Supplementary Table 1:

iSAFE on 8 well characterized selective sweeps. Number of haplotypes in CEU, CHB, JPT, FIN, and YRI populations are 198, 206, 208, 198, and 216, respectively. Computation of empirical P value is provided in Online Methods.

Gene	Target Population	Candidate SNP ID	Candidate SNP Function	Frequency	Selective Advantage	iSAFE Rank	P	Selection Reference	Functional Reference
SLC24A5	CEU	rs1426654	Missense	1	Light skin pigmentation	1	<1.3e-8	[1]	[2]
EDAR	CHB+JPT	rs3827760	Missense	0.87	Hair and teeth	1	<1.3e-8	[1]	[3,4]
LCT/MCM6	FIN	rs4988235	Intron	0.59	Lactase persistence	1	<1.3e-8	[5]	[6,7]
TLR1	CEU	rs5743618	Missense	0.77	Sepsis, leprosy, tuberculosis	1	1.0e-5	[8]	[9]
ACKR1/DARC	YRI	rs2814778	5′UTR	1	Malaria resistance	1	2.8e-5	[10]	[11]
ABCC11	CHB+JPT	rs17822931	Missense	0.93	Cold climate, earwax, body odour	2	<1.3e-8	[12]	[12]
HBB	YRI	rs334	Missense	0.14	Malaria resistance	4	1.6e-4	[13]	[14]
G6PD	YRI	rs1050828	Missense	0.21	Malaria resistance	13	7.3e-6	[5]	[15]

Supplementary Note 1:
The iSAFE statistic

1.1 iSAFE: Input, Output and Overview.

Consider a sample of phased haplotypes in a genomic region. We assume that all sites are biallelic and polymorphic in the sample. Thus, our input is in the form of a binary SNP matrix with each row corresponding to a haplotype and each column to a mutation, and entries corresponding to the allelic state, with 0 denoting the ancestral allele, and 1 denoting the derived allele. The output is a non-negative iSAFE-score for each mutation, with the highest score corresponding to the favored mutation.

At a high level, iSAFE uses a 2-step procedure to identify the favored variant, given a large region (5Mb) 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.

1.2 HAF: Haplotype Allele Frequency.

The HAF score for haplotype h is the sum of the derived allele counts of the mutations on h. Define the SNP matrix M such that, $M_{h,e} = 1$ if haplotype h carries the derived allele of SNP e, and 0 otherwise. The Haplotype Allele Frequency (HAF) score of haplotype h defined in Ronen et al. (2015)[16] as:

$$\text{HAF}(h) = \sum_{e} M_{h,e} \sum_{h'} M_{h',e} = \sum_{h'} [M \cdot M^T]_{h,h'}, \qquad \textbf{(SN1.1)}$$

where $\sum_{h'} M_{h',e}$ is derived allele count for SNP e, and $[M \cdot M^T]_{h,h'}$ is number of shared derived alleles (mutations) between haplotypes h and h' (see **Fig. 1a**).

1.3 SAFE: Selection of Allele Favored by Evolution.

For each SNP e, define ϕ as:

$$\phi(e) = \frac{\sum_h [M_{h,e} \cdot \text{HAF}(h)]}{\sum_h \text{HAF}(h)}, \qquad \textbf{(SN1.2)}$$

In other words, ϕ is sum of HAF scores of carriers of the derived allele e ($\sum_h [M_{h,e} \cdot \text{HAF}(h)]$), divided by sum of HAF scores of all haplotypes in the sample ($\sum_h \text{HAF}(h)$).

Similarly, for each SNP e, we define κ as:

$$\kappa(e) = \frac{\left|\bigcup_h \{M_{h,e} \cdot \text{HAF}(h)\}\right| - 1}{\left|\bigcup_h \{\text{HAF}(h)\}\right|}, \qquad \textbf{(SN1.3)}$$

implying that κ is the fraction of distinct non-zero values in HAF scores of SNP e carriers. κ is closely related, but not identical, to fraction of all distinct haplotypes that carry the mutation e.

We use ϕ and κ, to define the SAFE score of a SNP e as:

$$\mathrm{SAFE}(e) = \frac{\phi(e) - \kappa(e)}{\sqrt{f_e(1 - f_e)}} \tag{\textbf{SN1.4}}$$

where f_e is the derived allele frequency of SNP e.

To explain the behavior of the SAFE-score in pin-pointing the favored mutation, we describe a collection of theoretical and empirical observations that can be summarized as follows:

1. *Under neutrality, $\phi(e)$ and $\kappa(e)$ are (biased) estimators of f_e.*
2. *$\lambda f(1 - f)$ is a biased estimator for variance of $(\phi - \kappa)$, where λ is a positive constant.*
3. *The two points above allow the use of SAFE-score as a statistic that empirically follows a Gaussian distribution with mean 0 under neutrality.*
4. *For a population evolving under selection, ϕ and κ move in opposite directions. Specifically, for the favored mutation e, $\phi(e)$ increases, while $\kappa(e)$ decreases. The SAFE-score tends to be maximized for the favored mutation e.*

We elaborate on these points below.

