Identifying the favored mutation in a positive selective sweep

Ali Akbari1, Joseph J Vitti2,3, Arya Iranmehr1, Mehrdad Bakhtiar1, Pardis C Sabeti2,3, Siavash Mirarab1 & Vineet Bafna4

Most approaches that capture signatures of selective sweeps in population genomics data do not identify the specific mutation favored by selection. We present iSAFE (for “integrated selection of allele favored by evolution”), a method that enables researchers to accurately pinpoint the favored mutation in a large region (~5 Mbp) by using a statistic derived solely from population genetics signals. iSAFE does not require knowledge of demography, the phenotype under selection, or functional annotations of mutations.

Human genetic data have revealed a multitude of genomic regions believed to be evolving under positive selection. Methods for detecting regions under selection from genetic variations exploit a variety of genomic signatures: allele-frequency-based methods analyze the distortion in site frequency spectra, linkage-disequilibrium-based methods use extended homozygosity in haplotypes, other methods use differences in allele frequency between populations, and finally, composite methods combine multiple test scores to improve resolution. Currently, a lack of rare (singleton) mutations was used to detect very recent selection. The signature of a selective sweep can be captured even when standing variation or multiple de novo mutations create a ‘soft’ sweep of distinct haplotypes carrying the favored mutation. When paired with deep sequencing data, these methods have identified multiple regions believed to be under selection, and can provide a window into genetic adaptation and evolution and improve the overall understanding of diseases. For example, adaptation to chronic hypoxia at high altitude can suggest targets for cardiovascular and other ischemic diseases.

However, the regions encompassed by the selective sweep can be very large (up to a few megabases), making it difficult to pinpoint the favored mutation in a selective sweep. Grossman et al. note that different selection signals identify distinct but overlapping regions, and a composite of multiple signals (CMS) can localize the site of the favored mutation. An alternative strategy is to rank single-nucleotide polymorphisms (SNPs) on the basis of their functional annotations. However, the signal of selection is often spread over regions of up to 1–2 Mbp on either side, and the high linkage disequilibrium makes it difficult to pinpoint the favored mutation. Here we propose a method, iSAFE, that exploits coalescent-based signals in ‘shoulders’ of the selective sweep (genomic regions proximal to the region under selection, but carrying the selection signal) to rank all mutations within a large (5-Mbp) region on the basis of their contribution to the selection signal (https://github.com/alek0991/iSAFE and Supplementary Software).

The haplotype allele frequency (HAF) score is a haplotypic score that can be used to separate carrier haplotypes from non-carriers without knowledge of the favored mutation. Using properties of the HAF score, we developed a SAFE score (Online Methods, Supplementary Note 1, Fig. 1, and Supplementary Fig. 1) that tends to be maximized for the favored mutation in a small region (50 kbp), but shows decaying performance when larger regions are investigated. To address the more general case of large regions (~5 Mbp) under selection, we developed the iSAFE score, which uses a two-step procedure to identify the favored variant. In the first step, the best candidate mutations in small (low-recombination) windows are identified on the basis of the SAFE score. Then, the SAFE scores of all variants over all windows are combined to assign an iSAFE score to each variant in the large region (Online Methods, Supplementary Note 1, and Fig. 2a,b).

The main alternatives to iSAFE are CMS and selection detection by conditional coalescent tree (SCCT). CMS combines statistics from different selection tests, including the integrated haplotype score (iHS), so as to localize the signal. To develop a unified probabilistic model, CMS requires control populations as input, as well as demographic models, and cannot be used run using only the SNP matrix. Therefore, we first compared SAFE with iHS and SCCT in simulations. The median SAFE rank of the favored mutation in a 50-kbp region was 1 out of ~250 variants (Fig. 1c, left), and the favored mutation ranked among the top 5 in 91% of simulations. In comparison, the median ranks of iHS and SCCT were 6 and 3, respectively. SAFE-score performance is robust to a large range of parameter choices (Supplementary Fig. 2). However, in testing with increasing window sizes, we observed that the median rank increased for regions larger than 80 kbp, perhaps because of the confounding signal at the shoulders of the selective sweep (Fig. 2c).

