Title

Looking Into the Past: Identifying Genetic Mutations and Introgression Events that Shaped Human Adaptation

Author

Akbari, Ali

Publication Date

2018

UNIVERSITY OF CALIFORNIA SAN DIEGO

Looking Into the Past: Identifying Genetic Mutations and Introgression Events that Shaped Human Adaptation

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Intelligent Systems, Robotics, and Control)

by

Ali Akbari

Committee in charge:

Professor Vineet Bafna, Chair
Professor Siavash Mirarab, Co-Chair
Professor Glenn Tesler, Co-Chair
Professor Gabriel G. Haddad
Professor William S. Hodgkiss

2018

The dissertation of Ali Akbari is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____
Co-Chair

____
Co-Chair

____
Chair

University of California San Diego

2018

DEDICATION

To my loving and caring wife, Nasrin,

my hero, my mother,

and the memory of my father.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to my advisor, Vineet Bafna. Thank you for your insightful advice, constant support, and encouragement over the past years. I am very fortunate to have been mentored by such a superb scientist with an amazing personality. You are truly the best mentor I could ever ask for.

I would like to thank my thesis committee for their guidance and encouragement along the way; Vineet Bafna, Siavash Mirarab, Glenn Tesler, Gabriel Haddad, and William Hodgkiss. In particular, I would like to thank Siavash Mirarab for his support and valuable advice during the past years and for being a wonderful collaborator. I am also very thankful for having met with Gabriel Haddad, for his support and a wonderful collaboration on high-altitude adaptation in Andean highlanders. I deeply appreciate and thank Glenn Tesler for his constructive suggestions and for being a great collaborator.

I would like to thank my beloved wife, Nasrin. Thank you for your unconditional emotional support and encouragement for everything, and especially I cannot thank you enough for all the sacrifices you have made for me throughout this experience, with all ups and downs. Thank you for being such a wonderful friend and always cheering me up. I am so lucky and blessed to have you in my life.

Finally, everything I have accomplished so far would not have been possible without relentless devotion of my family, especially my mother, who is a fighter and thought me how to fight against all odds.

Chapter 2, in full, contains material from Roy Ronen*, Glenn Tesler*, Ali Akbari*, Shay Zakov, Noah A Rosenberg, and Vineet Bafna. “Predicting carriers of ongoing selective sweeps without knowledge of the favored allele” *PLoS genetics* (2015) [1]. I was a co-primary investigator and author of this paper. * denotes equal contributions.

Chapter 3, in full, contains material from Ali Akbari, Joseph J Vitti, Arya Iranmehr, Mehrdad Bakhtiari, Pardis C Sabeti, Siavash Mirarab, and Vineet Bafna. “Identifying the favored

mutation in a positive selective sweep" *Nature methods* (2018) [2]. I was the primary investigator and author of this paper.

Chapter 4, in full, contains material from Ali Akbari, Joseph J Vitti, Arya Iranmehr, Mehrdad Bakhtiari, Pardis C Sabeti, Siavash Mirarab, and Vineet Bafna. "Identifying the favored mutation in a positive selective sweep" *Nature methods* (2018) [2]. I was the primary investigator and author of this paper.

Chapter 5, in full, contains material from Ali Akbari and Vineet Bafna. "Detecting adaptive introgression in human populations without knowledge of the archaic samples", which is currently being prepared for submission for publication of the material. I was the primary investigator and author of this material.

VITA

2013	B.Sc. in Electrical Engineering, Sharif University of Technology, Iran
2013	B.Sc. in Computer Science, Sharif University of Technology, Iran
2016	M.Sc. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California San Diego, USA
2018	Ph.D. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California San Diego, USA

PUBLICATIONS

* denotes equal contributions

Ali Akbari and Vineet Bafna. “Detecting adaptive introgression in human populations without knowledge of the archaic samples.” (In preparation).

Ali Akbari, Joseph J Vitti, Arya Iranmehr, Mehrdad Bakhtiari, Pardis C Sabeti, Siavash Mirarab, and Vineet Bafna. “Identifying the favored mutation in a positive selective sweep.” *Nature methods* (2018).

Priti Azad, Tsering Stobdan, Dan Zhou, Iain Hartley, Ali Akbari, Vineet Bafna, and Gabriel G Haddad. “High-altitude adaptation in humans: from genomics to integrative physiology.” *Journal of Molecular Medicine* (2017).

Arya Iranmehr, Ali Akbari, Christian Schlötterer, and Vineet Bafna. “Clear: composition of likelihoods for evolve and resequence experiments.” *Genetics* (2017).

Tsering Stobdan*, **Ali Akbari***, Priti Azad, Dan Zhou, Orit Poulsen, Otto Appenzeller, Gustavo F Gonzales, Amalio Telenti, Emily HM Wong, Shubham Saini, Ewen F Kirkness, J Craig Venter, Vineet Bafna, Gabriel G Haddad. “New insights into the genetic basis of monges disease and adaptation to high- altitude.” *Molecular biology and evolution* (2017).

Erika L Flannery, Tina Wang, Ali Akbari, Victoria C Corey, Felicia Gunawan, A Taylor Bright, Matthew Abraham, Juan F Sanchez, Meddly L Santolalla, G Christian Baldeviano, Kimberly A Edgel, Luis A Rosales, Andres G Lescano, Vineet Bafna, Joseph M Vinetz, Elizabeth A Winzeler. “Next-generation sequencing of Plasmodium vivax patient samples shows evidence of direct evolution in drug-resistance genes.” *ACS infectious diseases* (2015).

Roy Ronen*, Glenn Tesler*, **Ali Akbari***, Shay Zakov, Noah A Rosenberg, and Vineet Bafna. “Predicting carriers of ongoing selective sweeps without knowledge of the favored allele.” *PLoS genetics* (2015).

Roy Ronen, Glenn Tesler, Ali Akbari, Shay Zakov, Noah A Rosenberg, and Vineet Bafna. "Haplotype Allele Frequency (HAF) score: predicting carriers of ongoing selective sweeps without knowledge of the adaptive allele." *International Conference on Research in Computational Molecular Biology* (2015).

ABSTRACT OF THE DISSERTATION

Looking Into the Past: Identifying Genetic Mutations and Introgression Events that Shaped Human Adaptation

by

Ali Akbari

Doctor of Philosophy in Electrical Engineering (Intelligent Systems, Robotics, and Control)

University of California San Diego, 2018

Professor Vineet Bafna, Chair
Professor Siavash Mirarab, Co-Chair
Professor Glenn Tesler, Co-Chair

Adaptation is the central evolutionary process and is at the core of some of the greatest challenges facing humanity. HIV would likely cause nothing more than a harmless fever without the ability of the virus to adapt and eventually destroy the immune system. Cancer would be much more straightforward to treat if not for its ability to adapt to anti-cancer drugs. Malaria could be treated with cheap drugs such as Quinine instead of being one of the world's worst killers. In disease and health, we are in an arms race without fully understanding the rules of engagement. Humans have also adapted to live in harsh ecological niches, allowing them for example to digest

milk sugars in adulthood, and to live at high altitudes with debilitating lack of Oxygen.

Over 83 million people live at altitudes above 2,500 meters (8,200 ft) where the oxygen levels are 25% lower than at sea level. If not adapted, residing for a long time at such a harsh environment with low oxygen level can be fatal. One of the most striking example of high-altitude adaptation is the adaptation of Tibetan highlanders, where the favored genetic material is introgressed from archaic humans similar to Denisovans. Introgression is the introduction of genetic material into a population via interspecies mating. The complex pathways involved in hypoxia tolerance also inform upon our ability to understand ischemic diseases (stroke, cardiovascular diseases), and new molecular targets for these diseases. Therefore, this natural human experiment is a wonderful system to work with.

When adaptation is genetic (inherited by offspring of adapted individuals), it leaves a variety of detectable signatures in genomes. Together with recent developments in DNA sequencing technologies in past decades, methods for detecting genomic *regions* under selection from population genomic data, have been actively developed. In contrast, little work has been done to identify the favored mutation in a selective sweep. Pinpointing the favored mutation among tens of thousands of other, hitchhiking, mutations is like a needle in a haystack problem.

Identifying the favored mutation can provide a more precise picture of the origin of the selection; and allows people to do functional studies to improve the overall understanding of diseases. For example, adaptation to chronic hypoxia at high altitude can suggest targets for cardiovascular and other ischemic diseases. Also, identifying the favored mutation gives a high resolution picture revealing complicated evolutionary scenarios like multiple favored mutations and adaptive introgression.

Here in this dissertation, we address the challenging but important problem of pinpointing the favored mutation in a selective sweep. We break the problem into smaller parts and very carefully craft them to accurately identify the favored mutation in a selective sweep, and also distinguish adaptively introgressed haplotypes from other models of selection.

Chapter 1

Introduction

The human DNA can be thought of as a book that each of us inherited from our parents, with a complete set of genetic instructions. This book is written out in three billion letters using a 4 letter alphabet denoted by $\{A, C, G, T\}$. It would take thirty years of non-stop typing to create such a gigantic book. In making a copy for our children, we change or *mutate* about one hundred letters of our book at random. Most of these mutations are neutral but sometimes they can have a significant effect, either harmful or beneficial. For example, a few fortuitous mutations allowed adults to become lactose persistent – gain an ability to digest milk sugar (lactose) properly after weaning. Through selection, about one-third of human populations are now lactose persistent. Many studies have shed light on the genetic bases of the lactose persistence and there are a few mutations, in different populations, that are associated with lactose persistence. The spread of these mutations is probably due to *natural selection* – an adaptive response to the availability of a dairy products in cultures that domesticated milk-producing herds [3].

The advancement in DNA sequencing technology is accelerating our understanding of complex biological systems, and as Dobzhansky famously said "Nothing in biology makes sense except in the light of evolution." The ever growing resolution of genetic variation within populations, due to the genomic data influx, empowers scientist to address fundamental evolutionary questions. In the field of population genetics, since its foundation in the early twentieth century, scientists have been standing on the shoulders of mathematical geniuses – like Fisher, Wright, and Haldane – to understand the forces that determine evolution, by using theoretically determined quantitative models. During the past decades, population genetics has been experiencing a great resurgence of interest due to availability of high resolution genomic data through space (different geographically located populations) and time (ancient populations). This massive amount and complexity of genomic data is creating a new set of questions that require new computational techniques.

Together with recent developments in genomics and computational power, we are able to look deeper into the past through the lens of population genomics like never before. Analyzing the pattern of similarity between different genomic sequences, using proper computational methods, has dramatically improved our understanding of the origin of humans and prehistoric population trends. Now in the light of these advancements, we can see that the *Y-chromosomal Adam* and *mitochondrial Eve*[1] of modern humans lived, a few hundred thousand years ago, in Africa [4, 5]. Modern humans migrated out of Africa and spread across the globe, over the past 100 thousand years, and experienced many different habitats, and along the way encountered other archaic humans like Neanderthals[2]. There is solid scientific evidence showing that *introgression* – the transfer of genetic information from one species to another as a result of interbreeding – happened from archaic humans like Neanderthal to modern humans; and approximately 2% of the genome of non-African populations are introgressed from archaic humans like Neandertals [6].

Eventually, transition of modern humans' lifestyle from hunting-gathering to practicing agriculture and pastoralism, increased the ability to support large populations and led to rapid population growth during the past few thousand years [7]. New conditions introduce new

[1] The Y-chromosome is normally present only in males and inherited solely from the father. The mitochondrial DNA is a tiny portion of the DNA that is inherited solely from the mother. Therefore, the most recent common ancestor of living humans with respect to Y-chromosome (respectively, mitochondrial DNA) is called *Y-chromosomal Adam* (respectively, *mitochondrial Eve*). Y-chromosomal Adam and the mitochondrial Eve are not required to have lived at the same time.

[2] Neanderthals went extinct a few thousand years after they met modern humans.

challenges, and those who respond better, tend to survive and reproduce more due to natural selection. It was not always easy for modern humans to adjust to new environment with new constraints for survival [8]. One of the examples of natural selection at work in modern humans is the adaptation to living at high altitudes, over 2,500 meters or 8,200 ft, where the oxygen level decreases $\sim 25\%$ compared to the sea level.

More than 83 million people all around the globe live permanently at high-altitude where low oxygen level makes most people sick [9]. If not adapted, residing for a long time at such a harsh environment can be fatal[3]. Most of the people living at high-altitudes for generations have been adapted and do not have any problem regarding the constant exposure to the low oxygen level. Many studies have shed light on the genetic bases of these adaptation in different populations, using genomes of many individuals of highlanders [10–13]. One of the most striking examples of high-altitude adaptation is the adaptation of Tibetan highlanders [14], where the favored genetic material is introgressed from archaic humans similar to Denisovans[4].