1.3.1 Behavior of ϕ, κ under neutrality, constant population size.

Consider a sample of size n selected from a population evolving neutrally according to the Wright Fisher model (constant population size, random mating, discrete generations, no recombination), with scaled mutation rate θ. Let ξ_i be the number of sites with derived allele count i. From Ronen et al.[16], the mean of the HAF scores of all n haplotypes in the sample is

$$\frac{1}{n}\sum_h \mathrm{HAF}(h) = \frac{1}{n}\sum_{i=1}^{n-1} \xi_i \cdot i^2. \tag{\textbf{SN1.5}}$$

Under the coalescent model, Eq. (22) of Fu 1995[17] shows that $\mathbb{E}[\xi_i] = \theta/i$ for all $1 \le i \le n-1$. By averaging over all haplotypes in all genealogies, the expected HAF score is computed as

$$\mathbb{E}[\mathrm{HAF}] = \frac{1}{n}\sum_{i=1}^{n-1} \mathbb{E}[\xi_i] \cdot i^2 = \frac{\theta}{n}\sum_{i=1}^{n-1} i\,. \tag{\textbf{SN1.6}}$$

Thus, the expected HAF score is,

$$\mathbb{E}[\mathrm{HAF}] = \frac{\theta(n-1)}{2}\,. \tag{\textbf{SN1.7}}$$

Therefore, the fraction of the total HAF-score of fn randomly chosen haplotypes is approximately f. A mutation e with derived allele frequency also has fn descendants (carriers). However, to compute the sum of the HAF-scores, we must consider a random coalescent process with a condition that carriers coalesce to a common ancestor before any carrier coalesces with a non-carrier.

This is harder, even though conditional coalescent processes have been studied extensively (e.g., Wiuf and Donnelly[18]). Empirical analysis on neutral coalescent simulations conditioned on the mutation e having fn carriers reveals that (**Supplementary Fig. 1a**)

$$\mathbb{E}[\phi(e)|f] \approx f\,.$$

While κ has not been studied previously, it is closely related to the fraction of distinct haplotypes in the sample. Empirically, for a mutation e, with fn descendants, we observe that (**Supplementary Fig. 1a**)

$$\mathbb{E}[\kappa(e)|f] \approx f\,.$$

and, for all e (**Supplementary Fig. 1b**),

$$\mathbb{E}[\phi(e) - \kappa(e)] \approx 0\,. \qquad \textbf{(SN1.8)}$$

1.3.2 Distribution of SAFE-scores in a neutrally evolving population.

The discussion above suggests that $\mathbb{E}(\text{SAFE}(e)) = 0$ for all derived alleles e. Additionally, empirical observations suggest that $\lambda f(1-f)$ is a biased estimator for variance of $(\phi-\kappa)$, where λ is a positive constant. We observed empirically that the distribution of the SAFE score of derived alleles in a neutrally evolving population is therefore approximated by a *Gaussian* distribution with mean 0 and unknown variance λ (see **Supplementary Fig. 1b**).

1.3.3 Behavior of ϕ, κ, and SAFE in a population under selection with a constant population size.

The dynamics of HAF-score for a haplotype carrying the favored mutation in an ongoing selective sweep was analyzed earlier[16]. It increases dynamically upto fixation of the favored allele, and then decreases dramatically.

Formally, let $\text{HAF}^{\,\text{car}}$ (respectively, $\text{HAF}^{\,\text{non}}$) denote the HAF score of a random haplotype carrier of the favored allele (respectively, a non-carrier) when a fraction f of the n sampled haplotypes carry the favored allele. In S1 Text of Ronen et al.(2015)[16], we show that under strong selection ($Ns \gg 1$) and no recombination,

$$\mathbb{E}[\text{HAF}^{\,\text{car}}] \approx \theta n \left(\frac{f+1}{2} - \frac{1}{(1-f)n+1} \right)\,, \qquad \textbf{(SN1.9)}$$

$$\mathbb{E}[\text{HAF}^{\,\text{non}}] \approx \theta n \left(\frac{1}{2} + \frac{1}{2n} - \frac{1}{(1-f)n+1} \right)\,. \qquad \textbf{(SN1.10)}$$

Because of the separation between carriers and non-carriers, the HAF-scores can be used to predict the carrier of ongoing selective sweeps without knowledge of the favored allele[16]. Moreover, for the favored allele e with fn descendants, in a hard selective sweep that is not very close to fixation, we can approximate $\phi(e)$ as

$$\phi(e) \approx \frac{fn\mathbb{E}[\text{HAF}^{\,\text{car}}]}{fn\mathbb{E}[\text{HAF}^{\,\text{car}}] + (1-f)n\mathbb{E}[\text{HAF}^{\,\text{non}}]} \approx \frac{f^2+f}{f^2+1} = f + \frac{f^2(1-f)}{f^2+1} \geq f\,. \qquad \textbf{(SN1.11)}$$

For a population undergoing a positive natural selection with favored mutation e, $\phi(e)$ overestimates the favored allele frequency f (**Fig. 1a,b** and Eq. **SN1.11**). On the other hand, $\kappa(e)$ underestimates f (**Fig. 1a,b**). Therefore, we expect the distribution of $(\phi - \kappa)$ for the favored allele to be skewed in positive direction.

SAFE score performs very well in identifying the favored variant within a small window (**Fig. 1c** and **Supplementary Fig. 2**); but the performance decays in larger windows (**Fig. 2c**); because in larger windows most of the haplotypes become unique and κ estimate f correctly, even for favored mutations of selective sweeps, while we expect it to underestimate the f for the favored mutations. Consequently, the estimator κ is no-longer useful for pinpointing the favored mutation.