iSAFE, unlike SAFE, is specifically designed to exploit signal from the shoulders of the sweep (Online Methods and Fig. 2a,b). iSAFE showed consistently high performance as the window size
Not surprisingly, iSAFE performance deteriorated when the favored mutation was fixed or near fixation (favored allele frequency $\nu > 0.9$ in Supplementary Fig. 3). To handle this special case, we included individuals from non-target populations, using a specific protocol (Online Methods). With this inclusion, performance remained unchanged for $\nu < 0.9$ and dramatically improved for high frequencies, including when the favored mutation was fixed in the target population (Supplementary Fig. 3). We also tested iSAFE against CMS, using a model of human demography. Although CMS showed excellent performance in localizing the favored mutation, iSAFE scoring greatly improved its ranking. For example, iSAFE ranked the favored mutation within the top 20 in 94% of the simulations of a 5-Mbp region (Fig. 3a and Supplementary Fig. 4), in contrast to CMS, which gave a top-20 ranking in 35% of cases.

iSAFE scores are not based on likelihood computations, and we use them primarily to rank-order mutations. However, iSAFE scores are normalized and can be compared across samples. Empirically computed $P$ values (Online Methods) for iSAFE indicated good performance at $P < 1 \times 10^{-4}$ (iSAFE $\geq 0.1$; Supplementary Fig. 7).

Figure 1 | Characterization of the SAFE method. (a) For any mutation $e$, let $f$ denote the mutation frequency, or the fraction of haplotypes carrying the mutation. The HAF score for haplotype $h$ is the sum of the derived allele counts on $h$. Carriers of the favored mutation have a higher fraction of the total HAF score of the sample (high $\phi$) and fewer distinct haplotypes compared with non-carriers (low $\kappa$). (b) Schematic genealogy under a selective sweep. Mutations on haplotypes carrying the favored mutation can arise before the favored mutation ("ancestral to favored") or after the favored mutation ("descendant to favored"). Right, simulations showing $\phi$ versus $\kappa$ for each variant under neutral evolution or a selective sweep (1,000 simulations; favored allele frequency $\nu = 0.5$, and default values for other parameters; see Online Methods). The joint distribution of $\phi$ and $\kappa$ in a selective sweep changes in a dramatic but predictable manner that separates non-carrier, descendant, and ancestral mutations from the favored mutations. The SAFE score presents a normalized difference of the two statistics, $\phi$ and $\kappa$. (c) Performance (favored mutation rank) of SAFE compared with that of iHS and SCCT on 50-kbp windows with 1,000 simulations per frequency bin and default parameter values (Online Methods) for a fixed population size with ongoing selective sweeps. The plot on the left combines all allele frequencies, and that on the right shows median and mean ranks for replicates divided into four bins, CDF, cumulative distribution function.

Figure 2 | Illustration of the iSAFE method. (a) Because different genomic windows ($w$) have different genealogies because of recombination, the SAFE score of a nonfavored mutation $e$ is relatively low when it is inserted in other windows. In contrast, the SAFE score of the favored mutation is likely to be dominant over those of other mutations (Supplementary Note 1). Identical haplotypes in each window are colored similarly. (b) The $\Psi e w$ matrix for a 5-Mbp region around the $LCT$ gene in the 1000GP FIN population shows that the ‘shoulder’ of selection can extend for a few megabase pairs. The blue circle indicates the location of putative favored mutation rs4988235. (c) SAFE and iSAFE performance (rank distribution of favored mutation) as a function of window size with 1,000 simulations per bin. Median and quartile values decay with increasing window size in SAFE, whereas iSAFE is robust to increases in window size.
We tested iSAFE performance on 22 human loci previously characterized as containing signatures of a selective sweep (Supplementary Note 2), with some evidence for the favored mutation. The list included eight ‘well-characterized’ cases with additional support for the favored mutation (Supplementary Table 1). Using genotype data from phase 3 of 1000 Genomes Project (1000GP) subpopulations, we used iSAFE to rank all variants (~21,000) in a 5-Mbp region surrounding each locus. Among the eight well-characterized cases \(^7,11-14\) (Fig. 3b and Supplementary Fig. 8), iSAFE ranked the candidate mutations as 1 in five cases (SLC2A4A3, LCT, EDAR, ACKR1, and TLR1) and ranked the remaining cases as 2 (ABCC11), 4 (HBB), and 13 (G6PD).

