panel, 200 samples were simulated ($N = 20000$, $n = 200$, $\theta = 48$, $\rho = 25$) while undergoing a hard sweep ($s = 0.01$) in a 50 kb window. Each panel shows balanced accuracy for a different position of the favored allele within the window, as the position varies from 0 to 25 kb.
(TIF)

S10 Fig. Balanced accuracy variation with different positions of favored allele: summary statistics. In each case, 5000 samples were simulated ($N = 20000$, $n = 200$, $\theta = 48$, $\rho = 25$) while undergoing a hard sweep ($s = 0.01$) in a 50 kb window. The mean, median and standard deviation of balanced accuracy of PreCIOSS was measured with the favored allele at the start of the window (0 kb, in blue) and at the middle of the window (25 kb, in red).
(TIF)

S11 Fig. A model of human demography described by Gravel et al. (2011) [50, Fig 4, Table 2]. The model assumes an out-of-Africa split at time T_B, with a bottleneck that reduced the effective population from N_{Af} to N_B, allowing for migrations at rate $m_{\text{Af-B}}$. The African population stays constant at N_{Af} up to the present generation. The model assumes a second split between European and Asian populations at time T_{EuAs}, with a bottleneck reducing the Asian and European populations to N_{As0} and N_{Eu0} respectively. The bottleneck was followed by exponential growth at rates r_{As} and r_{Eu}, as well as migrations among all three sub-populations, leading to current populations from which Asian (CHB+JPT), European (CEU), and Africans (YRI) individuals were sampled.
(TIF)

S12 Fig. 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 7 for further details on the conventions used.
(TIF)

S13 Fig. The coalescence of a sample of n individuals to their most recent common ancestor MRCA^{all}, during a hard sweep. We assume that the current time has vn carriers of the favored allele. These coalesce to MRCA^{car} in T^{car} generations. From that point, the coalescence to MRCA^{all} is governed by neutral coalescent theory. $T(k)$ is time to MRCA of k randomly chosen haplotypes in a neutrally evolving population.
(TIF)

S14 Fig. Partitioning the SNP matrix A of a sample of n individuals.
(TIF)

S15 Fig. Dynamics of expected 1-HAF score during a selective sweep. For each (θ, n, v) with $\theta \in \{24, 48\}$, $n \in \{100, 200, 300, 400\}$, $v \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\}$, $s = 0.08$, and $N = 2000$, we did 1500 trials. We plotted the mean value of (1-HAF)/(θn) as a function of v, for both carriers and non-

carriers, and compared against the theoretical expected value. The expected value of (1-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.
(TIF)

S1 Table. Simulation parameter sets used for generating S8 Fig. In simulations B through E, we changed one parameter (in boldface) at a time vs. simulation A.
(PDF)

S1 Text. Mathematical derivations of HAF score expected values, peak scores, and dynamics.
(PDF)

S1 Source Code. PreCIOSS. ZIP file with source code for PreCIOSS. See the project website for the latest version: http://bix.ucsd.edu/projects/precioss/
(ZIP)

Author Contributions

Conceived and designed the experiments: RR GT AA SZ VB. Performed the experiments: RR GT AA. Analyzed the data: RR GT AA VB. Contributed reagents/materials/analysis tools: RR GT AA. Wrote the paper: RR GT AA NAR VB. Developed the mathematical framework: RR GT AA SZ VB. Developed software tools: RR GT AA.

References

1. Fu W, Akey JM. Selection and adaptation in the human genome. Annu Rev Genomics Hum Genet. 2013; 14:467–489. doi: 10.1146/annurev-genom-091212-153509 PMID: 23834317
2. Lachance J, Tishkoff SA. Population Genomics of Human Adaptation. Annu Rev Ecol Evol Syst. 2013 Nov; 44:123–143. doi: 10.1146/annurev-ecolsys-110512-135833 PMID: 25383060
3. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013; 47:97–120. doi: 10.1146/annurev-genet-111212-133526 PMID: 24274750
4. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005 Nov; 15(11):1566–1575. doi: 10.1101/gr.4252305 PMID: 16251466
5. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 2009 May; 19(5):826–837. doi: 10.1101/gr.087577.108 PMID: 19307593
6. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res. 2010 Mar; 20(3):393–402. doi: 10.1101/gr.100545.109 PMID: 20086244
7. Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS Genet. 2014 Aug; 10(8): e1004412. doi: 10.1371/journal.pgen.1004412 PMID: 25102153
8. Jeong C, Di Rienzo A. Adaptations to local environments in modern human populations. Curr Opin Genet Dev. 2014 Dec; 29C:1–8. doi: 10.1016/j.gde.2014.06.011
9. Tekola-Ayele F, Adeyemo A, Chen G, Hailu E, Aseffa A, Davey G, et al. Novel genomic signals of recent selection in an Ethiopian population. Eur J Hum Genet. 2014 Nov; advance online publication. doi: 10.1038/ejhg.2014.233 PMID: 25370040
10. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science. 2010; 329(5987):75–78. Available from: http://www.sciencemag.org/content/329/5987/75.abstract. doi: 10.1126/science.1190371 PMID: 20595611
11. Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, et al. Genetic evidence for high-altitude adaptation in Tibet. Science. 2010 Jul; 329(5987):72–75. doi: 10.1126/science.1189406 PMID: 20466884