1.4 Illustration of iSAFE: integrated SAFE for large regions.

We devise iSAFE-score by extending the SAFE score to boost the performance in larger windows. We apply the SAFE score, as a kernel, on overlapping sliding windows. Define $\mathcal{S}$ as the set of all SNPs, $\mathcal{W}$ as the set of all sliding windows. Let $\mathcal{S}_1 \subseteq \mathcal{S}$ denote the subset of mutations that had the highest SAFE-score in their respective windows. For mutation $e \in \mathcal{S}$, and window $w \in \mathcal{W}$, let $\Psi_{e,w}$ denote the SAFE-score of e, when e is 'inserted' into window w if it is positive, 0 otherwise. **Fig. 2a** provides a cartoon illustration of windows w_1, w_2, w_3 and $\bigstar$, $\blacktriangle$, and $\blacksquare$, where $\bigstar$ denotes the favored mutation and is located in w_2.

We note the following:

- $\Psi_{\bigstar,w_2}$ is high for the favored mutation $\bigstar$. However, $\Psi_{\blacktriangle,w_1}$ and $\Psi_{\blacksquare,w_3}$ may be high even for hitchhiking mutations ($\blacktriangle$,$\blacksquare$) due to the genealogies of w_1 and w_3. Thus SAFE-score by itself may not be a reliable predictor over a large region containing multiple windows.

- When a non-favored mutation is inserted in a window with a different genealogy, it is not likely to have a high SAFE-score. When $\bigstar$ and $\blacktriangle$ are inserted into window w_3, $\Psi_{\bigstar,w_3} > \Psi_{\blacktriangle,w_3}$ because $\bigstar$ separates carriers from non-carriers and has high values for $\phi(\bigstar)$ and low values for $\kappa(\bigstar)$. On the other hand, $\kappa(\blacktriangle)$ is higher because its descendants include non-carriers which are typically distinct haplotypes. Similarly $\Psi_{\bigstar,w_1} > \Psi_{\blacksquare,w_1}$ because $\phi(\blacksquare)$ is lower in w_1. In other words, the weighted sum of $\Psi_{\bigstar,w}$ over all windows w is likely to dominate other mutations.

- Similarly, the window containing the favored mutation (w_2) has the appropriate genealogy, and is likely to give a high score to multiple candidate mutations.

Based on these considerations, we define the score α of window $w \in \mathcal{W}$ as:

$$\alpha(w) = \frac{\sum_{e\in\mathcal{S}_1} \Psi_{e,w}}{\sum_{w'\in\mathcal{W}} \sum_{e\in\mathcal{S}_1} \Psi_{e,w'}}, \qquad \textbf{(SN1.12)}$$

The window with the highest weight is the one which gets higher SAFE-scores for other mutations that are insrted into it. Finally, we define the score iSAFE of mutation $e \in \mathcal{S}$ as:

$$\text{iSAFE}(e) = \sum_{w \in \mathcal{W}} \Psi_{e,w} \cdot \alpha(w). \qquad \textbf{(SN1.13)}$$

where the mutation with the highest score is one that gives high scores when inserted into high weight windows.

1.5 MDDAF: Maximum Difference in Derived Allele Frequency.

We have shown that iSAFE is successful in pinpointing the favored variant in an ongoing selective sweep. When the favored mutation is near fixation ($\nu > 0.9$), iSAFE performance decays and when the favored variant is fixed ($\nu = 1$), iSAFE cannot detect the favored mutation because it is no longer a variant (**Supplementary Fig. 3e**). For the purpose of pinpointing the favored mutation in a fixed selective sweeps we add random samples from non-target population (outgroup) to the target population to constitute 10% of the sample.

To minimize the noise added to the data with random outgroup samples, we devise a simple method to decide whether to use outgroups or not. Our score is motivated by the work of Grossman et al.(2010)[1], who introduced the ΔDAF score of a mutation as $\Delta\text{DAF} = D_T - \text{mean}(\boldsymbol{D}_{NT})$, where D_T is the derived allele frequency in the target population and $\text{mean}(\boldsymbol{D}_{NT})$ is the *average* derived allele frequency in non-target populations. As it is possible that some of the non-target populations are also under selection, choosing the average derived allele frequency may lower ΔDAF, and weaken the signal of selection. Instead we define the Maximum Difference in Derived Allele Frequency (MDDAF) score as :

$$\text{MDDAF} = D_T - \min(\boldsymbol{D}_{NT}), \qquad \textbf{(SN1.14)}$$

where, D_T is the derived allele frequency in the target population and $\min(\boldsymbol{D}_{NT})$ is the *minimum* derived allele frequency over all non-target populations.

1.6 Adding Outgroup Samples.

Simulation of human population demography under neutral evolution (**Supplementary Fig. 14**), shows $P(\text{MDDAF} > 0.78 | D_T > 0.9) = 0.001$ (**Supplementary Fig. 15**) making it a rare event to have high MDDAF score even when the frequency is high in the Target population. Therefore, when there is a high frequency mutation ($D_T > 0.9$) with MDDAF > 0.78 in the target population, we add random outgroup samples to the data to constitute 10% of the data. For analysis on real data, where we looked at 1000GP populations, we randomly selected outgroup samples from non-target populations of 1000GP.

In **Supplementary Fig. 3c**, we compared the performance of iSAFE with or without having the option of using outgroup samples; we simulated 5 Mbp of human genome based on the human demography model described in **Supplementary Fig. 14**. The selection happens in a random time, with a distribution given in **Supplementary Fig. 14b**, after the out of Africa in the lineage of EUR population (as the target population). When the onset of selection is before split of EUR and EAS ($>$ 23kya), both (EUR and EAS) are under selection. When we have random sample option, we use the MDDAF criterion to decide whether we should use random sample or not. In case of adding random sample, we add a random subset of individuals from EAS+AFR to constitute 10% of the data (200 haplotypes from EUR and 22 from EAS+AFR).