We checked whether the other 14 loci \(^7,15-18\) under selection showed strong iSAFE signals (Supplementary Note 2). In 3 of the 14 loci (FUT2, F12, and ASPM; Supplementary Fig. 9), we observed weak signals and did not make a prediction (peak iSAFE score < 0.027). In other loci, iSAFE ranked the candidate mutations as 1 in the SLC45A2/MATP (1000GP population code CEU), MC1R (CHB and JPT populations), and ATXN2/SHB3 (GBR population) loci (Fig. 3c), and as 7, 8, and 12 in the PSCA (YRI population), ADH1B (CHB and JPT populations), and PCDH15 (CHB and JPT populations) loci, respectively. (Note: full definitions for the 1000GP population codes used in this paper can be found at http://www.internationalgenome.org/category/population/.) In each case, the iSAFE scores were high, with the exception of PSCA (peak iSAFE score of 0.04; Supplementary Fig. 9).

The other five putative selected loci struck us as interesting in that the mutations with the top iSAFE rankings had high scores but were distinct from the reported candidate mutations (Fig. 3c and Supplementary Note 2). Many of these loci are involved in pigmentation and determine skin, eye, and hair color. For example, the tyrosinase gene (TYR), which encodes an enzyme involved in the first step of melanin production, is considered to be under positive selection with a nonsynonymous mutation rs1042602 as a candidate favored variant \(^7\). A second intrinsic variant, rs10831496, in GRM5, 396 kbp upstream of TYR, has been shown to have a strong association with skin color \(^9\). In contrast, iSAFE ranked mutation rs672144 at the top not only in the CEU population sample (iSAFE = 0.48, \(P < 1.3 \times 10^{-9}\)) but also in the EUR, EAS, AMR, and SAS super-populations (iSAFE > 0.5, \(P < 1.3 \times 10^{-8}\); Supplementary Fig. 10), consistent with a signal of selection present in all populations except AFR. It might not have been reported previously because it is near fixation in all 1000GP populations except AFR (Supplementary Fig. 10). We found that two distinct haplotypes carry the rs672144 mutation, both of which have remained at high frequency, maintained...
across a large stretch of the region, suggestive of a soft sweep with standing variation (Fig. 3d). Similarly, for the loci TRPV6, KITLG, and OCA-HERC2 (Supplementary Note 2, Fig. 3c, and Supplementary Figs. 11–13), the mutations assigned top ranks by iSAFE were identical across all non-African populations, and supported an out-of-Africa onset of selection. For the one remaining gene (CYP1A2/CSK; Supplementary Note 2 and Fig. 3c), the iSAFE top-ranked mutation rs2470893 was previously found to be significant in a genome-wide association study and was tightly linked to the candidate mutation. To summarize, iSAFE analysis ranked the candidate mutation among the top 13 in 14 of 22 loci, did not show a strong signal in 3 loci, and identified plausible alternatives in the remaining 5 loci (Supplementary Note 2).

The identification of the favored allele in a selective sweep is a long-standing problem in population genomics. Our results suggest that statistics obtained from the coalescent structure of a region under a selective sweep can indeed pinpoint the favored mutation. iSAFE performance remained robust to a range of simulation parameters, including initial frequencies (standing variation) and the frequency of the favored mutation at the time of sampling. Although most results in this paper were obtained for human populations, iSAFE can be easily extended to other populations, as it is not highly parameterized.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This research was supported in part by the NSF (grant DBI-1458557 to A.A. and V.B.) and the NIH (grant R01GM114362 to V.B.).