The selection pressures for adapting to new environments have led to changes in the pattern of genetic variations of populations. Human genetic data have revealed a multitude of genomic regions believed to be evolving under positive selection in response to different selective pressure (such as dairy consumption, high-altitude, malaria, and many others) [8]. During the past decades, methods for detecting genomic regions under selection from population genetic data have been actively developed (Figure 1.1). In contrast, little work has been done to pinpoint the specific mutation favored by selection.

In this dissertation, we address the challenging but important problem of characterizing the favored allele in a selective sweep. We break the problem into smaller parts and very carefully craft them to accurately identify the favored mutation in a selective sweep and also distinguish

[3] Chronic mountain sickness (CMS or Monge's disease) is a disease that occurs in long-term high-altitude residents, and it can lead to heart failure.

[4] The Denisovans are an extinct archaic humans closely related to Neandertals. No one had suspected such a population existed until the DNA extracted from a tiny finger bone was sequenced. The finger bone belonged to a girl who lived about 41,000 years ago, in the Denisova cave in Siberia.

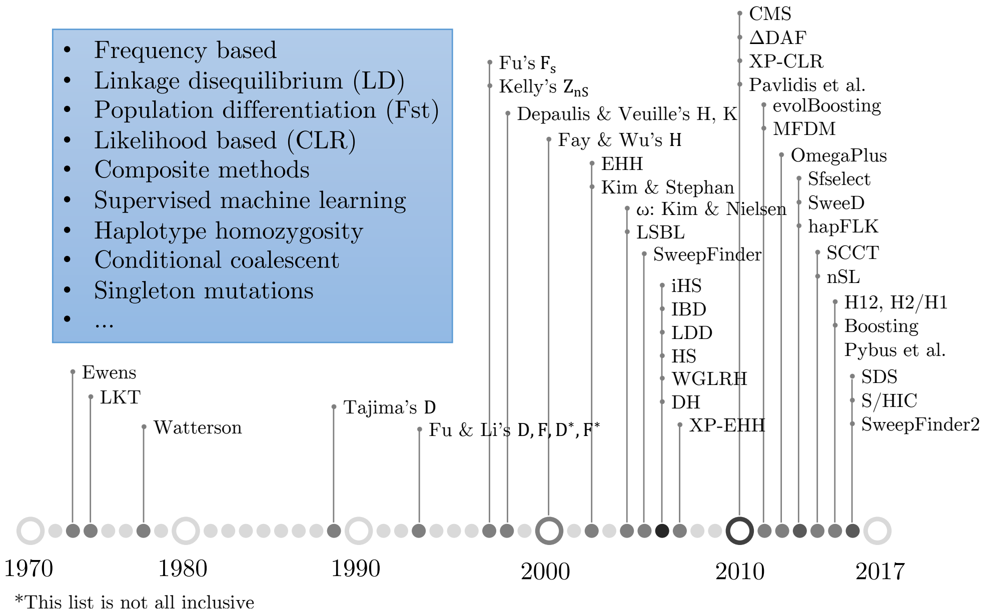


Figure 1.1: **Timeline of selection tests.** Methods to detect the *region* under selection have been actively developed during the past decades, together with the advent of deep sequencing, and some gained popularity. Each of these tests exploits a variety of genomic signatures, and can be fit into one or more of categories shown in the blue box. In contrast, little work has been done to identify the favored mutation in a selective sweep. In this dissertation, we address the challenging but important problem of characterizing the favored allele in a selective sweep.

adaptively introgressed haplotypes from other models of selection. Pinpointing the favored mutation 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 [10, 11].

1.1 Background

Analysis of the genomic data requires advanced data science skills, but also deeper insight into the underlying processes generating the data. Here I provide a short introduction to some of the technical terms used throughout this dissertation.

Neutral theory. The neutral theory of evolution is at the center of evolutionary studies at the molecular level [15]. The neutral theory holds that most variation at the molecular level – changes in genetic material itself – is neutral and best explained by a stochastic process called *genetic drift*, rather than natural selection. *Allele*[5] frequency changes are random in genetic drift and can lead to the fixation of some alleles and the loss of others. In contrast, in natural selection, allele frequency changes are forced in a specific direction due to the selective pressure. The neutral theory does not deny that natural selection occurs, but assumes only a tiny fraction of genomic mutations are adaptive – have significant impact on reproduction and survival.

Wright-Fisher model. The Wright-Fisher [16] model is a mathematical model of genetic drift, that assumes a fixed population-size with non-overlapping generations and the ancestors of the present generation are obtained by random sampling with replacement from the previous generation. The basic form of the model overlooks many realistic details like mutation, recombination[6], selection, population structure, and so forth.

Coalescent theory. Given two *haplotypes*[7] descended from a common ancestor in a generation in the past, the coalescent model [16–18] traces the ancestry of these two haplotypes, backwards in time, to the point where these two lineages coalesce in that generation. The coalescent model describes the ancestral relationships of samples and the time of coalescent events by a stochastic process. The model approximates the Wright-Fisher model, when the sample size is much smaller than the population size. Advanced models of coalescent theory include complex evolutionary and demographic models. The model has been widely used to simulate theoretical population [19] and provided a theoretical framework to infer population genetic parameters, such as migration, recombination, and population size [7].

Selective sweep. Genetic data from diverse human populations have revealed a multitude of genomic regions believed to be selective sweeps in response to a selective pressure. A mutation is *favored* when its carriers have higher fitness relative to non-carriers. A selective sweep can occur when a favored mutation rapidly increases in frequency due to natural selection. In the neighborhood of the favored mutation, neutral mutations on the same lineage as the favored

[5] An *allele* is a variant at a specific locus of the genome.

[6] *Recombination* is the rearrangement process of DNA pieces to produce new combinations of alleles.

[7] A *haplotype* is a set of DNA variations on a chromosome that are inherited together.

mutation hitchhike (are co-inherited) with the favored mutation, and increase in frequency. This hitchhiking effect leads to a loss of genetic diversity, increase in the *linkage disequilibrium* (LD)[8], and distortion in the pattern of allele frequencies [21–23].

Hard sweep. 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.

Soft sweep. Recently, the *soft sweep* model has generated significant interest [24–29]. A soft sweep occurs when multiple sets of hitchhiking alleles in a given region increase in frequency, rather than a single favored haplotype. Soft sweeps may take place by one or more of the following mechanisms: (i) selection from standing variation: a neutral segregating mutation, which exists on several haplotypic backgrounds, becomes favored due to a change in the environment; (ii) recurrent mutation: the favored mutation arises several times on different haplotypic backgrounds; or, (iii) multiple adaptations: multiple favored mutations occur on multiple haplotypic backgrounds. Several methods have been developed for detecting soft sweeps, as well as for distinguishing between soft and hard sweeps [28–33]. Throughout the dissertation, we restrict our attention to soft sweeps arising from standing variation.

Adaptive introgression. *Introgression* is the introduction of genetic material into a population via interspecies mating. Comparisons of DNA from archaic and modern humans suggest that archaic hominins (like Neanderthals and Denisovans) interbred with modern humans [34–41], and approximately 2% of the DNA in most of non-African modern populations comes from archaic hominins [6]. *Adaptive introgression* occurs when the introgressed haplotype is advantageous in the admixed population and rapidly rise in frequency due to positive selection. Adaptive introgression can create an evolutionary shortcut and lead to a faster response to selective pressure compared to classic models of selection [41]. The *EPAS1* locus in Tibetans highlanders is a striking example of adaptive introgression from a Denisovan-like archaic hominin [14]. The

[8] *Linkage disequilibrium* is the nonrandom association of alleles at different loci [20].

putative favored haplotype is at high frequency *only* in Tibetan highlanders in response to the hypoxic environments at high altitude. In chapter 5, we present a novel method to identify the adaptively introgressed haplotype without knowing the archaic samples.

1.2 Dissertation overview

Methods for detecting genomic regions under selection from population genetic data have been actively developed [42–63], and exploit a variety of genomic signatures (see Figure 1.1). Allele frequency based methods analyze the distortion in the site frequency spectrum [42–46]; Linkage Disequilibrium (LD) based methods use extended homozygosity in haplotypes [47, 48]; population differentiation based methods use difference in allele frequency between populations; and finally, composite methods [56, 57] combine multiple test scores to improve the resolution. Recently, a lack of rare (singleton) mutations has been used to detect very recent selection [58]. Together with the advent of deep sequencing, these methods have identified multiple regions believed to be under selection in humans and other organisms, and provide a window into genetic adaptation and evolution [64–73].

In contrast, little work has been done to identify the favored mutation in a selective sweep. The high LD around the favored locus makes it easier to detect the region under selection [47, 48], but harder to pinpoint the favored mutation [28]. Grossman et al. [56] note that different selection signals identify overlapping but different regions, and a composite of multiple signals (CMS) can localize the site of the favored mutation. An alternative strategy is to use functional information to annotate SNPs and rank them in order of their functional relevance. However, the signal of selection is often spread over a large region, up to 1–2 Mbp on either side [28], and the high LD makes it difficult to pinpoint the favored mutation.

Most approaches that capture signatures of selective sweeps in population genomics data do not identify the specific mutation favored by selection. In this dissertation, we address the challenging but important problem of characterizing the favored allele in a selective sweep. We break the problem into smaller parts and very carefully craft them to accurately identify the favored mutation in a selective sweep and also distinguish adaptively introgressed haplotypes from other models of selection. Following shows the step-by-step procedure:

Predicting carriers of the favored mutation. In Chapter 2, we present a haplotypic score, HAF (Haplotype Allele Frequency), that can be used to separate carrier haplotypes from non-carriers without knowledge of the favored mutation. The HAF score, assigned to individual haplotypes in a sample, naturally captures many of the properties shared by haplotypes carrying a favored allele. The HAF score is well-correlated with the relative fitness of individual haplotypes, and is very effective at predicting carriers of selective sweeps. We provide a theoretical framework for computing expected HAF scores under different evolutionary scenarios, which lay the mathematical foundation for the next steps.

Identifying the favored mutation in a small region. In Chapter 3, using properties of the HAF score, we present the SAFE (Selection of Allele Favored by Evolution) score to identify the favored mutation in a small region ($\sim$50 kbp), with few or no recombinations. The SAFE score tends to be maximized for the favored mutation in a small region, but shows decaying performance when larger regions are investigated, due to the proportional increase of recombination events with region size.

Identifying the favored mutation in a large region. In Chapter 4, to address the more general case of large regions under selection, we present iSAFE (integrated Selection of Allele Favored by Evolution), which uses the SAFE score as a building block to pinpoint the favored mutation within a 5 Mbp around the region under selection. iSAFE exploits coalescent based signals in 'shoulders' [28] of the selective sweep (genomic regions proximal to the region under selection, but carrying the selection signal) to rank all mutations within a large ($\sim$5 Mb) region based on their contribution to the selection signal. iSAFE proved to be very powerful to pinpoint the favored mutation in a selective sweep. iSAFE does not require knowledge of demography, the phenotype under selection, or functional annotations of mutations.

Identifying adaptively introgressed haplotypes. Finally, in Chapter 5, using the iSAFE score as the main feature of a supervised learning approach, we present a method, CHAI (Capturing Haplotypes Adaptively Introgressed), to identify adaptively introgressed haplotypes in human populations without having knowledge of the archaic samples.

1.3 Summary

Methods to detect the *region* under selection have been actively developed during the past decades. However, the identification of the *favored allele* in a selective sweep is a long-standing problem in population genomics. In my doctoral dissertation we address this challenging but important problem, using a step-by-step procedure. We show that statistics obtained from the coalescent structure of a region under a selective sweep can indeed pinpoint the favored mutation, and also distinguish adaptively introgressed haplotypes from other models of selection. Pinpointing the favored mutation 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.

Chapter 2

Predicting carrier haplotypes of the favored mutation

Methods for detecting the genomic signatures of natural selection have been heavily studied (Figure 1.1, [42–63]), 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 this chapter we present 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. Using both simulated and real data, we show that the HAF score is well-correlated with the *relative* fitness of individual haplotypes, and is very effective at predicting carriers of selective sweeps. Later, in Chapters 3 and 4 we use some properties of HAF score distribution and present a new statistics to pinpoint the favored mutation in a selective sweep.

2.1 Motivation

With advances in genome sequencing, we now have an opportunity to more completely sample genetic diversity in human populations, and probe deeper for signatures of adaptive evolution [61–63]. Genetic data from diverse human populations in recent years have revealed a multitude of genomic regions believed to be evolving under recent positive selection [64–73].

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 [21, 22]. The site frequency spectrum (SFS) within and across populations is often used as a basis for such inference [42–46]. Methods based on haplotype structure have been developed using a variety of approaches, including the frequency of the most common haplotype [74], the number and diversity of distinct haplotypes [75], the haplotype frequency spectrum [76], and the popular approach of long-range haplotype homozygosity [47–49].