The performance of iSAFE for sweeps with $\nu < 0.9$ did not change with or without having outgroup sample option (**Supplementary Fig. 3e**). When frequency of the favored mutation is near fixation ($\nu > 0.9$) having the outgroup sample option is helpful and increase the performance of the iSAFE. When the sweep is fixed ($\nu = 1$), iSAFE is no longer capable of detecting the favored mutation without having outgroup samples because the favored mutation is no longer a variant in the target population. However, with the outgroup sample option, iSAFE can successfully pinpoint the Favored mutation even in a fixed selective sweep (see **Supplementary Fig. 3e**).

Supplementary Note 2:
Results on selective sweeps in human populations

2.1 Well characterized selective sweeps.

We examined 8 well characterized selective sweeps with strong candidate mutation. These loci are LCT, SLC24A5, TLR1, EDAR, ACKR1/DARC, ABCC11, HBB, and G6PD[1,13,6,12,15,8]. iSAFE results for these loci are summarized in **Fig. 3b** and **Supplementary Fig. 8** and **Supplementary Table 1**.

We also examined 14 other loci reported to be under selection with one or more candidate favored mutations[19,20,5,1,21].

2.2 Pigmentation genes.

SLC45A2/MATP. This region is involved in human pigmentation pathways and is a target of selective sweep in European population[19]. A nonsynonymous mutation rs16891982 is associated with light skin pigmentation and is believed to be the favored variant[1,19]. This mutation is also ranked first by iSAFE out of ~21,000 mutations (5 Mbp) in CEU population with a significant score (see **Fig. 3c**, iSAFE = 0.32, $P < 1.3\text{e-}8$). This mutation is almost fixed in European; frequency in AFR, EAS, SAS, AMR, and EUR is 0.04, 0.01, 0.06, 0.45, and 0.94, respectively.

MC1R. The MC1R gene is implicated in many skin color phenotypes, including red hair, fair skin, freckles, poor tanning response and higher risk of skin cancer. It is is a target of positive selection in East Asian populations, with a non-synonymous mutation (rs885479) suggested as a candidate favored mutation[20]. This mutation is ranked first by iSAFE in CHB+JPT (see **Fig. 3c**, iSAFE = 0.24, $P = 1.4\text{e-}6$) out of ~16,000 mutations (2.8 Mbp). The putative selected region is 300 kbp away from the telomere of chromosome 16.

GRM5-TYR. The Tyrosinase (TYR) gene, encoding an enzyme involved in the first step of melanin production is present in a large region under selection. A nonsynonymous mutation rs1042602 in TYR gene is reported as a candidate favored variant[19]. A second intronic variant rs10831496 in GRM5 gene, 396 kbp upstream of TYR, has been shown to have a strong association with skin color[22].

In contrast, iSAFE ranks mutation rs672144 as the top candidate for the favored variant region out of ~22,000 mutations (5 Mbp). This variant was the top ranked mutation not only in CEU (iSAFE = 0.48, $P \ll 1.3\text{e-}8$), but also the top ranked mutation for EUR, EAS, AMR, and SAS (see **Fig. 3c** and **Supplementary Fig. 10**). The signal of selection is strong in all populations (iSAFE > 0.5, $P \ll 1.3\text{e-}8$ for all of) except AFR, which does not show a signal of selection in this region. It may not have been reported earlier because it is near fixation in all populations of 1000GP except for AFR ($f = 0.27$), as seen in **Supplementary Fig. 10**. We plotted the haplotypes carrying rs672144 and found (**Fig. 3d**) that two distinct haplotypes carry the mutation, both with high frequencies maintained across a large stretch of the region, suggestive of a soft sweep with standing variation.

The previously suggested candidates rs1042602, rs10831496 are fully linked to rs672144 (**Supplementary Fig. 10c**), but not to each other. The EUR haplotypes can be partitioned into 4 clusters (**Supplementary Fig. 10c**). Each of the 4 haplotypes show high homozygosity, suggestive of selection. However, rs1042602 can only explain the sweep in clusters C1+C2. rs10831496 can only explain C1+C3. Only rs672144 explains all 4 clusters, providing a simpler explanation of selection in this region. GTEx eQTL analysis on TYR gene for the tissue 'Skin - Sun Exposed (Lower leg)' showed $P = 0.61$ for rs1042602, $P = 0.15$ for rs10831496, and $P = 0.08$ for rs672144. While the P value does not rise to a level of significance due to sample size issues, it is indicative of a regulatory function for the mutation.

OCA2-HERC2. This region is suggested as a target of selection in European[1,23,19], and several mutations in this region are associated with hair, eye, and skin pigmentation. For example, rs12913832 is considered to be the main determinant of iris pigmentation (brown/blue) and is also associated with skin and hair pigmentation and the propensity to tan[19]. rs1667394 is also linked to blond hair and blue eyes[23]. Some other mutations, many fully linked, (rs4778138, rs4778241, rs7495174, rs1129038, rs916977) are also associated with blue eyes[23]. This region is also suggested to be a target of selection in East Asia with rs1800414 suggested as a candidate for light skin pigmentation in that population. We applied iSAFE on this region to all 1000GP super-populations.

iSAFE selected a single variant rs1448484 in OCA2 (with high confidence, $P < 1.34\text{e-}8$ for EUR, EAS, AMR and $P = 2.13\text{e-}6$ for SAS) as the favored variant in all 1000GP populations (EUR, EAS, SAS, AMR) except for AFR that showed no signal of selection in this region (see **Supplementary Fig. 11** and **Fig. 3c**). This variant is close to fixation in all populations except for AFR, where $\nu = 20\%$ (see **Supplementary Fig. 11**). iSAFE result along with the frequency pattern of the top ranked variant, suggests an out of Africa selection, probably on light skin color, on this region. The other candidate variants are all ranked high, and tightly linked with the top-ranked variant (**Supplementary Table SN2.1**).