AUTHOR CONTRIBUTIONS

A.A., S.M., and V.B. conceived and designed the experiments and wrote the manuscript with input from all other authors; A.A., J.J.V., and A.I. performed the experiments; A.A. analyzed the data; A.A. and M.B. developed software tools; and P.C.S., S.M., and V.B. provided guidance throughout the study.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Vitti, J.J., Grossman, S.R. & Sabeti, P.C. Annu. Rev. Genet. 47, 97–120 (2013).
2. Fan, S., Hansen, M.E.B., Lo, Y. & Tishkoff, S.A. Science 354, 54–59 (2016).
3. Schrider, D.R., Mendes, F.K., Hahn, M.W. & Kern, A.D. Genetics 200, 267–284 (2015).
4. Field, Y. et al. Science 354, 760–764 (2016).
5. Azad, P. et al. J. Mol. Med. (Berl.) 95, 1269–1282 (2017).
6. Stobdan, T. et al. Mol. Biol. Evol. 34, 3154–3168 (2017).
7. Grossman, S.R. et al. Science 337, 883–886 (2010).
8. Ronen, R. et al. PLoS Genet. 11, e1005527–e1005527 (2015).
9. Wang, M. et al. Mol. Biol. Evol. 31, 3068–3080 (2014).
10. Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. PLoS Biol. 4, e72 (2006).
11. Sabeti, P.C. et al. Science 312, 1614–1620 (2006).
12. Ohashi, J., Naka, I. & Tsuchiya, N. Mol. Biol. Evol. 28, 849–857 (2011).
13. Tishkoff, S.A. et al. Science 293, 455–462 (2001).
14. Heffelfinger, C. et al. Eur. J. Hum. Genet. 22, 551–557 (2014).
15. Wilde, S. et al. Proc. Natl. Acad. Sci. USA 111, 4832–4837 (2014).
16. Coop, G. et al. PLoS Genet. 5, e1000500 (2009).
17. Campbell, C.D. et al. Nat. Genet. 44, 1277–1281 (2012).
18. Galinsky, K.J., Loh, P.-R., Mallick, S., Patterson, N.J. & Price, A.L. Annu. Rev. Genet. 59, 1130–1139 (2016).
19. Beleza, S. et al. PLoS Genet. 9, e1003372 (2013).
20. Cornelis, M.C. et al. Mol. Psychiatry 20, 647–656 (2015).
**ONLINE METHODS**

A comprehensive explanation of the method is provided in Supplementary Note 1.

**Method overview.** Methods to identify signatures of selective sweeps in population genomics data have been actively developed\(^3\) but mostly do not identify the specific mutation favored by selection. iSAFE uses a statistical derived solely from population genetics signals to pinpoint the favored mutation in a large region (5 Mbp) without any knowledge of demography, specific phenotype under selection, and functional annotations of mutations. iSAFE uses a two-step procedure to identify the favored variant, given a large region (5 Mbp) under selection. In the first step, it finds the best candidate mutations in small (low-recombination) windows. Finally, it combines the evidence to give an iSAFE score to all variants in the large region. It considers only biallelic sites, taking as input a binary SNP matrix with each row corresponding to a haplotype, and each column to a site. Entries in the matrix correspond to the allelic state, with 0 denoting the ancestral allele, and 1 denoting the derived allele.

**Haplotype allele frequency.** A haplotype contains/carry a mutation if it has the derived allele at site. Recently, we devised the HAF score to capture the dynamics of a selective sweep\(^5\). The HAF score for a haplotype \(h\) (\(HAF(h)\)) is the sum of the derived allele counts of the mutations on \(h\) (Supplementary Note 1 and Fig. 1a). It has been shown that when \(h\) is a carrier of the favored allele, \(HAF(h)\) increases with the frequency of the favored mutation (equation (SN1.9) of Supplementary Note 1), in contrast to HAF scores of non-carriers (equation (SN1.10) of Supplementary Note 1), and this can be used to separate carrier haplotypes from non-carriers without knowledge of the favored mutation\(^8\).

**Selection of allele favored by evolution.** Denote two haplotypes as ‘distinct’ if they have different HAF scores. For any mutation \(e\), let \(f(e)\) denote the mutation frequency, or the fraction of haplotypes carrying the mutation. Let \(k(e)\) (Fig. 1a) denote the fraction of distinct haplotypes that carry mutation \(e\),

\[
k(e) = \frac{\text{Number of distinct haplotypes carrying mutation } e}{\text{Number of distinct haplotypes in sample}}
\]

Similarly, let \(\phi(e)\) denote the normalized sum of HAF scores of all haplotypes carrying the mutation \(e\),

\[
\phi(e) = \frac{\text{Sum of HAF scores of haplotypes carrying mutation } e}{\text{Sum of HAF scores of all haplotypes}}
\]