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.

2.2 Methods

Consider a sample of haplotypes in a genomic region. We assume that all sites are biallelic, and at each site, we denote ancestral alleles by 0 and derived alleles by 1. We also assume that all sites are polymorphic in the sample. The *HAF vector* of a haplotype *h*, denoted **c**, is obtained by taking the binary haplotype vector and replacing non-zero entries (derived alleles carried by the haplotype) with their respective frequencies in the sample (Fig 2.1**a**). We define the HAF *score* of **c** as:

$$\mathrm{HAF}(\mathbf{c}) = \sum_j \mathbf{c}_j \tag{2.1}$$

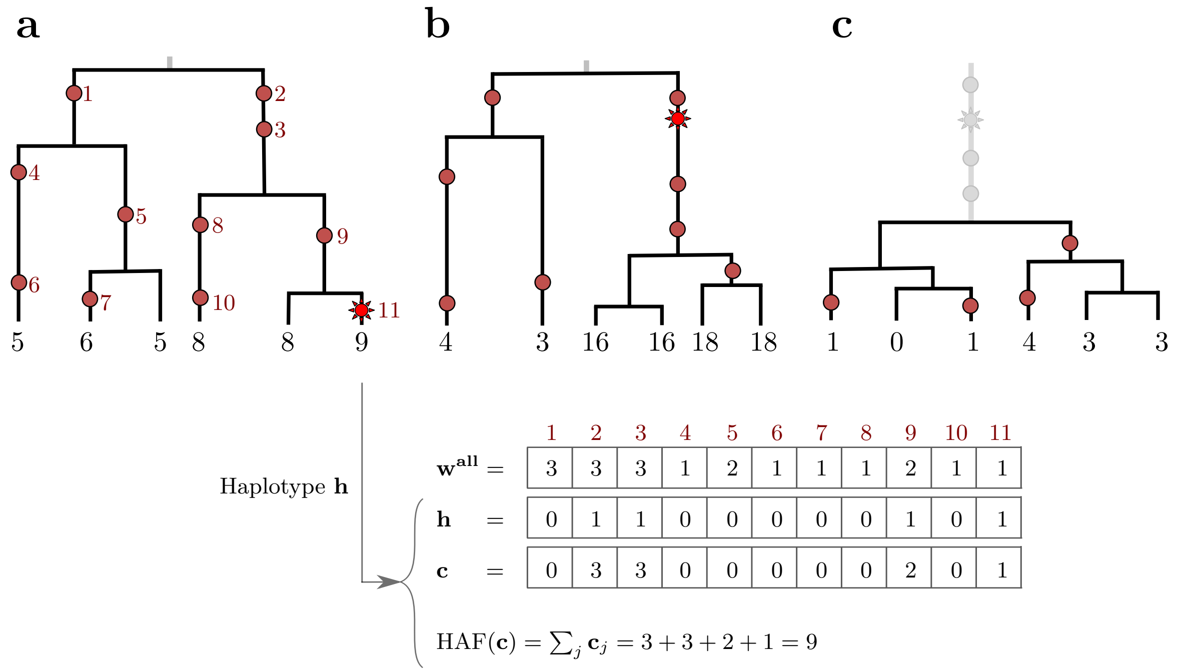
where the sum proceeds over all segregating sites j in the genomic region. The HAF score of a haplotype amounts to the sum of frequencies of all derived alleles carried by the haplotype.

Below, we start with a theoretical explanation of the behavior of the HAF score under different evolutionary scenarios, validating our results using simulation. 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.

2.3 Results

Theoretical 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 $\alpha = 2N_0 r$



	1	2	3	4	5	6	7	8	9	10	11
$\mathbf{w}^{\text{all}} =$	3	3	3	1	2	1	1	1	2	1	1
$\mathbf{h} =$	0	1	1	0	0	0	0	0	1	0	1
$\mathbf{c} =$	0	3	3	0	0	0	0	0	2	0	1

$\text{HAF}(\mathbf{c}) = \sum_j \mathbf{c}_j = 3+3+2+1 = 9$

Figure 2.1: **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 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 HAF score. Vector $\mathbf{w}^{\text{all}}$ lists the frequencies of all mutations. (**a**) The favored allele appears on a single haplotype. At this point in time, both the genealogy and the HAF score distribution are largely neutral. (**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.

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 HAF score. We use simulations to demonstrate the concordance of theoretical and empirical values of the HAF score, and show that the values are robust to the presence of recombination (see Methods section of Ronen et al. [1] for parameter choices).

Expected HAF score under neutrality

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 to the HAF score of each of the w haplotypes with the mutation, and contributes 0 for each of the other $n - w$ haplotypes. The mean of the HAF scores of all n haplotypes in the sample is

$$\frac{1}{n}\sum_{v=1}^{n} \mathrm{HAF}(\mathbf{c}(v)) = \frac{1}{n}\sum_{w=1}^{n-1} \xi_w \cdot w \cdot w = \frac{1}{n}\sum_{w=1}^{n-1} \xi_w \cdot w^2. \tag{2.2}$$

Under the coalescent model, [77, 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 HAF score is computed as

$$\mathbb{E}[\mathrm{HAF}] = \frac{1}{n}\sum_{w=1}^{n-1} \mathbb{E}[\xi_w] \cdot w^2 = \frac{\theta}{n}\sum_{w=1}^{n-1} w = \frac{\theta(n-1)}{2}. \tag{2.3}$$

HAF score dynamics in selective sweeps

We now consider the dynamics of HAF scores in a population undergoing a selective sweep. Fig 2.2 illustrates the HAF score dynamics in a single simulated population undergoing a hard sweep. Initially (leftmost, time 0) the HAF scores of carriers and non-carriers of the favored allele are similar. As the sweep progresses (times 100–500), carrier HAF scores increase. Soon after fixation (time ∼500), we observe a sharp decline in HAF scores, followed by slow and steady recovery due to new mutation and drift (times 600–50,000).

Below, we provide a theoretical description of the HAF score dynamics during an ongoing selective sweep, as well as empirical validation using simulations. As the selective sweep progresses, the value of the HAF score of haplotypes carrying the favored allele increases. Consider n haplotypes sampled from a fixed population of N haploid individuals under a selective sweep and assume that there is no recombination.

We let ν denote the fraction of carrier haplotypes in the sample. When $\nu \leq 1/n$ (i.e., 0 or 1 carrier haplotypes), there is no selection going back in time, and the time to MRCA can be

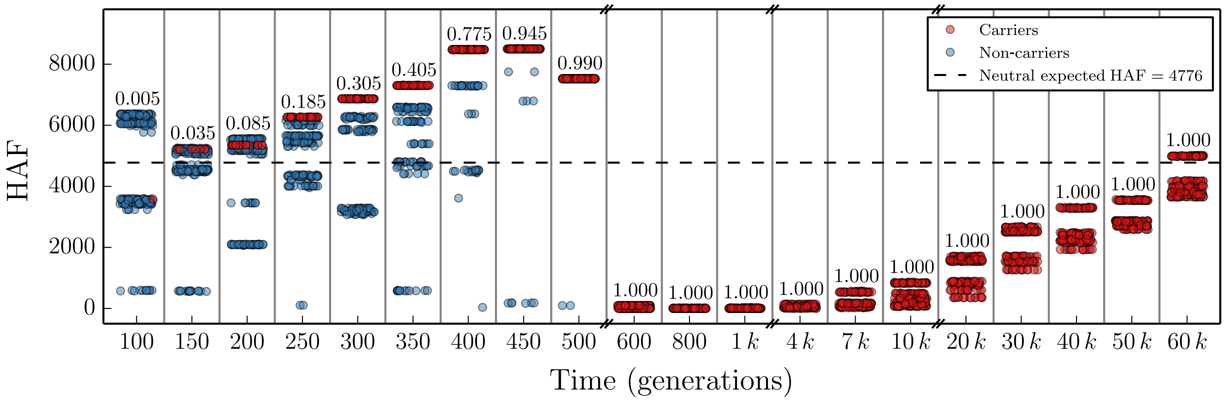


Figure 2.2: **Schematic of HAF score dynamics**. We consider HAF scores in 50 kb segments, examining $n = 200$ haplotypes sampled from a constant-sized ($N = 20000$ haploids) population, evolving with population-scaled mutation rate $\theta = 48$, selection coefficient $s = 0.05$, and no recombination ($\rho = 0$). 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 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 HAF score in the neutral population.

computed using the neutral Wright-Fisher model [16]. The expected HAF scores for carriers and non-carriers are identical (Eq. (2.3)). At the time when ν first equals 1, there are no non-carriers, and the HAF scores are given by the exponential growth model (See Figure 3 of Ronen et al. [1]). Below, we model the HAF scores for all intermediate values of ν.

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

$$\mathbb{E}[\text{HAF}^{\text{car}}] \approx \theta n \left(\frac{\nu + 1}{2} - \frac{1}{(1-\nu)n + 1} \right), \tag{2.4}$$

$$\mathbb{E}[\text{HAF}^{\text{non}}] \approx \theta n \left(\frac{1}{2} + \frac{1}{2n} - \frac{1}{(1-\nu)n + 1} \right). \tag{2.5}$$

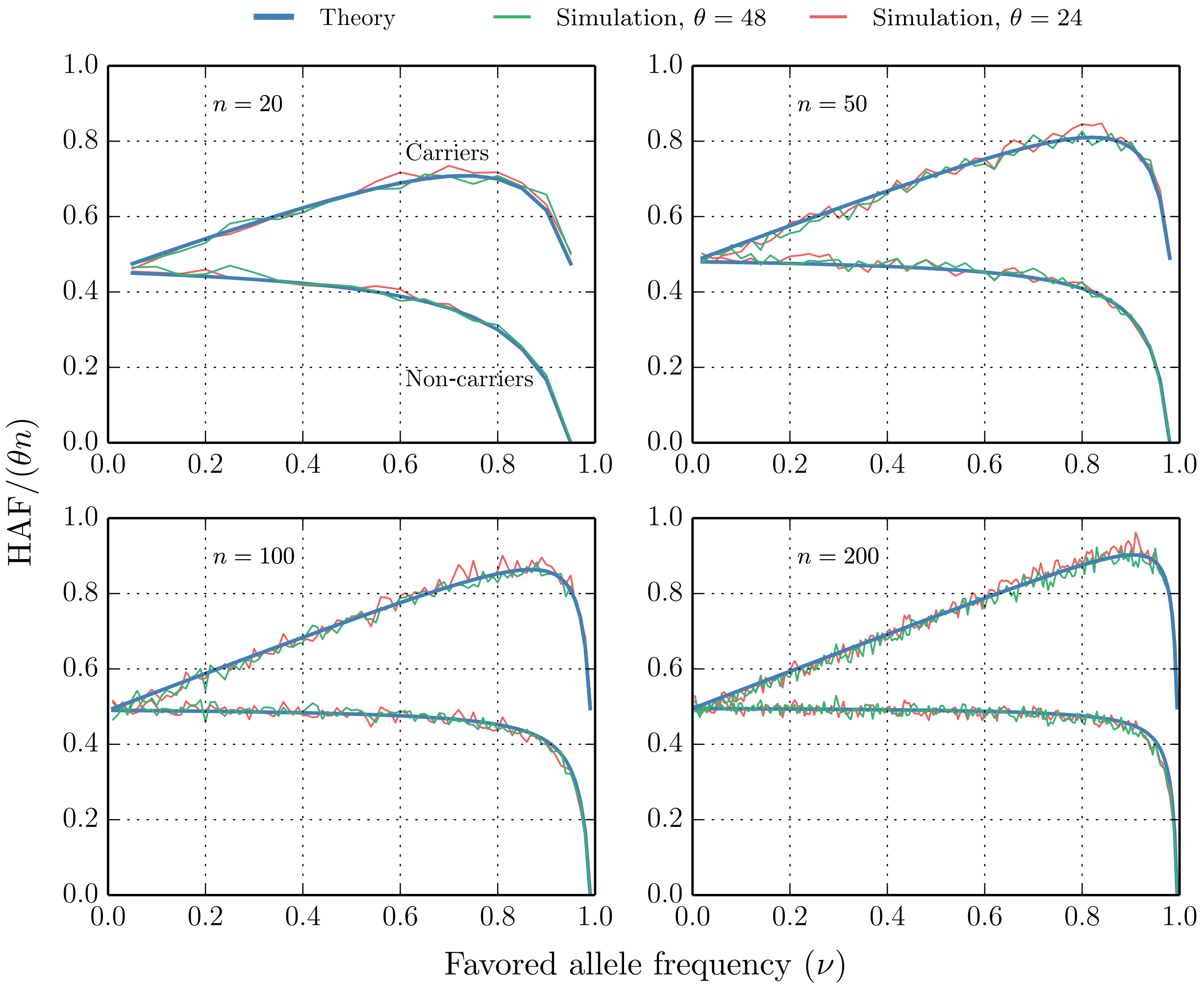


Figure 2.3: **Dynamics of expected HAF score during a selective sweep.** For each (θ, n, ν) with $\theta \in \{24, 48\}$, $n \in \{20, 50, 100, 200\}$, $\nu \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\}$, $s = 0.01$, and $N = 20,000$, we plotted the mean value of $\mathrm{HAF}/(\theta n)$ over 1000 trials, for both carriers and non-carriers, and compared against the theoretical values (Eqs. (2.4), (2.5)).