KITLG. This genomic region has been linked to skin pigmentation[24] in European and East Asian populations, and shows a strong signature of selective sweep on regulatory regions surrounding the gene in all non-African populations[20], with a candidate variant rs642742, that is associated with skin pigmentation[24].

iSAFE analysis identified the same mutations gaining the top rank in multiple populations (**Supplementary Fig. 12**). Top rank mutations in EUR, SAS, EAS, and AMR populations are shown in **Supplementary Table SN2.2**. The top ranked mutation in EUR and CEU populations (rs405647) was ranked 1, 2, 3 in AMR, SAS, and EAS, respectively, and is tightly linked to rs642742 ($D' = 0.92$). Mutation rs661114 is ranked 2 in EUR, 5 in CEU, 6 in SAS, and 20 in AMR, and lies in a region with H3K27 acetylation that is associated with enhanced expression.

TRPV6. This region has been reported a target of selection in CEU population[5]. TRPV6 is involved in calcium absorption. It has been suggested that “Individuals with lighter skin pigmentation might have produced too much 1,25-dihydroxyvitamin D, resulting in an increased intestinal Ca2+ absorption. Thus, to reduce the risk of absorptive hypercalciuria with kidney stones, the derived haplotype would have spread only among individuals with lighter skin pigmentation”[25]. iSAFE suggests 10 strongly linked mutations located along a 9 kbp region located 84 kbp downstream of TRPV6 (see **Supplementary Fig. SN2.1**). These mutations are ranked in the top 10 in all non-African populations (**Supplementary Table SN2.3**). There is no signal of selection in this region in AFR. The pattern of selection in this region in global population along with the confidence and consistency of iSAFE results in all non-African populations is consistent with an out of Africa selection on this region with the favored mutation being near fixation in all non-African populations (**Supplementary Fig. 13**).

2.3 Population specific selection: East Asian.

PCDH15. This gene plays a role in development of inner-ear hair cells and maintaining retinal photoreceptors and is reported to be under selection in East Asian and a nonsynonymous mutation rs4935502 is proposed to be the favored variant[1]. This mutation is ranked 12 by iSAFE in CHB+JPT (see **Supplementary Fig. 9**, iSAFE = 0.45, $P < $1.34e-8). All top mutations are highly linked.

ADH1B. ”The ADH1B gene encodes one of three subunits of the Alcohol dehydrogenase (ADH1) protein, a major enzyme in the alcohol degradation pathway that catalyzes the oxidization of alcohols into aldehydes.” This region is a target of positive selection in East Asian population[5]. A non-synonymous mutation in this gene is associated with Alcohol dependence[26]. We tested this gene in CHB+JPT populations. iSAFE rank, in 2 Mbp around ADH1B gene, for the candidate mutation (rs1229984) is 8 (see **Supplementary Fig. 9**). The top rank mutation is an upstream mutation (rs3811801) 5 kbp upstream of the candidate mutation rs1229984 and highly linked to it ($D' = 0.99$). The second rank mutation (rs284787) is a 3′-UTR of ADH7 which is shown to be associated with Upper Aerodigestive Tract Cancers in a Japanese Population[27].

2.4 Population specific selection: UK

The UK Biobank project was recently investigated for regions under selection. The regions were reported as a target of a recent selection by analyzing the structure of UK Biobank and Ancient Eurasians[21]. We applied iSAFE on GBR (British in England and Scotland) population in 1000GP to check if the favored mutation could be confirmed.

ATXN2-SH2B3. Galinsky et al. proposed a nonsynonymous mutation (rs3184504) as a candidate that is associated to blood pressure[28]. We tested this region in GBR population of 1000GP. This candidate mutation is jointly ranked first with two other mutations rs7137828, rs7310615 (see **Fig. 3c**, iSAFE = 0.27, $P =$1.6e-7). rs7137828 is an intronic mutation in ATXN2 that is associated with Primary Open Angle Glaucoma that is a leading cause of blindness worldwide[29]. The other first rank mutation (rs7310615) is associated with blood expression levels of SH2B3[30]. Surprisingly, all of the top 10 mutations, ranked by iSAFE have a known association to a phenotype (**Supplementary Table SN2.4**), and are highly linked (**Supplementary Fig. SN2.2**).

CYP1A2/CSK. We tested a 5 Mbp region around these genes in GBR population of 1000GP. The proposed mutation rs1378942 by[21] with frequency 0.69 in GBR population is ranked 89 by iSAFE (iSAFE = 0.13, $P =$7.0e-5). The top-ranked mutation rs2470893 (**Fig. 3c**, iSAFE = 0.16, $P =$2.7e-5) is between CYP1A1 and CYP1A2 with frequency 0.40 in GBR and is associated with Caffeine metabolism[31]. rs2470893 and rs1378942 are in a strong LD ($D' = 0.91$).

FUT2. The signal of selection on 5 Mbp around this region in GBR population is very weak (**Supplementary Fig. 9**), with peak iSAFE = 0.026, $P = 0.009$. There is a very weak peak in 400 kbp around FUT2 gene (chr:49077276-49475876). The stop gained mutation rs601338 proposed as a candidate mutation by[21] is ranked 4 ($P = 0.1$).