12. Scheinfeldt LB, Soi S, Thompson S, Ranciaro A, Woldemeskel D, Beggs W, et al. Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biol. 2012; 13(1):R1. doi: 10.1186/gb-2012-13-1-r1 PMID: 22264333

13. Alkorta-Aranburu G, Beall CM, Witonsky DB, Gebremedhin A, Pritchard JK, Di Rienzo A. The genetic architecture of adaptations to high altitude in Ethiopia. PLoS Genet. 2012; 8(12):e1003110. doi: 10.1371/journal.pgen.1003110 PMID: 23236293

14. Huerta-Sanchez E, Degiorgio M, Pagani L, Tarekegn A, Ekong R, Antao T, et al. Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. Mol Biol Evol. 2013 Aug; 30(8): 1877–1888. doi: 10.1093/molbev/mst089 PMID: 23666210

15. Udpa N, Ronen R, Zhou D, Liang J, Stobdan T, Appenzeller O, et al. Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. Genome Biol. 2014 Feb; 15(2): R36. doi: 10.1186/gb-2014-15-2-r36 PMID: 24555826

16. Zhou D, Udpa N, Ronen R, Stobdan T, Liang J, Appenzeller O, et al. Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in Andean highlanders. Am J Hum Genet. 2013 Sep; 93(3):452–462. doi: 10.1016/j.ajhg.2013.07.011 PMID: 23954164

17. Kaplan NL, Hudson RR, Langley CH. The "hitchhiking effect" revisited. Genetics. 1989 Dec; 123(4): 887–899. PMID: 2612899

18. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974 Feb; 23(1):23–35. doi: 10.1017/S0016672300014634 PMID: 4407212

19. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989 Nov; 123(3):585–595. PMID: 2513255

20. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000 Jul; 155:1405–1413. PMID: 10880498

21. Pavlidis P, Jensen JD, Stephan W. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics. 2010 Jul; 185(3):907–922. doi: 10.1534/genetics.110.116459 PMID: 20407129

22. Lin K, Li H, Schlotterer C, Futschik A. Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics. 2011 Jan; 187(1):229–244. doi: 10.1534/genetics.110.122614 PMID: 21041556

23. Ronen R, Udpa N, Halperin E, Bafna V. Learning natural selection from the site frequency spectrum. Genetics. 2013 Sep; 195(1):181–193. doi: 10.1534/genetics.113.152587 PMID: 23770700

24. Simonsen KL, Churchill GA, Aquadro CF. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics. 1995 Sep; 141(1):413–429. PMID: 8536987

25. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 1995 Jun; 140(2):783–796. PMID: 7498754

26. Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ. Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster. Genetics. 1994 Apr; 136(4):1329–1340. PMID: 8013910

27. Depaulis F, Mousset S, Veuille M. Haplotype tests using coalescent simulations conditional on the number of segregating sites. Mol Biol Evol. 2001 Jun; 18(6):1136–1138. doi: 10.1093/oxfordjournals.molbev.a003885 PMID: 11371602

28. Innan H, Zhang K, Marjoram P, Tavare S, Rosenberg NA. Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. Genetics. 2005 Mar; 169(3):1763–1777. doi: 10.1534/genetics.104.032219 PMID: 15654103

29. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002 Oct; 419(6909):832–837. doi: 10.1038/nature01140 PMID: 12397357

30. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006 Mar; 4(3):e72. doi: 10.1371/journal.pbio.0040072 PMID: 16494531

31. Toomajian C, Hu TT, Aranzana MJ, Lister C, Tang C, Zheng H, et al. A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome. PLoS Biol. 2006 May; 4(5):e137. doi: 10.1371/journal.pbio.0040137 PMID: 16623598

32. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007 Oct; 449(7164):913–918. doi: 10.1038/nature06250 PMID: 17943131

33. Kim Y, Stephan W. Selective sweeps in the presence of interference among partially linked loci. Genetics. 2003 May; 164(1):389–398. PMID: 12750349

34. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol (Amst). 2013 Nov; 28(11):659–669. doi: 10.1016/j.tree.2013.08.003

35. Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005 Apr; 169(4):2335–2352. doi: 10.1534/genetics.104.036947 PMID: 15716498

36. Pennings PS, Hermisson J. Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol. 2006 May; 23(5):1076–1084. doi: 10.1093/molbev/msj117 PMID: 16520336

37. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol. 2014 May; 31(5):1275–1291. doi: 10.1093/molbev/msu077 PMID: 24554778

38. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. 2015 Feb; 11(2):e1005004. doi: 10.1371/journal.pgen.1005004 PMID: 25706129

39. Peter BM, Huerta-Sanchez E, Nielsen R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet. 2012; 8(10):e1003011. doi: 10.1371/journal.pgen.1003011 PMID: 23071458

40. Schrider DR, Mendes FK, Hahn MW, Kern AD. Soft Shoulders Ahead: Spurious Signatures of Soft and Partial Selective Sweeps Result from Linked Hard Sweeps. Genetics. 2015 Feb; advance online publication.

41. Wilson BA, Petrov DA, Messer PW. Soft selective sweeps in complex demographic scenarios. Genetics. 2014 Oct; 198(2):669–684. doi: 10.1534/genetics.114.165571 PMID: 25060100

42. Fu YX. Statistical properties of segregating sites. Theor Popul Biol. 1995 Oct; 48(2):172–197. doi: 10.1006/tpbi.1995.1025 PMID: 7482370

43. Hudson RR. Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J, editors. Oxford Surveys in Evolutionary Biology. Oxford: Oxford University Press; 1990. p. 1–44.

44. Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 1991 Oct; 129(2):555–562. PMID: 1743491

45. Graham R, Knuth DE, Patashnik O. Concrete Mathematics: A Foundation for Computer Science. 2nd ed. Reading, Mass: Addison-Wesley; 1994.

46. Nordborg M. Coalescent Theory. In: Balding DJ, Bishop M, Cannings C, editors. Handbook of statistical genetics. 3rd ed. John Wiley & Sons, Ltd; 2008. p. 843–877.

47. Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics. 2010 Aug; 26(16):2064–2065. doi: 10.1093/bioinformatics/btq322 PMID: 20591904

48. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. In: Pattern Recognition (ICPR), 2010 20th International Conference on; 2010. p. 3121–3124.

49. Grossman SR, Shlyakhter I, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science. 2010 Feb; 327(5967):883–886. doi: 10.1126/science.1183863 PMID: 20056855

50. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci USA. 2011 Jul; 108(29):11983–11988. doi: 10.1073/pnas.1019276108 PMID: 21730125

51. Altshuler DM, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep; 467(7311):52–58. doi: 10.1038/nature09298 PMID: 20811451

52. Sequencing TC, Consortium A. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005 Sep; 437(7055):69–87. doi: 10.1038/nature04072

53. Kuokkanen M, Enattah NS, Oksanen A, Savilahti E, Orpana A, Jarvela I. Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. Gut. 2003 May; 52(5):647–652. doi: 10.1136/gut.52.5.647 PMID: 12692047

54. Olds LC, Sibley E. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Hum Mol Genet. 2003 Sep; 12(18):2333–2340. doi: 10.1093/hmg/ddg244 PMID: 12915462

55. Troelsen JT, Olsen J, Møller J, Sjöström H. An upstream polymorphism associated with lactase persistence has increased enhancer activity. Gastroenterology. 2003 Dec; 125(6):1686–1694. doi: 10.1053/j.gastro.2003.09.031 PMID: 14724821

56. Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, et al. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2004 Oct; 2(10):e286. doi: 10.1371/journal.pbio.0020286 PMID: 15361935

57. Stajich JE, Hahn MW. Disentangling the effects of demography and selection in human history. Mol Biol Evol. 2005 Jan; 22(1):63–73. doi: 10.1093/molbev/msh252 PMID: 15356276

58. Akey JM, Swanson WJ, Madeoy J, Eberle M, Shriver MD. TRPV6 exhibits unusual patterns of polymorphism and divergence in worldwide populations. Hum Mol Genet. 2006 Jul; 15(13):2106–2113. doi: 10.1093/hmg/ddl134 PMID: 16717058