In Figure 2.3, we simulated selective sweeps for a variety of parameters and compared their trajectories against these results. The results show a tight correspondence between theory and empirical observations.

2.4 Summary and discussion

In this chapter we only focused on the behavior of the HAF score in a constant population size. In the Ronen et al. [1], we provide a theoretical framework for computing expected HAF scores under different evolutionary scenarios, including variable population size, and we validate the theoretical predictions with simulations. In the same paper, 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.

This chapter introduced 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 is very effective at predicting carriers of selective sweeps. Later in the Chapter 3, we use some properties of HAF score distribution, for carrier and non-carrier haplotypes, and present a new statistic to identify the favored mutation in a small window ($\sim$50 kbp).

2.5 Acknowledgments

Chapter 2, in full, contains material from Roy Ronen*, Glenn Tesler*, Ali Akbari*, Shay Zakov, Noah A Rosenberg, and Vineet Bafna. “Predicting carriers of ongoing selective sweeps without knowledge of the favored allele” *PLoS genetics* (2015) [1]. I was a co-primary investigator and author of this paper. * denotes equal contribution.

Chapter 3

Identifying the favored mutation in a *small* region

Methods to identify signatures of selective sweeps in population genomics data have been actively developed (Figure 1.1, [42–63]), but mostly do not identify the specific mutation favored by the selective sweep. We present SAFE (Selection of Allele Favored by Evolution), a method that uses a statistic derived solely from population genetics signals to pinpoint the favored mutation in a small region (∼50 kbp). SAFE was tested extensively on simulated data; the median SAFE rank of the favored mutation in a 50kbp region was 1 out of ∼250 variants, and the favored mutation was in the top 5 in 91% of simulations. In comparison, the median ranks of iHS (integrated Haplotype Score) [47] and SCCT (Selection detection by Conditional Coalescent Tree) [59] were 6 and 3, respectively. SAFE does not require knowledge of demography, the phenotype under selection, or functional annotations of mutations. The more general case of large regions (∼5 Mbp) under selection is addressed in Chapter 4.

3.1 Motivation

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 [42–46], linkage-disequilibrium-based methods use extended homozygosity in haplotypes [47, 48], other methods use differences in allele frequency between populations, and finally, composite methods combine multiple test scores to improve resolution [56, 57]. Recently, a lack of rare (singleton) mutations was used to detect very recent selection [58].

The signature of a selective sweep can be captured even when standing variation or multiple de novo mutations create a ‘soft’ sweep [24–29] 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 [64–73]. For example, adaptation to chronic hypoxia at high altitude can suggest targets for cardiovascular and other ischemic diseases [10, 11].

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 and conduct follow-up investigations. Not much work has been done to identify the favored mutation in a selective sweep. Grossman et al. [56] 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 12 Mbp on either side [28], and the high linkage disequilibrium makes it difficult to pinpoint the favored mutation.

In Chapter 2 we showed that 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 [1]. In this chapter, using properties of the HAF score, we developed a method, SAFE, that tends to be maximized for the favored mutation in a small region (50 kbp), but shows decaying performance when larger regions are investigated. In Chapter 4, we will further refine our method to address the more general case of large regions ($\sim$5 Mbp) under selection.

3.2 Methods

We consider only biallelic sites, taking as input a binary SNP matrix with each row corresponding to a haplotype h, each column to a site e. Entries in the matrix correspond to the allelic state, with 0 denoting the ancestral allele, and 1 denoting the derived allele. A haplotype 'contains/carries a mutation e' if it has the derived allele at site e.

In Chapter 2 we discussed that the HAF score is a haplotypic score that can be used to separate carrier haplotypes from non-carriers without knowledge of the favored mutation. The HAF score for a haplotype h ($\mathrm{HAF}(h)$) is the sum of the derived allele counts of mutations in h (Figure 3.1). It has been shown that, when h is a carrier of the favored allele, $\mathrm{HAF}(h)$ increases with the frequency of the favored mutation (Eq. 3.6), in contrast to HAF scores of non-carriers (Eq. 3.7), and this can be used to separate carrier haplotypes from non-carriers without knowing the favored mutation.

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 $\kappa(e)$

(Figure 3.1) denote the fraction of distinct haplotypes that carry mutation e.

$$\begin{aligned}\kappa(e) &= \frac{\left|\bigcup_h \{M_{h,e} \cdot \mathrm{HAF}(h)\}\right| - 1}{\left|\bigcup_h \{\mathrm{HAF}(h)\}\right|} \\ &= \frac{\text{\# of distinct haplotypes carrying mutation } e}{\text{\# of distinct haplotypes in sample}}\ . \end{aligned} \tag{3.1}$$

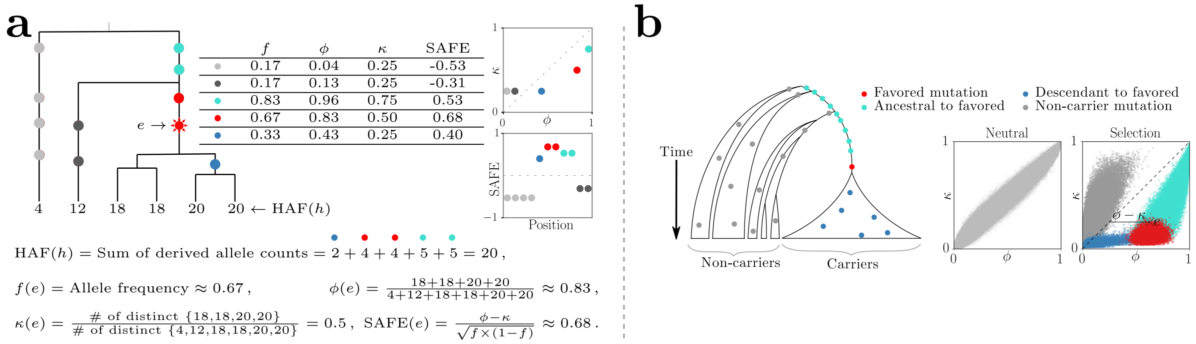
Similarly, let $\phi(e)$ denote the normalized sum of HAF scores of all haplotypes carrying the mutation e.

$$\begin{aligned}\phi(e) &= \frac{\sum_h [M_{h,e} \cdot \mathrm{HAF}(h)]}{\sum_h \mathrm{HAF}(h)} \\ &= \frac{\text{sum of HAF scores of haplotypes carrying mutation } e}{\text{sum of HAF scores of all haplotypes}}\ . \end{aligned} \tag{3.2}$$

We observe empirically that in a region evolving according to a neutral Wright-Fisher model [16], $\kappa(e)$ and $\phi(e)$ are both estimators of f_e. Moreover, empirical results (see Supplementary Figure C.1) suggest that the expected value of $\phi(e) - \kappa(e)$ is 0, and variance is proportional to $f_e(1 - f_e)$. Based on these observations, we define the SAFE score of mutation e as

$$\mathrm{SAFE}(e) = \frac{\phi(e) - \kappa(e)}{\sqrt{f_e(1 - f_e)}}\ . \tag{3.3}$$

Empirically, SAFE(e) behaves like a Gaussian random variable, with mean 0, under neutrality (Supplementary Figure C.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 (Figure 3.1). 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,



	f	ϕ	κ	SAFE
●	0.17	0.04	0.25	-0.53
●	0.17	0.13	0.25	-0.31
●	0.83	0.96	0.75	0.53
●	0.67	0.83	0.50	0.68
●	0.33	0.43	0.25	0.40

$$\mathrm{HAF}(h) = \text{Sum of derived allele counts} = 2+4+4+5+5 = 20\,,$$

$$f(e) = \text{Allele frequency} \approx 0.67\,, \qquad \phi(e) = \frac{18+18+20+20}{4+12+18+18+20+20} \approx 0.83\,,$$

$$\kappa(e) = \frac{\#\text{ of distinct }\{18,18,20,20\}}{\#\text{ of distinct }\{4,12,18,18,20,20\}} = 0.5\,,\ \ \mathrm{SAFE}(e) = \frac{\phi-\kappa}{\sqrt{f\times(1-f)}} \approx 0.68\,.$$

Figure 3.1: **Characterization of the SAFE method.** (**a**) The HAF score for haplotype *h* is the sum of the derived allele counts of the mutations on *h*. Carriers of the favored mutation have a higher fraction of the total HAF score of the sample (high ϕ) and fewer distinct haplotypes compared with non-carriers (low κ). (**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 ϕ versus κ values 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 Supplementary Note B.1). The joint distribution of ϕ and κ 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, ϕ and κ.

or 'descendant', if they arise after (Figure 3.1). Representing each mutation as a point in a 2-dimensional plot of ϕ, κ values, these classes are clustered differentially (Figure 3.1**b**). The selective sweep reduces the number of distinct haplotypes carrying the favored mutation (lower κ), leaving non-carrier mutations with an increased fraction of distinct haplotypes (higher κ). On the other hand, increased HAF scores in carrier haplotypes reduces the proportion of total HAF score contributed by non-carrier haplotypes (lower ϕ). In contrast, the favored mutation has high positive value of $\phi - \kappa$ due to high HAF scores for carriers (higher ϕ), and the reduced number of distinct haplotypes among its descendants (lower κ). As we go up to ancestral mutations, the number of non-carrier haplotype descendants increase, and κ grows faster than ϕ. As we go down to descendant mutations, there is a reduction in the already small number of distinct haplotypes. However, ϕ decreases sharply, reducing $\phi - \kappa$ (see Figure 3.1). Thus, we expect that the mutation with the highest SAFE score is a strong candidate for the favored mutation.

Theoretical and empirical modeling of the SAFE

To explain the behavior of the SAFE score in pinpointing 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.

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 [16] (constant population size, random mating, discrete generations, no recombination), with scaled mutation rate θ. From Eq. 2.3, the expected HAF score is,

$$\mathbb{E}[\text{HAF}] = \frac{\theta(n-1)}{2}. \tag{3.4}$$

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 [78]).

Empirical analysis on neutral coalescent simulations conditioned on the mutation e having fn carriers reveals that (Supplementary Figure C.1**a**)

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

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 Figure C.1**a**)

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

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

$$\mathbb{E}[\phi(e) - \kappa(e)] \approx 0\,. \tag{3.5}$$

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 Figure C.1**b**).

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 in Chapter 2. It increases dynamically upto fixation of the favored allele, and then decreases dramatically. Formally, let HAF^{car} (respectively, HAF^{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 Chapter 2 we showed that under strong selection ($Ns \gg 1$) and no recombination,

$$\mathbb{E}[\mathrm{HAF}^{\mathrm{car}}] \approx \theta n \left(\frac{f+1}{2} - \frac{1}{(1-f)n+1} \right), \tag{3.6}$$

$$\mathbb{E}[\mathrm{HAF}^{\mathrm{non}}] \approx \theta n \left(\frac{1}{2} + \frac{1}{2n} - \frac{1}{(1-f)n+1} \right). \tag{3.7}$$

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 [1]. 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

$$\phi(e) \approx \frac{fn\mathbb{E}[\mathrm{HAF}^{\mathrm{car}}]}{fn\mathbb{E}[\mathrm{HAF}^{\mathrm{car}}] + (1-f)n\mathbb{E}[\mathrm{HAF}^{\mathrm{non}}]} \approx \frac{f^2+f}{f^2+1} = f + \frac{f^2(1-f)}{f^2+1} \geq f. \tag{3.8}$$

For a population undergoing a positive natural selection with favored mutation e, $\phi(e)$ overestimates the favored allele frequency f (Figure 3.1 and Eq. 3.8). On the other hand, $\kappa(e)$ underestimates f (Figure 3.1). Therefore, we expect the distribution of $(\phi - \kappa)$ for the favored allele to be skewed in positive direction.

3.3 Results

We performed extensive simulations to test SAFE on samples evolving neutrally and under positive selection. We varied one parameter in each run (see Supplementary Note B.1), including window size ($L = 50$ kbp), number of individual haplotypes ($n = 200$) chosen from a larger effective population size ($N = 20,000$), scaled selection coefficient ($Ns = 500$), and initial and final favored mutation frequencies ($\nu_0 = 1/N$, and ν).