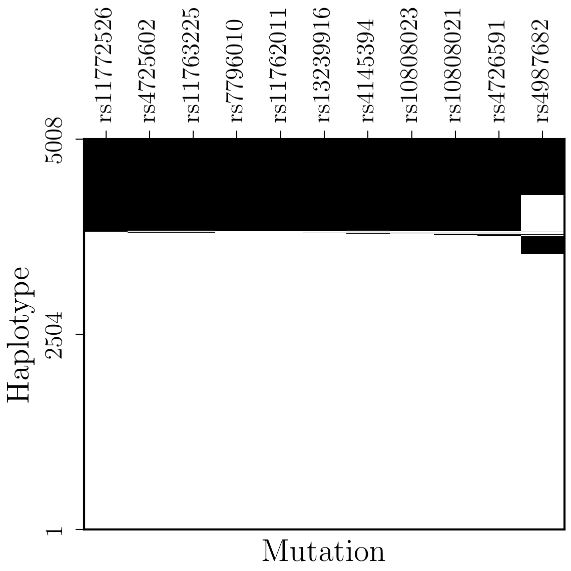
F12. The signal of selection on 5 Mbp around this region in GBR population is very weak (**Supplementary Fig. 9**, peak iSAFE = 0.027, $P = 0.008$). The proposed mutation rs2545801 has a very weak signal ($P = 0.2$).

Other genes

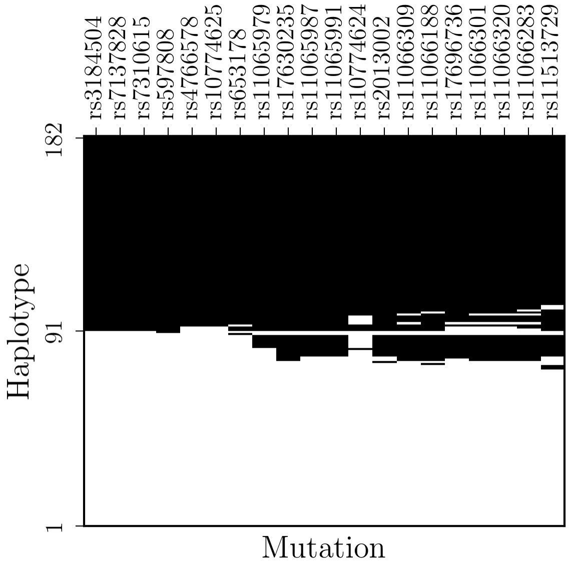
PSCA. This gene has been reported as a target of selection in YRI population[5]. A 5′UTR mutation rs2294008 proposed as a candidate favored mutation in this region that is associated with urinary bladder and gastric cancers[32,33]. The signal of iSAFE in 5 Mbp around this gene in YRI population is weak (see **Supplementary Fig. 9**, peak iSAFE = 0.04, $P =$2.4e-3). The proposed mutation rs2294008 is ranked 7 in 5 Mbp region surrounding this region. The local rank in 400 kbp around this gene is joint-first with 8 other mutations including rs2976392 which is also associated with diffuse-type gastric cancer[33]. Other mutations are rs2978979, rs2920279, rs2978980, rs2920282, rs2294010, rs2717562, rs2978982. This 9 mutation are fully linked in YRI population in a 20 kbp region that cover PSCA from upstream regulatory region to its down stream (chr8:143757286-143776668, GRCh37/hg19).

ASPM. This gene is reported to be a target of weak selection in GBR population[5]. The signal in 2 Mbp around this gene is very weak (see **Supplementary Fig. 9**, peak-iSAFE = 0.025,

$P = 0.01$). The proposed mutation rs41310927 has a very weak signal ($P = 0.4$). However, we do see a strong iSAFE signal 1.3 Mbp away from the ASPM gene.



Supplementary Fig. SN2.1:
SNP matrix of TRPV6 top candidates. Haplotypes of top 10 iSAFE mutations, and the proposed mutation (rs4987682) by[5], in 5 Mbp around TRPV6 in 2504 × 2 haplotypes of 1000GP are shown. These mutations are sorted by their iSAFE rank from left to right. iSAFE top 10 mutations span a 9 kbp region(chr7:142476441-142485399, GRCh37/hg19). White is derived and black is ancestral allele.



Supplementary Fig. SN2.2:
SNP matrix of ATXN2-SH2B3 top candidates. Haplotypes of top 20 iSAFE mutations in 5 Mbp around ATXN2-SH2B3 in GBR population are shown. These mutations are sorted by their iSAFE rank from left to right. They span a 1.07 Mbp region around ATXN2-SH2B3 region (chr12:111833788-112906415, GRCh37/hg19). White is derived and black is ancestral allele. Most of these mutations are associated to a phenotype (see **Supplementary Table SN2.4**).

Supplementary Table SN2.1:
iSAFE rank of putative favored variants of OCA2-HERC2. iSAFE rank of candidate mutations proposed by[19,23] in 1 Mbp region around OCA2-HERC2 that are associated with eye, hair, and skin pigmentation. Number of haplotypes in CEU, CHB, and JPT populations are 198, 206, and 208, respectively. Computation of empirical P value is provided in Online Methods.