59. Bhatia G, Patterson N, Pasaniuc B, Zaitlen N, Genovese G, Pollack S, et al. Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am J Hum Genet. 2011 Sep; 89(3):368–381. doi: 10.1016/j.ajhg.2011.07.025 PMID: 21907010

60. Sakamoto H, Yoshimura K, Saeki N, Katai H, Shimoda T, Matsuno Y, et al. Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. Nat Genet. 2008 Jun; 40(6):730–740. doi: 10.1038/ng.152 PMID: 18488030

61. Wu X, Ye Y, Kiemeney LA, Sulem P, Rafnar T, Matullo G, et al. Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. Nat Genet. 2009 Sep; 41(9): 991–995. doi: 10.1038/ng.421 PMID: 19648920

62. Whitfield JB. Alcohol dehydrogenase and alcohol dependence: variation in genotype-associated risk between populations. Am J Hum Genet. 2002 Nov; 71(5):1247–1250. doi: 10.1086/344287 PMID: 12452180

63. Peng Y, Shi H, Qi XB, Xiao CJ, Zhong H, Ma RL, et al. The ADH1B Arg47His polymorphism in east Asian populations and expansion of rice domestication in history. BMC Evol Biol. 2010; 10:15. doi: 10.1186/1471-2148-10-15 PMID: 20089146

64. Osier MV, Pakstis AJ, Soodyall H, Comas D, Goldman D, Odunsi A, et al. A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. Am J Hum Genet. 2002 Jul; 71(1):84–99. doi: 10.1086/341290 PMID: 12050823

65. Eng MY, Luczak SE, Wall TL. ALDH2, ADH1B, and ADH1C genotypes in Asians: a literature review. Alcohol Res Health. 2007; 30(1):22–27. PMID: 17718397

66. Li H, Mukherjee N, Soundararajan U, Tarnok Z, Barta C, Khaliq S, et al. Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia. Am J Hum Genet. 2007 Oct; 81(4):842–846. doi: 10.1086/521201 PMID: 17847010

67. McGovern PE, Zhang J, Tang J, Zhang Z, Hall GR, Moreau RA, et al. Fermented beverages of pre- and proto-historic China. Proc Natl Acad Sci USA. 2004 Dec; 101(51):17593–17598. doi: 10.1073/pnas.0407921102 PMID: 15590771

68. Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, Tsunoda T, et al. A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. Hum Genet. 2008 Sep; 124(2):179–185. doi: 10.1007/s00439-008-0537-1 PMID: 18704500

69. Kimura R, Yamaguchi T, Takeda M, Kondo O, Toma T, Haneji K, et al. A common variation in EDAR is a genetic determinant of shovel-shaped incisors. Am J Hum Genet. 2009 Oct; 85(4):528–535. doi: 10.1016/j.ajhg.2009.09.006 PMID: 19804850

70. Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M, et al. Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. PLoS ONE. 2008; 3(5):e2209. doi: 10.1371/journal.pone.0002209 PMID: 18493316

71. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007 Oct; 449(7164):913–918. doi: 10.1038/nature06250 PMID: 17943131

72. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante 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. doi: 10.1073/pnas.0502300102 PMID: 15905331

73. Luksza M, Lassig M. A predictive fitness model for influenza. Nature. 2014 Mar; 507(7490):57–61. doi: 10.1038/nature13087 PMID: 24572367

74. Lee MC, Lopez-Diaz FJ, Khan SY, Tariq MA, Dayn Y, Vaske CJ, et al. Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing. Proc Natl Acad Sci USA. 2014 Nov; 111(44):E4726–4735. doi: 10.1073/pnas.1404656111 PMID: 25339441

75. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000 Sep; 156(1):297–304. PMID: 10978293

76. Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, Han L, et al. Estimating the human mutation rate using autozygosity in a founder population. Nat Genet. 2012 Nov; 44(11):1277–1281. doi: 10.1038/ng.2418 PMID: 23001126

77. Hey J, Wakeley J. A coalescent estimator of the population recombination rate. Genetics. 1997 Mar; 145(3):833–846. PMID: 9055092

78. Szpiech ZA, Hernandez RD. selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. Mol Biol Evol. 2014 Oct; 31(10):2824–2827. doi: 10.1093/molbev/msu211 PMID: 25015648

79. Frazer KA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct; 449(7164):851–861. doi: 10.1038/nature06258 PMID: 17943122