We observe empirically that in a region evolving according to a neutral Wright–Fisher model, \(k(e)\) and \(\phi(e)\) are both estimators of \(f(e)\) (Supplementary Fig. 1). Moreover, empirical results suggest that the expected value of \(\phi - k\) is 0, and variance is proportional to \(f(1 - f)\). On the basis of these observations, we define the SAFE score of mutation \(e\) as

\[
\text{SAFE}(e) = \frac{\phi - k}{\sqrt{f(1 - f)}}
\]

Empirically, SAFE(e) behaves like a Gaussian random variable, with mean 0, under neutrality (Supplementary Fig. 1), and it can be used to test departure from neutrality. However, its real power appears during positive selection, when SAFE scores change in a dramatic, but predictable, manner (Fig. 1a,b). Assuming a no-recombination scenario (only for visual exposition), label mutations as ‘non-carrier’ if they are carried only by haplotypes not carrying the favored allele. The remaining mutations can be labeled as ‘ancestral,’ if they arise before the favored mutation, or ‘descendant’, if they arise after (Fig. 1b). When each mutation is represented as a point in a two-dimensional plot of (\(\phi, \kappa\)) values, these classes are clustered differentially (Fig. 1b). The selective sweep reduces the number of distinct haplotypes carrying the favored mutation (lower \(\kappa\)), leaving non-carrier mutations with an increased fraction of distinct haplotypes (higher \(\kappa\)). Increased HAF scores in carrier haplotypes reduce the proportion of the total HAF score contributed by non-carrier haplotypes (lower \(\phi\)). In contrast, the favored mutation has a high positive value of \(\phi - \kappa\) owing to the high HAF scores for carriers (higher \(\phi\)) and the reduced number of distinct haplotypes among its descendants (lower \(\kappa\)). As one follows values in the plot up to ancestral mutations, the number of non-carrier haplotype descendants increases, and \(\kappa\) grows faster than \(\phi\). For descendant mutations, there is a reduction in the already small number of distinct haplotypes. However, \(\phi\) decreases sharply, reducing \(\phi - \kappa\) (Fig. 1a,b). Thus, one can expect that the mutation with the highest SAFE score is a strong candidate for the favored mutation.

We performed extensive simulations to test SAFE on samples evolving neutrally and under positive selection. We varied one parameter in each run (Supplementary Fig. 2), including window size (\(L = 50\) kbp), number of individual haplotypes (\(n = 200\)) chosen from a larger effective population size (\(N = 20,000\)), and initial and final favored mutation frequencies (\(v_0 = 1/N\) and \(v\)). Whereas the standing variation, \(v_0 > 1/N\), generally weakened the selection signal, the performance of SAFE remained relatively robust to variation in \(v_0\). The median SAFE rank of the favored allele was at most 3 out of ~250 variants in all cases except when \(v_0 > 1,000/N\) (Supplementary Fig. 2). Similarly, the performance was robust to selection pressure, with only a slight degradation at weak selection (\(N_s = 50\)) (Supplementary Fig. 2) when the median rank went to 9 (3.5 percentile), whereas for \(N_s \geq 200\) the median rank was at most 2. As expected, the performance improved with increasing sample size (Supplementary Fig. 2). We also tested SAFE on a model of European demography and observed similar results (Supplementary Fig. 2). These tests used \(L = 50\) kbp, chosen so as to minimize the effects of recombination.

**Integrated selection of allele favored by evolution.** Next, we tested SAFE with increasing window sizes, and observed that the median rank of the favored mutation increased with increasing window size (Fig. 2c). The deterioration for larger windows probably occurs because most haplotypes are becoming unique, and \(\kappa\) loses its utility in pinpointing the favored mutation. However, the selective sweep signal is known to extend to large, linked regions, as far as 1 Mbp on either side of the favored allele. These ‘shoulders’\(^3\) of selective sweeps are helpful in identifying the region under selection, but make it harder to pinpoint the favored mutation. We further refined our method to exploit the signal from shoulders.