Only a few tests have been developed to identify or localize the favored mutation: Composite of Multiple Signals (CMS) [56], and Selection detection by Conditional Coalescent Tree (SCCT) [59]. CMS combines statistics from different selection tests, including the integrated Haplotype Score (iHS) [47], so as to localize the signal. In order to develop a unified probabilistic model, CMS expects control populations as input, as well as demographic models, and cannot be used directly on data based solely on coalescent simulations. Therefore, we compared SAFE against iHS and SCCT to obtain a baseline comparison here. The median SAFE rank of the favored mutation in a 50kbp region was 1 out of ~250 variants (left panel of Figure 3.2), and the favored mutation was in the top 5 in 91% of simulations. In comparison, the median ranks of iHS and SCCT were 6 and 3, respectively. The comparisons to CMS using simulated models of human demography are described later, in Chapter 4.

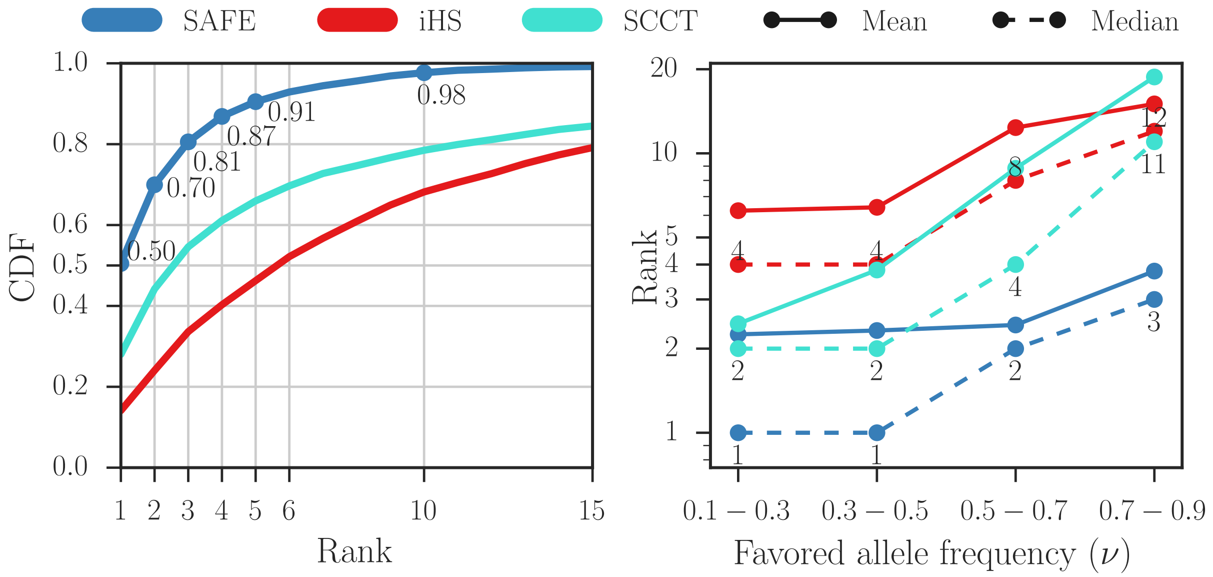


Figure 3.2: **SAFE performance.** Performance (favored mutation rank) of SAFE compared with that of iHS and SCCT on 50 kbp windows with 1,000 simulations per frequency bin and default parameter values (Supplementary Note B.1) for a fixed population size with ongoing selective sweeps. The plot on the 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.

While standing variation, $\nu_0 > 1/N$, generally weakens the selection signal, the performance of SAFE remains relatively robust to variation in ν_0. The median SAFE rank of the favored allele is at most 3 out of ~250 variants in all cases except when $\nu_0 \geq 1000/N$ (Figure C.2).

Similarly, the performance is robust to selection pressure, with only a slight degradation at weak selection ($Ns = 50$) (Figure C.2) where the median rank goes to 9 (3.5%-ile), while for $Ns \geq 200$ the median rank is at most 2. As expected, the performance improves with increasing sample size (Figure C.2). We also tested SAFE on a model of European demography and observed similar results (Figure C.2). These tests used $L = 50$ kbp, chosen so as to minimize the effects of recombination.

Next, we tested SAFE with increasing window sizes, and observed that while the median rank of the favored mutation increases with increasing window size, the percentile rank improves up to 80kbp and then degrades to 3%-ile around 1Mbp (Figure C.2). The deterioration for larger windows is due to most haplotypes becoming 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.

3.4 Summary and discussion

The SAFE score performs very well in identifying the favored variant within a small window (Figure 3.2); but the performance decays in larger windows (Figure C.2). The deterioration for larger windows is due to most haplotypes becoming unique, and κ losing its utility in pinpointing the favored mutation. However, the selective sweep signal is known to extend to large, linked regions, as far as 1Mbp on either side of the favored allele [28]. These ‘shoulders’ of selective sweeps are helpful in identifying the region under selection, but make it harder to pinpoint the favored mutation. To address the more general case of large regions under selection in Chapter 4, we present a method, that is using SAFE score as a building block, to pinpoint the favored mutation even when the signature of selection extends to 5 Mbp, by exploiting the signal from shoulders of the selective sweep.

3.5 Acknowledgments

Chapter 3, in full, contains material from Ali Akbari, Joseph J Vitti, Arya Iranmehr, Mehrdad Bakhtiari, Pardis C Sabeti, Siavash Mirarab, and Vineet Bafna. “Identifying the favored mutation in a positive selective sweep” *Nature methods* (2018) [2]. I was the primary investigator and author of this paper.

Chapter 4

Identifying the favored mutation in a *large* region

Methods to identify signatures of selective sweeps in population genomics data have been actively developed (Figure 1.1, [42–63]), but mostly do not identify the specific mutation favored by the selective sweep. We present iSAFE (integrated Selection of Allele Favored by Evolution), a method that uses a statistic derived solely from population genetics signals to pinpoint the favored mutation even when the signature of selection extends to 5Mbp. iSAFE was tested extensively on simulated data and in human populations from the 1000 Genomes Project (1000GP), at 22 loci with previously characterized selective sweeps. For 14 of the 22 loci, iSAFE ranked the previously characterized candidate mutation among the 13 highest scoring (out of $\sim 21,000$ variants). Three loci did not show a strong signal. For the 5 remaining loci, iSAFE identified previously unreported mutations as being favored. In these regions, all of which involve pigmentation related genes, iSAFE identified identical selected mutations in multiple non-African populations suggesting an out-of-Africa onset of selection.

4.1 Motivation

Genetic data from diverse human populations have revealed a multitude of genomic regions believed to be evolving under positive selection [8, 63]. We consider a regime where a single, *favored*, mutation increases in frequency in response to a selective pressure. The favored mutation either exists as standing variation at the onset of selection pressure, or arises *de novo*, after the onset. Neutral mutations on the same lineage as the favored mutation, hitchhike (are co-inherited) with the favored mutation, and increase in frequency, leading to a loss of genetic diversity.

Methods for detecting genomic regions under selection from population genetic data exploit a variety of genomic signatures. Allele frequency based methods analyze the distortion in the site frequency spectrum [42–46]; Linkage Disequilibrium (LD) based methods use extended homozygosity in haplotypes [47, 48]; population differentiation based methods use difference in allele frequency between populations; and finally, composite methods combine multiple test scores to improve the resolution [56, 57]. Recently, a lack of rare (singleton) mutations has been used to detect very recent selection [58]. The signature of a selective sweep can be captured even when standing variation or multiple *de novo* mutations create a *soft sweep* [24–29] of distinct haplotypes carrying the favored mutation. Together with the advent of deep sequencing, these methods have identified multiple regions believed to be under selection in humans and other organisms, and provide a window into genetic adaptation and evolution [64–73].

In contrast, little work has been done to identify the favored mutation in a selective sweep. Grossman et al. [56] note that different selection signals identify overlapping but different regions, and a composite of multiple signals (CMS) can localize the site of the favored mutation. An alternative strategy is to use functional information to annotate SNPs and rank them in order of their functional relevance. However, the signal of selection is often spread over a large region, up to 1–2 Mbp on either side [28], and the high LD makes it difficult to pinpoint the favored mutation.

Here, we propose a method, iSAFE (integrated Selection of Allele Favored by Evolution), that exploits coalescent based signals in ‘shoulders’ [28] of the selective sweep (genomic regions proximal to the region under selection, but carrying the selection signal) to rank all mutations within a large (5Mb) region based on their contribution to the selection signal. iSAFE does not depend on knowledge of the specific phenotype under selection, and does not rely on functional annotations of mutations, or knowledge of demography.

4.2 Methods

iSAFE: integrated Selection of Allele Favored by Evolution

In Chapter 2 we discussed that the HAF score is a haplotypic score that can be used to separate carrier haplotypes from non-carriers without knowledge of the favored mutation. In Chapter 3, using properties of the HAF score, we developed the SAFE score (Figure 3.1) to identify the favored mutation of an ongoing selective sweep in a small region ($\sim$50 kbp). The SAFE score tends to be maximized for the favored mutation in a small region, but shows decaying performance when larger regions are investigated (Supplementary Figure C.2). To address the more general case of large regions ($\sim$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.

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

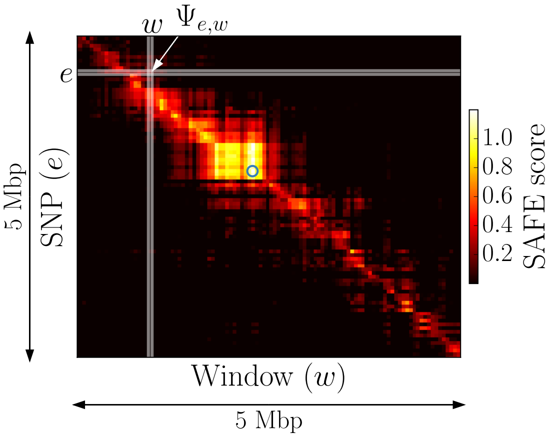


Figure 4.1: **The Ψ matrix.** The $\Psi_{e,w}$ matrix for a 5-Mbp region around the LCT gene in the 1000GP FIN population shows that the 'shoulder' of selection can extend for a few megabase pairs. The blue circle indicates the location of putative favored mutation rs4988235.

corresponding to the favored mutation.

For larger regions, we considered a set of 50% overlapping windows of fixed size (300 SNPs). Define $\mathcal{S}$ as the set of all SNPs, and $\mathcal{W}$ as the set of all sliding windows. For each window, we applied SAFE and chose the mutation with the highest SAFE score. Let $\mathcal{S}_1 \subseteq \mathcal{S}$ denote the set of selected mutations. Mutations in $\mathcal{S}_1$ are likely to contain either the favored mutation itself or mutations linked to it. For mutation $e \in \mathcal{S}$, and window $w \in \mathcal{W}$, let $\Psi_{e,w'}$ denote either the SAFE score of e when e is inserted into window $w' \in \mathcal{W}$, or 0, whichever is larger (Figure 4.1). As different windows have different genealogies due to recombination, $\Psi_{e,w'}$ is relatively high when e is the favored mutation and the genealogies of w, w' are identical or very similar, but not otherwise. In contrast, the SAFE score of a non-favored mutation e is relatively low when inserted in other windows. Define the weight of a window $w \in \mathcal{W}$ as

$$\alpha(w) = \frac{\sum_{e \in \mathcal{S}_1} \Psi_{e,w}}{\sum_{w' \in \mathcal{W}} \sum_{e \in \mathcal{S}_1} \Psi_{e,w'}}. \tag{4.1}$$

Windows that contain the favored mutation and those sharing its genealogy are expected to have