ID	Association	Population	iSAFE Rank	P
rs916977	Blue eye	CEU	15	4.1E-5
rs1667394	Blue eye & blond hair	CEU	16	4.3E-5
rs1129038	Blue eye	CEU	21	6.2E-5
rs12913832	Blue eye, skin & hair	CEU	21	6.2E-5
rs4778138	Blue eye	CEU	70	1.6E-4
rs4778241	Blue eye	CEU	72	1.8E-4
rs1800414	Skin	CHB+JPT	122	2.6E-3

Supplementary Table SN2.2:
KITLG candidate variants. iSAFE rank of top mutations in 2 Mbp around KITLG gene. sorted by their Mean Reciprocal Ranks, calculated over EUR, SAS, EAS, and AMR. Only those with Mean Reciprocal Rank greater than 0.1 are shown (the candidate mutation rs642742 proposed by[24] is also reported in the last row). Frequency and iSAFE score for this region in all the 1000GP populations are provided in **Supplementary Fig.12**. Number of haplotypes in CEU, EUR, SAS, EAS, and AMR populations are 198, 1006, 978, 1008, and 694, respectively.

ID	iSAFE Rank EUR	iSAFE Rank SAS	iSAFE Rank EAS	iSAFE Rank AMR	Mean Reciprocal Rank EUR, SAS, EAS, AMR	iSAFE Rank CEU
rs405647	1	2	3	1	0.71	1
rs496859	4	1	2	12	0.46	7
rs61942772	10	57	1	94	0.28	22
rs560859	2	4	152	20	0.2	5
rs661114	2	6	151	20	0.18	5
rs11105020	8	3	32	5	0.17	23
rs10506957	17	22	46	2	0.16	2
rs7979311	5	5	156	20	0.11	3
rs1907702	22	20	45	3	0.11	8
rs642742	30	49	64	166	0.02	94

Supplementary Table SN2.3:
TRPV6 candidate variants. iSAFE rank of top mutations in 5 Mbp around TRPV6 gene, sorted by their Mean Reciprocal Ranks, calculated over EUR, SAS, EAS, and AMR. Number of haplotypes in CEU, EUR, SAS, EAS, and AMR populations are 198, 1006, 978, 1008, and 694, respectively.

ID	iSAFE Rank EUR	iSAFE Rank SAS	iSAFE Rank EAS	iSAFE Rank AMR	Mean Reciprocal Rank EUR, SAS, EAS, AMR	iSAFE Rank CEU
rs11772526	4.0	1.0	1.0	1.0	0.81	4.0
rs4725602	1.0	4.0	1.0	2.0	0.69	1.0
rs11763225	1.0	4.0	5.0	2.0	0.49	1.0
rs7796010	4.0	1.0	3.0	6.0	0.44	4.0
rs11762011	4.0	3.0	3.0	6.0	0.27	4.0
rs13239916	4.0	6.0	6.0	4.0	0.21	4.0
rs4145394	3.0	8.0	10.0	5.0	0.19	1.0
rs10808023	8.0	7.0	7.0	8.0	0.13	4.0
rs10808021	9.0	10.0	8.0	8.0	0.12	10.0
rs4726591	10.0	9.0	9.0	10.0	0.11	4.0

Supplementary Table SN2.4:
ATXN2-SH2B3 candidate variants. iSAFE rank of top 20 mutations in GBR population (182 haplotypes) of 1000GP in 5 Mbp around ATXN2-SH2B3 region and their association to diseases. Computation of empirical *P* value is provided in Online Methods.

ID	Rank	*P*	Gene	Function	GBR Frequency	Association	Reference
rs3184504	1	2.2e-7	SH2B3	missense	0.5	Blood pressure and hypertension, Coronary artery disease, & more	[34]
rs7137828	1	2.2e-7	ATXN2	intron	0.5	Primary open-angle glaucoma	[29]
rs7310615	1	2.2e-7	SH2B3	intron	0.5	Fibrinogen levels	[30]
rs597808	4	2.7e-7	ATXN2	intron	0.49	Systemic lupus erythematosus	[35]
rs4766578	5	3.0e-7	ATXN2	intron	0.51	Vitiligo	[36]
rs10774625	5	3.0e-7	ATXN2	intron	0.51	Systemic lupus erythematosus, Retinal vascular caliber	[35]
rs653178	7	3.1e-7		regulatory	0.5	Blood pressure and hypertension, Myocardial infarction, & more	[34]
rs11065979	8	4.4e-7		intergenic	0.47	Cancer (pleiotropy)	[37]
rs17630235	9	4.6e-7	TRAFD1	downstream	0.43	Body mass index	[38]
rs11065987	10	4.9e-7		intergenic	0.45	Tetralogy of Fallot, Coronary artery disease, & more	[39]
rs11065991	10	4.9e-7	BRAP	intron	0.45		
rs10774624	12	5.2e-7	RP3-473L9.4	intron,nc	0.52	Rheumatoid arthritis	[40]
rs2013002	13	8.2e-7	ALDH2	upstream	0.44		
rs11066309	14	1.1e-6	PTPN11	intron	0.45		
rs11066188	15	1.5e-6			0.43		
rs17696736	16	1.5e-6	NAA25	intron	0.46	Ischemic stroke, Type 1 diabetes, & more	[41]
rs11066301	17	1.9e-6	PTPN11	intron	0.46	Hematological parameters	[42]
rs11066320	17	1.9e-6	PTPN11	intron	0.46		
rs11066283	19	2.1e-6	RPL6	downstream	0.46		
rs11513729	20	2.2e-6	MAPKAPK5-AS1	downstream	0.45		

References

[1] Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883–886 (2010).