For larger regions, we considered a set of 50% overlapping windows (\(W\)) of fixed size (300 SNPs). For each window, we applied

\[\text{doi:10.1038/nmeth.4606}\]
SAFE and chose the mutation with the highest SAFE score. Let \( S_1 \) denote the set of selected mutations. Mutations in \( S_1 \) are likely to contain either the favored mutation itself or mutations linked to it. For mutation \( e \) in window \( w \), let \( \Psi_{e,w} \) denote either the SAFE score of \( e \) when \( e \) is ‘inserted’ into window \( w’ \), or 0, whichever is larger (Fig. 2a,b). Because different windows have different genealogies owing to recombination, \( \Psi_{e,w} \) is relatively high when \( e \) is the favored mutation and the genealogies of \( w \) and \( w’ \) are identical or very similar, but not otherwise. In contrast, the SAFE score of a non-favored mutation \( e \) is relatively low when it is inserted in other windows (Supplementary Note 1 and Fig. 2a). The window containing the favored mutation is likely to give a high score to mutations in \( S_1 \). Therefore, we assigned the weight of a window \( w \) as

\[
\alpha(w) = \frac{\sum_{e \in S_1} \Psi_{e,w}}{\sum_{w’ \in W} \sum_{e \in S_1} \Psi_{e,w’}}
\]

(4)

Windows that contain the favored mutation and those sharing its genealogy are expected to have high \( \alpha \) values. We defined the iSAFE score for all mutations \( e \) (including those not in \( S_1 \)) as

\[
iSAFE(e) = \sum_{w \in W} \Psi_{e,w} \alpha(w)
\]

(5)

iSAFE scores are not based on likelihood computations, and the distribution of scores depends on largely unknown factors including demography, time since onset of selection, selection coefficient, and other parameters. Nevertheless, they can be used to rank-order the mutations. Additionally, iSAFE scores are normalized and can be compared across samples. We found distinct differences in performance below a score threshold of 0.1. The median rank of the favored mutation was 4 when the peak iSAFE score exceeded 0.1, versus a median rank of 10 along with a longer tail when the peak iSAFE score was below 0.1 (Supplementary Fig. 7). Empirically computed \( P \) values for iSAFE indicated good performance when \( P < 1 \times 10^{-4} \) (Supplementary Fig. 7).

Adding outgroup samples. Not surprisingly, iSAFE performance deteriorates when the favored mutation is fixed or near fixation (\( v > 0.9 \) in Supplementary Fig. 3). To handle this special case, we included individuals from non-target populations. For a mutation, define the maximum difference in derived allele frequency score (MDDAF) as

\[
\text{MDDAF} = D_T - \min(D_{NT})
\]

(6)

where \( D_T \) is the derived allele frequency in the target population and \( \min(D_{NT}) \) is the minimum derived allele frequency over all non-target populations. Simulations of human population demography under neutral evolution showed \( P(\text{MDDAF} > 0.78 | D_T > 0.9) = 0.001 \) (Supplementary Fig. 15). Therefore, when we observed the rare event of high-frequency mutations in target (\( D_T > 0.9 \) with \( \text{MDDAF} > 0.78 \), we added random outgroup samples to the data to constitute 10% of the data (Supplementary Note 1). In testing on the phase 3 1000GP data, we chose outgroup samples from non-target 1000GP populations. The addition of outgroup samples using the MDDAF criterion was tested in extensive simulations. Although the performance did not change for \( v < 0.9 \), it dramatically improved for high frequencies, including when the favored mutation was fixed in the target population (Supplementary Fig. 3).

iSAFE evaluation. In testing on models of human demography, we also compared iSAFE with CMS. Although CMS showed excellent performance in localizing the favored mutation, iSAFE scoring greatly improved the ranking of the mutation. For example, iSAFE ranked the favored mutation within the top 20 in 94% of the simulations of a 5-Mb region (Fig. 3a and Supplementary Fig. 4), in contrast to CMS, which had a top-20 ranking in 35% of cases.

We note that in testing instances of previously characterized sweeps in 1000GP data, performance is difficult to characterize because of many complicating factors. Multiple sweeps could be occurring in response to different selection events, including background selection in the same region, or polygenic selection may dilute the selection signal at any one locus. Moreover, the favored mutation is well characterized in only a few instances. We looked for genes/regions that showed the signature of a selective sweep in one of the 1000GP subpopulations and had additional evidence pointing to the favored mutation. We identified 22 genes with some evidence, but only 8 ‘well-characterized’ cases with additional support for the favored mutation (Supplementary Note 2 and Supplementary Table 1).