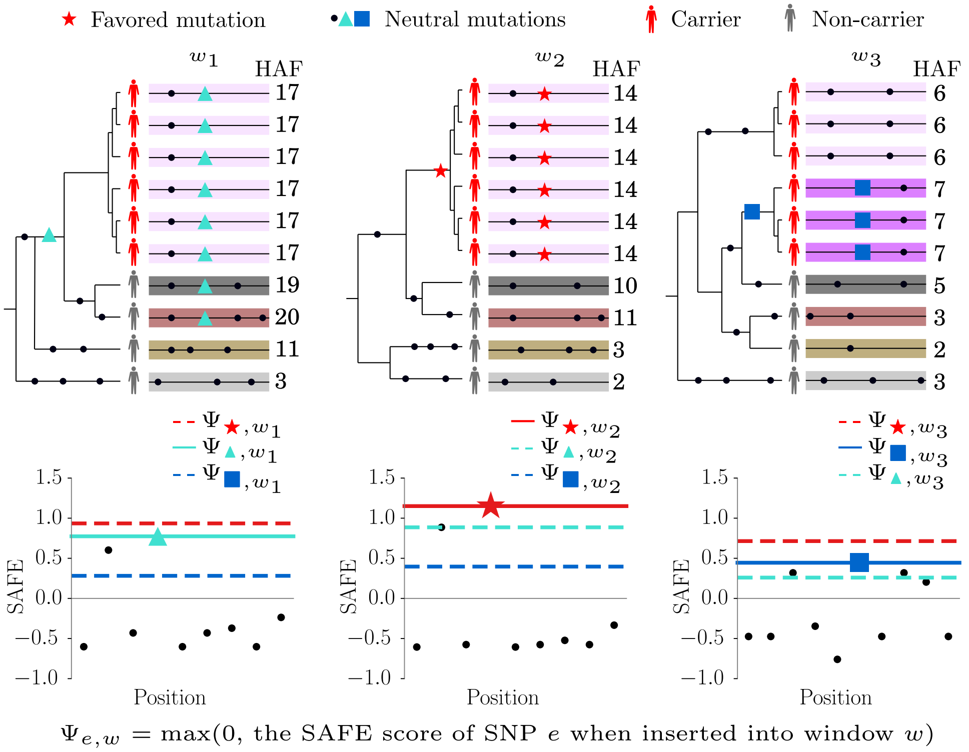


Figure 4.2: **Illustration of the iSAFE method.** Different genomic windows (w) have different genealogies because of recombination. The SAFE score of a non-favored mutation e is relatively low when it is inserted in other windows. In contrast, the SAFE score of the favored mutation is likely to be dominant over those of other mutations. Identical haplotypes in each window are colored similarly.

high α values. We define the iSAFE score of a mutation $e \in \mathcal{S}$ as:

$$\text{iSAFE}(e) = \sum_{w \in \mathcal{W}} \Psi_{e,w} \cdot \alpha(w). \tag{4.2}$$

A cartoon illustration of the iSAFE method

Figure 4.2 provides a cartoon illustration of the iSAFE method. In this simplified toy example, $\mathcal{W} = \{w_1, w_2, w_3\}$ and $\mathcal{S}_1 = \{▲, ★, ■\}$, where ★ denotes the favored mutation and is located in w_2. We note the following:

- $\Psi_{★,w_2}$ is high for the favored mutation ★. However, $\Psi_{▲,w_1}$ and $\Psi_{■,w_3}$ may be high even for hitchhiking mutations (▲,■) 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.

MDDAF: Maximum Difference in Derived Allele Frequency

Not surprisingly, iSAFE performance deteriorates when the favored mutation is fixed, or near fixation ($\nu > 0.9$ in Supplementary Figure C.3**e**). To handle this special case, we include individuals from non-target populations, using a specific protocol (see Supplementary Note B.3). For a mutation, define the Maximum Difference in Derived Allele Frequency score (MDDAF) as

$$\text{MDDAF} = D_T - \min(D_{NT}) \, , \tag{4.3}$$

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 (Supplementary Figure B.1), shows $P(\text{MDDAF} > 0.78 | D_T > 0.9) = 0.001$ (see Supplementary Figure C.4). Therefore, when we observe the rare event of high frequency mutations in target ($D_T > 0.9$) with MDDAF > 0.78, we add random outgroup samples to the data to constitute 10% of the data (see Supplementary Note B.3).

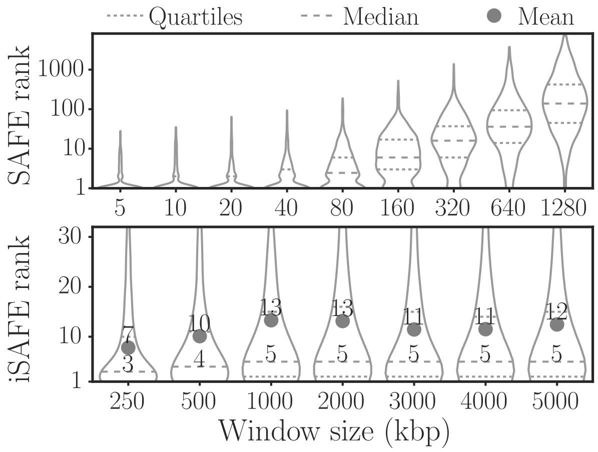


Figure 4.3: **SAFE versus iSAFE.** SAFE and iSAFE performance (rank distribution of favored mutation) as a function of window size with 1,000 simulations per bin. Median and quartile values decay with increasing window size in SAFE, whereas iSAFE is robust to increases in window size.

4.3 Results

Simulations

iSAFE, unlike SAFE, is specifically designed to exploit signal from the shoulders of the sweep (Figure 4.2). iSAFE showed consistently high performance as the window size was increased from 250 kbp to 5 Mbp (Figure 4.3). The median rank remained between 3 and 5 for windows up to 5 Mbp in size, and the performance remained robust to a large range of parameters (Supplementary Figs. C.2, C.3). iSAFE showed greatly improved performance compared with that of iHS and SCCT, placing the favored mutation within the top 20 in 88% of cases, in contrast to iHS (39%) and SCCT (34%), for an ongoing selective sweep with fixed population size (Supplementary Figure C.3**d**).

iSAFE scores are not based upon likelihood computations, and the distribution of scores depends upon 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

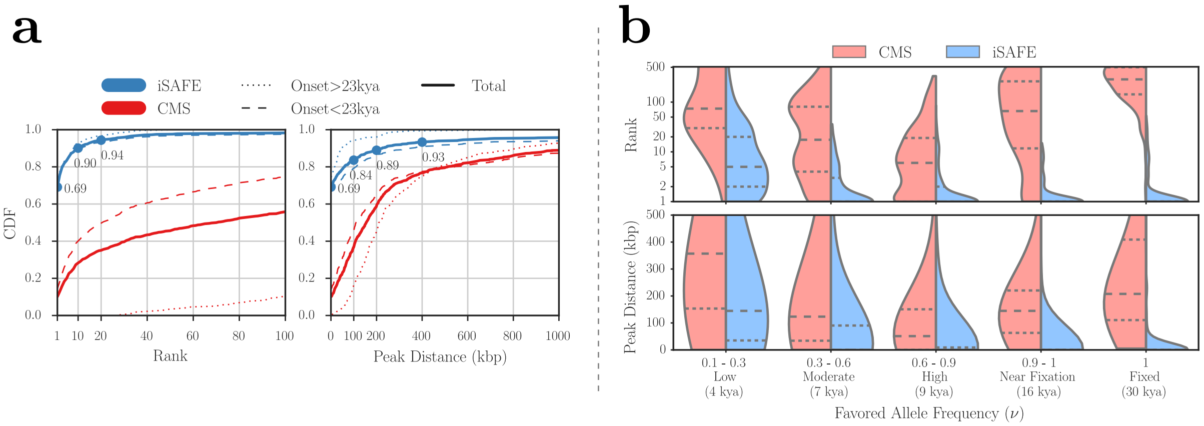


Figure 4.4: **iSAFE performance.** (**a**) The cumulative distribution function (CDF) of the favored mutation rank (left) and peak distance (right) for iSAFE and CMS scores. (**b**) Rank and peak distance distributions of the favored mutation as a function of favored allele frequency (ν) in the target population (EUR). In the (**b**), the dashed (dotted) line represents the median (quartiles). All data are based on 1,000 simulations of 5-Mbp genomic regions using a model of human demography (Supplementary Note B.2). The time of onset of selection was chosen at random from the distribution in Supplementary Figure B.1, after the out-of-Africa event, in the EUR (target) population lineage. When the onset of selection is before the EUR-EAS split (>23 kya), both EUR and EAS are under selection. kya, kiloyears ago.

the favored mutation is 4 when peak iSAFE score exceeds 0.1 versus a median rank of 10 along with a longer tail, when peak iSAFE score is below 0.1 (Supplementary Figure C.5). Empirically computed *P* values (see Supplementary Note B.4) on iSAFE indicate good performance when *P* value $<$ 1e-4 (Supplementary Figure C.5).

Not surprisingly, iSAFE performance deteriorates when the favored mutation is fixed, or near fixation ($\nu > 0.9$ in Supplementary Figure C.3**e**). To handle this special case, we include individuals from non-target populations, using a specific protocol (see Supplementary Note B.3). 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 Figure C.3**e**). 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 (Figure 4.4 and Supplementary Figure C.6), in contrast to CMS, which gave a top-20 ranking in 35% of cases.

1000 Genomes Project data

In testing instances of previously characterized sweeps in 1000GP data, we note that performance is difficult to characterize due to 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 also 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 sub-populations, 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 (Table 4.1).

We used iSAFE to rank all variants ($\sim$21,000) in a 5Mbp region surrounding the gene. Among the 8 well characterized cases, (Table 4.1 and Figure 4.5), iSAFE ranked the candidate mutations as 1 in five cases (SLC24A5, EDAR, LCT, TLR1, ACKR1) and ranked the remaining cases as 2 (ABCC1), 4 (HBB), and 13 (G6PD).

We checked whether the other 14 loci [56, 79–82] under selection showed strong iSAFE signals (Supplementary Note A). We checked to see if the other 14 regions under selection showed a strong iSAFE signal. In 3 of the 14 regions (FUT2, F12, ASPM; Supplementary Figure A.1), we only observed weak signals, and did not make a prediction (peak iSAFE < 0.027). In other loci, iSAFE ranked the candidate mutations as 1 in the SLC45A2/MATP (CEU population), MC1R (CHB and JPT populations), and ATXN2/SHB3 (GBR population) loci (Figure 4.6), and as 7, 8, and 12 in the PSCA (YRI population), ADH1B (CHB and JPT populations), and PCDH15 (CHB and JPT populations) loci, respectively. In each case, the iSAFE scores were high with the exception of PSCA (peak iSAFE $= 0.04$, Supplementary Figure A.1).

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 (Figure 4.6 and Supplementary Note A). 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 [79]. A second intronic variant, rs10831496, in GRM5, 396kbp upstream of TYR, has been shown to have a strong association with skin color [83]. In contrast, iSAFE ranks mutation rs672144 at the top not only in the CEU population sample (iSAFE $= 0.48$, *P* value$\ll$1.3e-8), but also in EUR, EAS, AMR, and SAS (iSAFE $>$0.5, *P* value$\ll$1.3e-8; Figs. 4.6, 4.8), 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 (Figure 4.8). We plotted the haplotypes carrying rs672144 and found that two distinct haplotypes carry the mutation, both remaining high frequency, maintained across a large stretch of the region, suggestive of a soft sweep with standing variation (Fig 4.7). A similar analysis applied to genes TRPV6, KITLG, OCA-HERC2, where in each case, the top iSAFE mutations were identical across all non-African populations (Supplementary Note A.2), and supported an out-of-Africa onset of selection. In the one remaining gene (CYP1A2/CSK; Figure 4.6; Supplementary Note A.4), the top ranked iSAFE mutation rs2470893 was previously found significant in a genome wide association study [84], and was tightly linked to the candidate mutation. To summarize, iSAFE analysis ranked the candidate mutation among the top 13 in 14 of the 22 loci, did not show a strong signal in 3, and identified plausible alternatives in the remaining 5. See Supplementary Note A for more detailed analysis of these 22 loci.

Table 4.1: **iSAFE on 8 well characterized selective sweeps.** We used iSAFE to rank all variants (∼21,000) in a 5Mbp region surrounding the gene. Among the 8 well characterized cases, (Figure 4.5), iSAFE ranked the candidate mutations as 1 in five cases (SLC24A5, EDAR, LCT, TLR1, ACKR1) and ranked the remaining cases as 2 (ABCC1), 4 (HBB), and 13 (G6PD). 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 Supplementary Note B.4.