[2] Sturm, R. A. & Duffy, D. L. Human pigmentation genes under environmental selection. *Genome biology* **13**, 248 (2012).

[3] Fujimoto, A. *et al.* A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *Human genetics* **124**, 179–185 (2008).

[4] Bryk, J. *et al.* Positive selection in East Asians for an EDAR allele that enhances NF-κB activation. *PLoS One* **3**, e2209 (2008).

[5] Peter, B. M., Huerta-Sanchez, E. & Nielsen, R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genet* **8**, e1003011 (2012).

[6] Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nature genetics* **30**, 233–237 (2002).

[7] Olds, L. C. & Sibley, E. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Human molecular genetics* **12**, 2333–2340 (2003).

[8] Heffelfinger, C. *et al.* Haplotype structure and positive selection at TLR1. *European Journal of Human Genetics* **22**, 551–557 (2014).

[9] Wong, S. H. *et al.* Leprosy and the adaptation of human toll-like receptor 1. *PLoS Pathog* **6**, e1000979 (2010).

[10] McManus, K. F. *et al.* Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. *PLoS genetics* **13**, e1006560 (2017).

[11] Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The resistance factor to Plasmodium vivax in blacks: the Duffy-blood-group genotype, FyFy. *New England Journal of Medicine* **295**, 302–304 (1976).

[12] Ohashi, J., Naka, I. & Tsuchiya, N. The impact of natural selection on an ABCC11 SNP determining earwax type. *Molecular biology and evolution* **28**, 849–857 (2011).

[13] Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *science* **312**, 1614–1620 (2006).

[14] Network, M. G. E. Reappraisal of known malaria resistance loci in a large multicenter study. *Nature genetics* **46**, 1197–1204 (2014).

[15] Tishkoff, S. A. *et al.* Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**, 455–462 (2001).

[16] Ronen, R. *et al.* Predicting carriers of ongoing selective sweeps without knowledge of the favored allele. *PLoS Genet* **11**, e1005527 (2015).

[17] Fu, Y.-X. Statistical properties of segregating sites. *Theoretical population biology* **48**, 172–197 (1995).

[18] Wiuf, C. & Donnelly, P. Conditional genealogies and the age of a neutral mutant. *Theoretical population biology* **56**, 183–201 (1999).

[19] Wilde, S. *et al.* Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proceedings of the National Academy of Sciences* **111**, 4832–4837 (2014).

[20] Coop, G. *et al.* The role of geography in human adaptation. *PLoS Genet* **5**, e1000500 (2009).

[21] Galinsky, K. J., Loh, P.-R., Mallick, S., Patterson, N. J. & Price, A. L. Population structure of UK Biobank and ancient Eurasians reveals adaptation at genes influencing blood pressure. *The American Journal of Human Genetics* **99**, 1130–1139 (2016).

[22] Beleza, S. *et al.* Genetic architecture of skin and eye color in an African-European admixed population. *PLoS Genet* **9**, e1003372 (2013).

[23] Donnelly, M. P. *et al.* A global view of the OCA2-HERC2 region and pigmentation. *Human genetics* **131**, 683–696 (2012).

[24] Miller, C. T. *et al.* cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell* **131**, 1179–1189 (2007).

[25] Suzuki, Y. *et al.* Gain-of-function haplotype in the epithelial calcium channel TRPV6 is a risk factor for renal calcium stone formation. *Human molecular genetics* **17**, 1613–1618 (2008).

[26] Park, B. L. *et al.* Extended genetic effects of ADH cluster genes on the risk of alcohol dependence: from GWAS to replication. *Human genetics* **132**, 657–668 (2013).

[27] Oze, I. *et al.* Impact of multiple alcohol dehydrogenase gene polymorphisms on risk of upper aerodigestive tract cancers in a Japanese population. *Cancer Epidemiology and Prevention Biomarkers* **18**, 3097–3102 (2009).

[28] for Blood Pressure Genome-Wide Association Studies, I. C. *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).

[29] Bailey, J. N. C. *et al.* Genome-wide association analysis identifies TXNRD2, ATXN2 and FOXC1 as susceptibility loci for primary open-angle glaucoma. *Nature genetics* (2016).

[30] De Vries, P. S. *et al.* A meta-analysis of 120,246 individuals identifies 18 new loci for fibrinogen concentration. *Human molecular genetics* ddv454 (2015).

[31] Cornelis, M. C. *et al.* Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Molecular psychiatry* **20**, 647–656 (2015).

[32] Wu, X. *et al.* Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nature genetics* **41**, 991–995 (2009).

[33] Sakamoto, H. *et al.* Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nature genetics* **40**, 730–740 (2008).

[34] Levy, D. *et al.* Genome-wide association study of blood pressure and hypertension. *Nature genetics* **41**, 677–687 (2009).

[35] Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature genetics* (2015).

[36] Jin, Y. *et al.* Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nature genetics* **44**, 676–680 (2012).

[37] Fehringer, G. *et al.* Cross-Cancer Genome-Wide Analysis of Lung, Ovary, Breast, Prostate, and Colorectal Cancer Reveals Novel Pleiotropic Associations. *Cancer research* **76**, 5103–5114 (2016).

[38] Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

[39] Cordell, H. J. *et al.* Genome-wide association study identifies loci on 12q24 and 13q32 associated with tetralogy of Fallot. *Human molecular genetics* dds552 (2013).

[40] Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

[41] Dichgans, M. *et al.* Shared genetic susceptibility to ischemic stroke and coronary artery disease. *Stroke* **45**, 24–36 (2014).

[42] Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature genetics* **41**, 1182–1190 (2009).