Default simulation parameters. Neutral and sweep samples were generated with the simulator msms\(^{33}\). By default, simulated populations are haploid with sample size of \( n = 200 \) haplotypes from a larger effective population of \( N = 20,000 \) haplotypes, each of length \( L \), with default values of 50 kbp for SAFE and 5 Mbp for iSAFE. For human populations, a mutation rate of approximately \( \mu = 2.5 \times 10^{-8} \) mutations per base pair per generation\(^{17,34} \) and a recombination rate of approximately \( r = 1.25 \times 10^{-8} \) per base pair per generation\(^{35} \) have been proposed. For SAFE simulations, we used a scaled mutation rate \( \theta = 2N\mu = 1 \) mutation per kilobase pair per generation and scaled recombination rate \( \rho = 2Nr = 0.5 \) crossovers per kilobase pair per meiosis to approximate human rates. The rates were scaled linearly by \( L \). In the case of positive selection, the default scaled selection strength of the favored allele was set at \( Ns = 500 \), with the favored mutation located at a random position uniformly distributed on the range \([1, L] \). The default value for favored mutation starting frequency \( v_0 = 1/N \) (hard sweep), and the frequency of the favored mutation \( v \) at the time of sampling was a random value uniformly distributed on the range \([0.1, 0.9] \). We used the default parameters for all simulations unless otherwise stated.

A model of human demography. We simulated the demography of 1000GP AFR, EUR, and EAS populations with the parameters shown in Supplementary Figure 14, based on a popular demographic model of human population\(^{36} \). In the case of positive selection, the selection coefficient was set to \( s = 0.05 \), and the starting favored allele frequency \( v_0 = 0.001 \). The time of onset of selection was chosen at random (using the distribution in Supplementary Fig. 14) after the out-of-Africa event, in the lineage of the EUR population (as the target population). When the onset of selection was before the split of EUR and EAS (>23,000 years ago), both populations (EUR and EAS) were under selection.
We used the SCCT (v1.1) software. Next, we normalized the iHS score by estimating the distribution of raw iHS scores on 1,000 neutral simulations with the same simulation parameters. The iHS scores were always computed on a 5-Mbp window. When comparing results with SAFE on a 50-kbp window, we used the corresponding iHS scores in the identical 50-kbp region surrounding the favored variant (Fig. 1c and Supplementary Fig. 2). In considering 5-Mbp windows (Supplementary Fig. 3), we compared the iHS scores on all variants for iHS against iSAFE.

Computing CMS scores. CMS requires a control population as well as a demographic model, in addition to the target population under selection. All CMS comparisons on simulated data were performed with a model of human demography with a random onset of selection (Supplementary Fig. 14). We used the CMS (v2.0) software available at https://github.com/broadinstitute/cms, disabling the default allele frequency filter to allow for more direct comparison with iSAFE SNP ranking.

Computing empirical $P$ values. We applied iSAFE on a neutrally evolving simulated population with window size of 5 Mbp, based on the European demography shown in Supplementary Figure 14. A $P$ value was calculated on the basis of the empirical distribution of iSAFE on these simulated populations. We limited the number of samples to ~74,800,000 for efficiency, and this allowed us to get a $P$ value as low as $1.34 \times 10^{-8}$ for an iSAFE score of 0.304. Scores higher than this cutoff were considered to have $P < 1.34 \times 10^{-8}$.

Putative selective sweeps in human populations. We examined eight well-characterized selective sweeps with strong candidate mutations. These genes are LCT, SLC24A5, TLR1, EDAR, ACKR1 (DARC), ABCC11, HBB, and G6PD. iSAFE results for these genes are summarized in Figure 3b, Supplementary Figure 8, and Supplementary Table 1. We also examined 14 other regions reported to be under selection with one or more candidate favored mutations. A detailed report for each of these 14 loci is provided in Supplementary Note 2.

Code availability. The iSAFE software and instructions are available at https://github.com/alek0991/iSAFE and as Supplementary Software.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. For all the following datasets, the genome build is GRCh37/hg19. We downloaded the phased haplotypes of the 1000GP (phase 3) dataset from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. The ancestral allele dataset from Ensembl (release 75) was downloaded from http://ftp.ensembl.org/pub/release-75/fasta/ancestral_alleles/. The physical position was converted into genetic position using the HapMap II genetic map downloaded from http://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20110106_recombination_hotspots/.