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	[56]	[85]
EDAR	CHB+JPT	rs3827760	Missense	0.87	Hair and teeth	1	<1.3e-8	[56]	[86, 87]
LCT/MCM6	FIN	rs4988235	Intron	0.59	Lactase persistence	1	<1.3e-8	[32]	[88, 89]
TLR1	CEU	rs5743618	Missense	0.77	Sepsis, leprosy, tuberculosis	1	1.0e-5	[90]	[91]
ACKR1/DARC	YRI	rs2814778	5′UTR	1	Malaria resistance	1	2.8e-5	[92]	[93]
ABCC11	CHB+JPT	rs17822931	Missense	0.93	Cold climate, earwax, body odour	2	<1.3e-8	[94]	[94]
HBB	YRI	rs334	Missense	0.14	Malaria resistance	4	1.6e-4	[95]	[96]
G6PD	YRI	rs1050828	Missense	0.21	Malaria resistance	13	7.3e-6	[32]	[97]

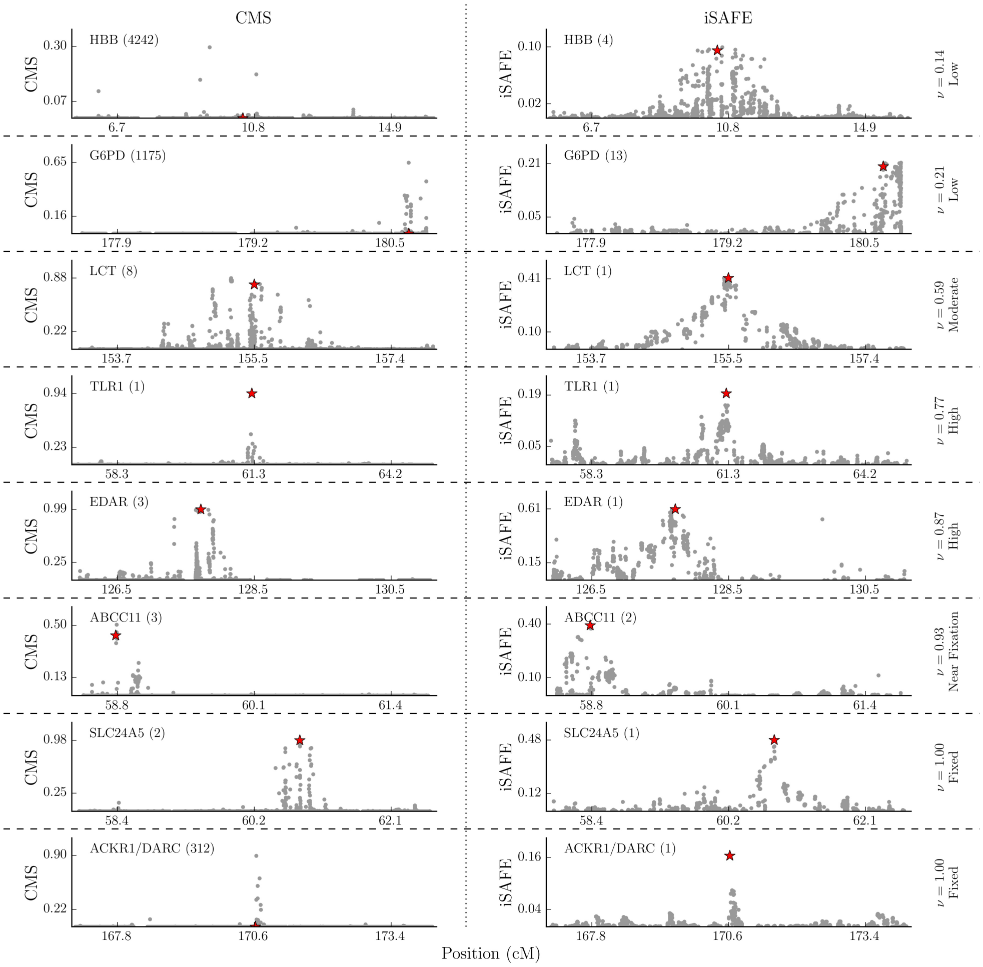


Figure 4.5: **iSAFE and CMS on 8 well-characterized selective sweeps.** We used iSAFE to rank all variants (∼21,000) in a 5Mbp region surrounding the gene. Among the 8 well characterized cases, (Table 4.1), iSAFE ranked the candidate mutations as 1 in five cases (SLC24A5, EDAR, LCT, TLR1, ACKR1) and ranked the remaining cases as 2 (ABCC1), 4 (HBB), and 13 (G6PD). The rank of the putative favored mutation in the 5-Mbp region is shown in the top left corner in each plot. cM, centimorgan.

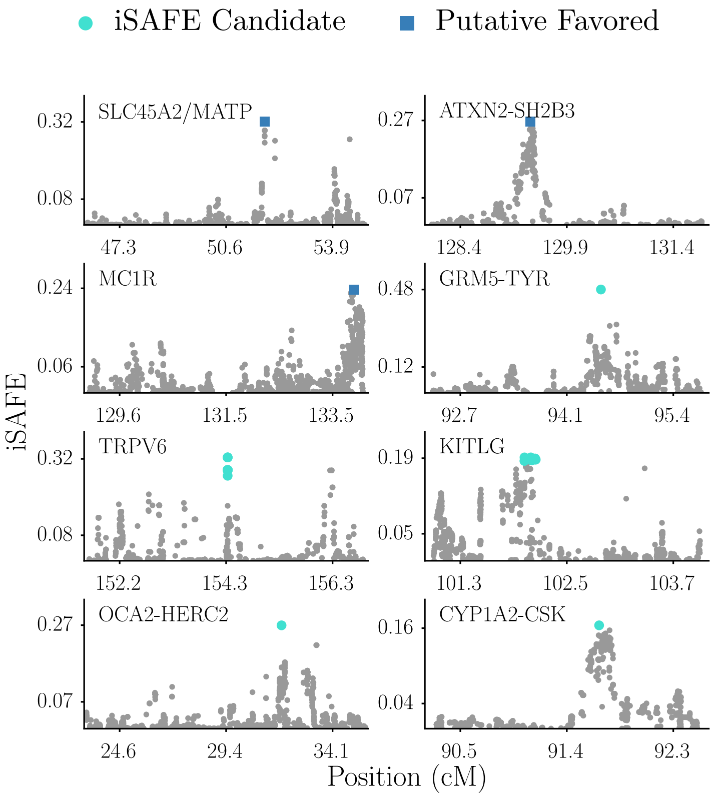


Figure 4.6: **iSAFE scores for regions under selection.** Top-ranked iSAFE candidates that match reported favored mutations (putative favored) or are newly suggested by iSAFE (iSAFE candidate) are indicated. All datasets consisted of a 5-Mbp window around the selected region, unless one side reached the telomere or centromere.

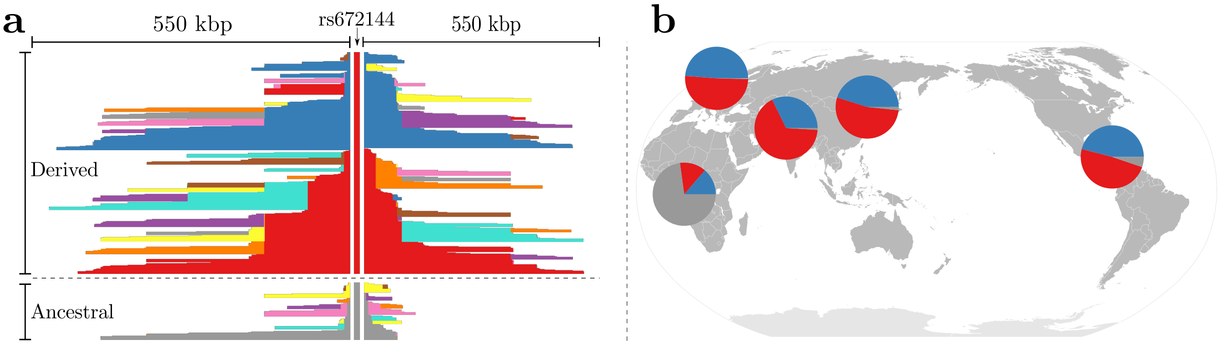


Figure 4.7: **The GRM5-TYR region.** The mutation rs672144 was ranked first by iSAFE and is very well separated from other mutations in the surrounding 5 Mbp, in all non-African populations, with high confidence (iSAFE score > 0.5, P $\ll$ 1.3e-8 ; Figure 4.8). (**a**) All 5,008 haplotypes (2,504 samples) from 1000GP carrying core mutation rs672144 (red/blue) are conserved over a longer distance than haplotypes in non-carriers (gray), which is a signal of selection [47]. (**b**) Global frequencies of haplotypes carrying (red/blue) and not carrying (gray) mutation rs672144 are consistent with out-of-Africa selection on standing variation (soft sweep), with mutation rs672144 as the favored variant.

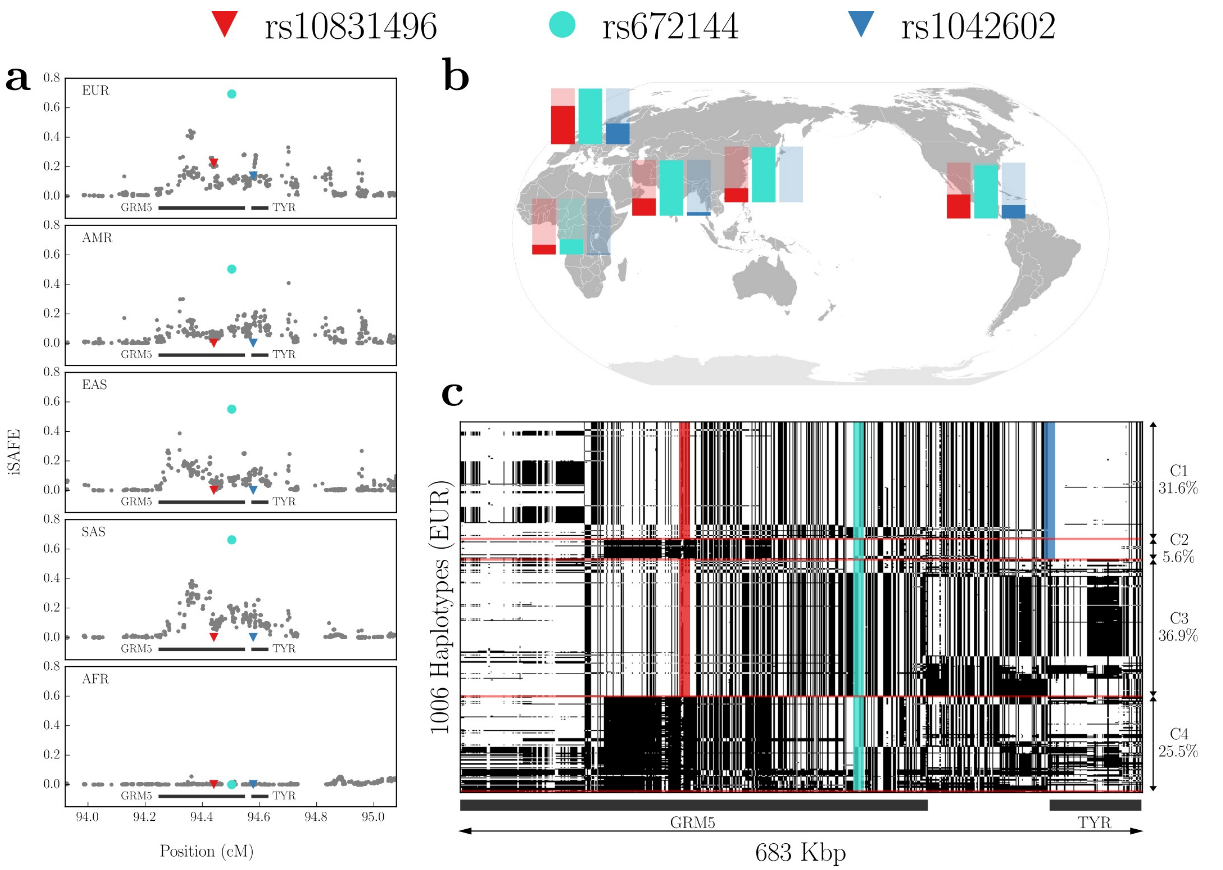


Figure 4.8: **iSAFE on the GRM5-TYR locus.** The Tyrosinase (TYR) gene, encoding an enzyme involved in the first step of melanin production is present in a large region under selection. A nonsynonymous mutation rs1042602 (blue) in TYR gene is reported as a candidate favored variant. A second intronic variant rs10831496 (red) in GRM5 gene, 396 kbp upstream of TYR, has been shown to have a strong association with skin color. In contrast, iSAFE ranks mutation rs672144 (turquoise) as the top candidate for the favored variant region out of 22,000 mutations (5 Mbp; see Section A.2). (**a**) This variant was the top ranked mutation not only in CEU (Fig 4.6), but also the top ranked mutation for EUR, EAS, AMR, and SAS. The signal of selection is strong in all populations (iSAFE $>$ 0.5, P $\ll$ 1.3e-8 for all of) except AFR, which does not show a signal of selection in this region. We plotted the haplotypes carrying rs672144 in all 5008 haplotypes (2504 samples) of 1000GP and found (Fig 4.7) 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. (**b**) Frequency of derived alleles of rs10831496, rs672144, and rs1042602, are shown in red, turquoise, and blue, respectively. iSAFE candidate (rs672144) may not have been reported earlier because it is near fixation in all populations of 1000GP except for AFR ($f = 0.27$). (c) Each row is a haplotype and each column is a variant in EUR populations of 1000GP. In total we have 1006 haplotypes (503 samples). Carrier haplotypes of derived alleles of rs10831496, rs672144, and rs1042602, are shaded by red, turquoise, and blue, respectively. For making the plot sensible, we removed low frequency SNPs $f_{EUR} < 0.2$ and SNPs that are near fixation in the whole 1000GP, $f_{1000GP} > 0.95$. The previously suggested candidates rs1042602, rs10831496 are fully linked to rs672144, but not to each other. The EUR haplotypes can be partitioned into 4 clusters. 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.