21. Ferrer-Admetlla, A., Liang, M., Korneliussen, T. & Nielsen, R. Mol. Biol. Evol. 31, 1275–1291 (2014).
22. Pybus, M. et al. Bioinformatics 31, 3946–3952 (2015).
23. Garud, N.R., Messer, P.W., Buzbas, E.O. & Petrov, D.A. PloS Genet. 11, e1005004 (2015).
24. DeGiorgio, M., Huber, C.D., Hubisz, M.J., Hellmann, I. & Nielsen, R. Bioinformatics 32, 1895–1897 (2016).
25. Ronen, R., Udpa, N., Halperin, E. & Bafna, V. Genetics 195, 181–193 (2013).
26. Pavlidis, P., Živkovic, D., Stamatakis, A. & Alachiotis, N. Mol. Biol. Evol. 30, 2224–2234 (2013).
27. Chen, H., Patterson, N. & Reich, D. Genome Res. 20, 393–402 (2010).
28. Sabeti, P.C. et al. Nature 449, 913–918 (2007).
29. Sabeti, P.C. et al. Nature 419, 832–837 (2002).
30. Nielsen, R. et al. Genome Res. 15, 1566–1575 (2005).
31. Kim, Y. & Nielsen, R. Genetics 167, 1513–1524 (2004).
32. Shriver, M.D. et al. Hum. Genomics 1, 274–286 (2004).
33. Ewing, G. & Hermisson, J. Bioinformatics 26, 2064–2065 (2010).
34. Nachman, M.W. & Crowell, S.L. Genetics 156, 297–304 (2000).
35. Jensen-Seaman, M.I. et al. Genome Res. 14, 528–538 (2004).
36. Gravel, S. et al. Proc. Natl. Acad. Sci. USA 108, 11983–11988 (2011).
37. Szpiech, Z.A. & Hernandez, R.D. Mol. Biol. Evol. 31, 2824–2827 (2014).
38. 1000 Genomes Project Consortium. Nature 526, 68–74 (2015).
39. Zerbino, D.R. et al. Nucleic Acids Res. 46, D754–D761 (2018).
40. International HapMap Consortium. Nature 449, 851–861 (2007).

doi:10.1038/nmeth.4606
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

### Experimental design

1. **Sample size**
   
   Describe how sample size was determined.

   We used the 1000 Genomes Project whole genome samples where there is 26 populations and each population have on average 100 individuals. Therefore for simulation we chose 100 individuals (200 haplotypes).

2. **Data exclusions**
   
   Describe any data exclusions.

   No data were excluded.

3. **Replication**
   
   Describe the measures taken to verify the reproducibility of the experimental findings.

   We have done extensive simulations with different parameters that are explained in details in the Supplementary information. We also tested the performance of the method in well-characterized examples of human population.

4. **Randomization**
   
   Describe how samples/organisms/participants were allocated into experimental groups.

   As provided by 1000 Genomes Project, samples are grouped by population. For example, British in England and Scotland (GBR), Yoruba in Ibadan, Nigeria (YRI), and so forth.

5. **Blinding**
   
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Data is provided by 1000 Genomes Project and we didn't change anything regarding the data.

   Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. **Statistical parameters**
   
   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

   - n/a  
     - [ ] Confirmed
   - [ ] The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
   - [ ] A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
   - [ ] A statement indicating how many times each experiment was replicated
   - [ ] The statistical test(s) used and whether they are one- or two-sided
     - *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
   - [ ] A description of any assumptions or corrections, such as an adjustment for multiple comparisons
   - [ ] Test values indicating whether an effect is present
     - *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
   - [ ] A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
   - [ ] Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

   *See the web collection on statistics for biologists for further resources and guidance.*
7. Software

Describe the software used to analyze the data in this study.

- selscan (v1.1.0a); Szpiech et al. (2014); https://github.com/szpiech/selscan
- SCCT (v1.1); Wang et al. (2014); https://github.com/wavefancy/scct
- CMS (v2.0); Grossman et al. (2010); https://github.com/broadinstitute/cms

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

- no unique material were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

- no antibodies were used.

10. Eukaryotic cell lines

- State the source of each eukaryotic cell line used.
- Describe the method of cell line authentication used.
- Report whether the cell lines were tested for mycoplasma contamination.
- If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

- no eukaryotic cell line were used

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

- no animals were used

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

- The study didn’t involve human research participants