4.4 Summary and discussion

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.

An important challenge was that regions undergoing a selective sweep also present a signal far away from the favored mutation, making it harder to pinpoint the favored mutation. We observe that when a true favored mutation is inserted into a shoulder region, it gets higher SAFE scores on average, in contrast to the insertion of a hitchhiking mutation. The iSAFE technique uses this idea to exploit the shoulders and rank mutations according to the weighted sum of their SAFE scores in all windows.

We also use a cross-population technique in a limited manner by using the frequency differential of mutations in high frequency scenarios to get representative non-carrier haplotypes in the sample, and show its power in identifying nearly fixed favored mutations. We do assume a model with a single, favored variant, and future work could contribute to identify multiple interacting loci favored by selection. Finally, in Chapter 5, we use the iSAFE score as the main feature of a supervised learning approach to identify adaptively introgressed haplotypes in human populations without having knowledge of the archaic samples.

4.5 Acknowledgments

Chapter 4, in full, contains material from Ali Akbari, Joseph J Vitti, Arya Iranmehr, Mehrdad Bakhtiari, Pardis C Sabeti, Siavash Mirarab, and Vineet Bafna. “Identifying the favored

mutation in a positive selective sweep" *Nature methods* (2018) [2]. I was the primary investigator and author of this paper.

Chapter 5

Capturing haplotypes adaptively introgressed

We present CHAI (Capturing Haplotypes Adaptively Introgressed), a method that enables researchers to accurately identify the adaptively introgressed haplotypes without knowledge of the archaic samples. The CHAI method is an extension of iSAFE, a novel method we proposed in Chapter 4 to accurately pinpoint the favored mutation in a large region ($\sim$ 5 Mbp) by using a statistic derived solely from population genetics signals. Results on both simulations and real data are very promising.

5.1 Motivation

There is a growing body of evidence showing that *introgression* – the transfer of genetic information from one species to another as a result of interbreeding – happened from archaic hominins like Neanderthals and Denisovans, into modern humans. Approximately 2% of the genome of non-African populations are introgressed from archaic hominins like Neandertals [6]. A striking aspect of introgression is it can provide an evolutionary shortcut for adaptation to changing selection pressures. The most well known example of adaptive introgression (AI) is at the *EPAS1* locus in Tibetan highlanders, where the favored haplotype was introgressed from Denisovan-like archaic hominins, and carriers of the introgressed haplotype adapted better to the hypoxic environment at high altitudes [14, 98].

These discoveries have spurred the sequencing of archaic hominin genomes [34–37], along with the development of methods that detect introgression [6, 37, 99] by comparing the reference human sequence to the genomes of the archaic hominins. An introgressed haplotype at high frequency, relative to other populations, can be a signal of adaptive introgression. Introgression increases LD and also distorts the pattern of allele frequency distribution, which both are used by statistics such as integrated haplotype score (iHS) [47] and Tajima's D [42] to detect region under selection. Therefore, using standard tests of selection on region with introgression can easily lead to false inference of selection [39]. Racimo et al. [41], consider the joint dynamic of selection and introgression and also incorporate the reference archaic hominin genome sequences to detect adaptive introgression.

However, several recent studies indicate introgression in African populations from unknown archaic hominins [100–103]. The number of archaic honinin species remains unknown. It is feasible that introgression from unknown archaic homins has also helped humans adapt as they migrated out of Africa. Even in the absence of a reference hominin genome, methods like S^* [104, 105] and Sprime [106] have been developed that use linkage disequilibrium (LD) patterns to detect introgressed haplotypes. However, there is no current test for AI without knowledge of archaic reference.

Here, we present CHAI, a method to capture adaptively introgressed haplotypes without knowledge of archaic samples. CHAI uses a supervised-learning approach to score each mutation according to its probability of being adaptively introgressed. It is designed to minimize false inference due to confounding by other events, including hard/soft sweeps and adaptation on regions with an active recombination suppressor mechanism (RSM), for example, due to a

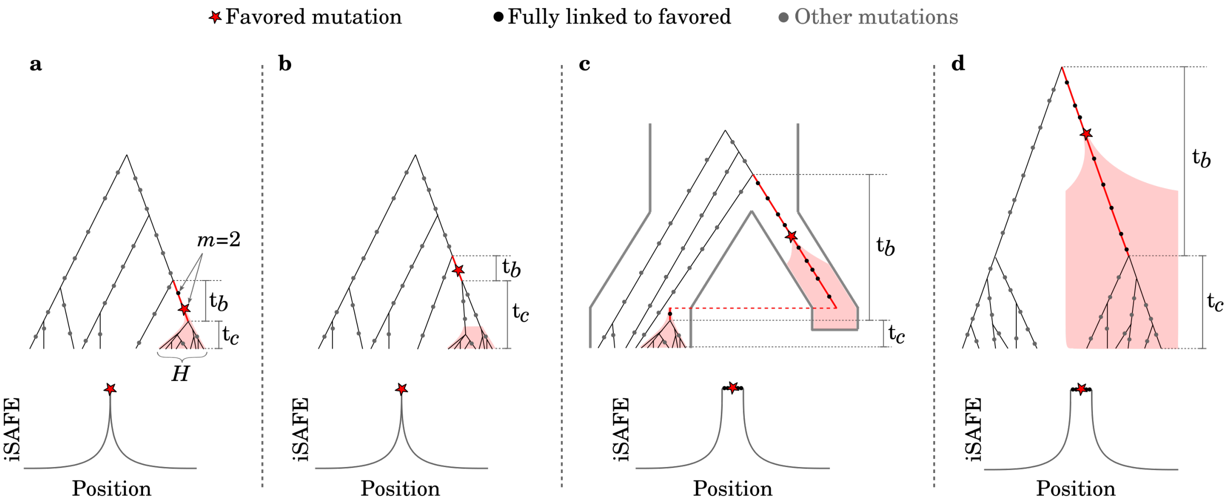


Figure 5.1: **Pattern of iSAFE in different evolutionary scenarios.** Evolutionary scenarios explaining mutations fully-linked to the favored mutation in a selective sweep. **(a)** Hard sweep. **(b)** Soft sweep on standing variation. **(c)** Adaptive introgression (AI). **(d)** Recombination suppressor mechanisms (RSM) like chromosomal inversion with balancing selection. t_b is the length of the branch where the mutation arises and t_c is the coalescent time of carriers of the mutation. Red-shaded area showing the spread of the favored mutation after the onset of selection pressure. The lower graphs are showing the pattern of iSAFE signal on a large window (∼5 Mbp) around the favored mutation and trees are showing a simplified genealogy of different scenarios on a small window (∼50 kbp) around the favored mutation.

chromosomal inversion event.

Given a population sample of n individuals, and a candidate mutation, CHAI is inferred using only 3 features. First, and most important, it uses the iSAFE [2] ("integrated selection of allele favored by evolution") score for identifying the favored mutation in a selective sweep. Second, it uses t_b, the length of the branch where the mutation arises. Third, it uses t_c, the time to coalescence of the subset of haplotypes carrying the mutation. For each SNP, the CHAI score is defined as the probability estimate of a Logistic Regression classifier with a quadratic decision boundary [107, 108], and iSAFE, t_a, and t_b as features.

5.2 Methods

We showed, in Chapter 4 with an extensive analysis, that the iSAFE score gives a sharp peak, usually with the favored mutation on top, in a region under hard/soft selective sweep. iSAFE treats each SNP as a binary classifier, carriers and non-carriers. If two SNPs are fully-linked together (in complete linkage disequilibrium) they get the exact same iSAFE score regardless of their position in the window. This unique feature of the iSAFE score helps to identify the haplotype under selection in scenarios more complex than a hard/soft selective sweep, such as adaptive introgression. The pattern of iSAFE signal in a region undergoing a recent adaptive introgression is flattened as the branch length of the favored mutation is longer (higher t_b), and all the mutations on this branch get the exact same iSAFE score as the favored mutation (see Figure 5.1). We utilize this pattern to distinguish recent adaptive introgression from a hard/soft selective sweep.

Confounding factors. As described, we look for long branch lengths t_b – the length of the branch where the mutation arises (see Figure 5.1). Such long branch length could also arise due to recombination suppressor mechanisms (RSM) like chromosomal inversion with balancing selection (Figure 5.1d). However, this is only true (high t_b) if the inversion happened a long time ago. By incorporating the t_c – an estimate of the time to coalescence of the subset of haplotypes carrying the mutation – we are able to rule out these cases as t_c is high for old adaptation and low for a recent adaptation.

Estimating branch length. Let H denote a subset of haplotypes carrying a mutation in a sample of size n, and m denote number of mutations fully-linked to it (the number of mutations shared by all, and only the haplotypes in H). Let w be the minimal region demarcated by these m fully-linked mutations and assume within this, usually small, region is recombination free. Define t_b as length of the branch where the mutation arises (Figure 5.1). We assume mutations occur as a Poisson process along all branches [109] (constant molecular clock). Therefore, we estimate t_b by:

$$t_b \approx \frac{m}{\mu l}, \tag{5.1}$$

where l is the length of region w in bp and μ is mutation rate per bp per generation.

Estimating coalescent time. Let t_c denote the coalescent time – the time to the most recent common ancestor (MRCA) – of the haplotypes in H (Figure 5.1a). Then an estimator [109, 110] of t_c is

$$t_c = \frac{\sum_{h \in H} x_h}{\mu l \cdot |H|}, \tag{5.2}$$

where x_h is number of mutations on haplotype h that are polymorphic in H, and $|H|$ is number of haplotypes in H (cardinality of set H).

CHAI: Capturing Haplotypes Adaptively Introgressed

We usually do not know the favored mutation. Therefore, we calculate iSAFE, t_a, and t_c scores for all mutations in the target region. As mentioned before, iSAFE treats each SNP as a binary classifier, carriers and non-carriers, and if two SNPs are fully-linked together they get the exact same iSAFE score regardless of their position in the window. Similar to iSAFE, if two SNPs are fully-linked together they get the exact same t_b and t_c scores regardless of their position in the target window. Because two fully-linked SNPs are on the same branch and consequently their branch length t_b is identical; also, their carrier haplotypes are the same, which leads to identical carriers-coalescent-time t_c. This unique feature helps to identify the haplotype under selection in scenarios more complex than a hard/soft selective sweep, like adaptive introgression and adaptation on a recombination suppressor mechanisms like chromosomal inversion.

The combination of these three scores (iSAFE, t_b, and t_c) provides enough information to identify adaptively introgressed haplotypes without having archaic samples. iSAFE tends to be maximized for the favored mutation. In an adaptive introgression, the branch length score t_b of the favored mutation is expected to be large compared to hard/soft selective sweeps. Such long branch length could also arise due to recombination suppressor mechanisms like chromosomal inversion with balancing selection, if happened a long time ago. Therefore, we are able to rule out these cases by incorporating the carriers-coalescent-time t_c.

In a recent adaptive introgression, we expect the favored mutation, and all other fully-linked mutations that represent the favored haplotype, to have significantly high iSAFE (to capture the favored haplotypes and rule out neutrals), high t_b (to rule out hard/soft sweeps), and low t_c (to rule out recombination suppressor mechanism). We use a supervised learning approach to combine information of these three scores and devise the CHAI score to capture adaptively introgressed haplotypes. For each SNP, the CHAI score is defined as the probability estimate of a Logistic Regression classifier with a quadratic decision boundary [107, 108], and iSAFE, t_a, and t_b as features.

5.3 Results

Simulation. CHAI is a reference-free method for detecting adaptive introgression. An introgressed haplotype at high frequency, relative to other populations, can be a signal of adaptive introgression. The Sprime method is proposed by Browning et al. 2018 [106] and predicts introgressed alleles without a reference archaic sequence. Sprime is similar to S^* [104, 105], with a superior performance. As an evaluation for our method (CHAI) we designed the following experiment. We simulated 5 different scenarios including, adaptive introgression (AI), adaptation on recombination suppressor mechanisms (RSM), hard sweep, soft sweep, and neutral. As Sprime is not specifically designed to capture adaptive introgression we assume that we know the region is under selection and simply use Sprime to capture introgression. Therefore, if Sprime maximum value in the 1 Mbp region around the selected locus is above a specific threshold (5% FDR in neutral scenario) we label the region as adaptive introgression. For each scenario we simulated 1000 replicates. Figure 5.2 shows CHAI is very powerful in detecting the AI and distinguishing it from other evolutionary scenarios, specially RSM. CHAI detected AI in 94% of the cases compared to 30% for Sprime with default parameters ($maxfreq = 0.01$). The rate of false inference of RSM for CHAI was 3% compared to 23% for Sprime with default parameters.

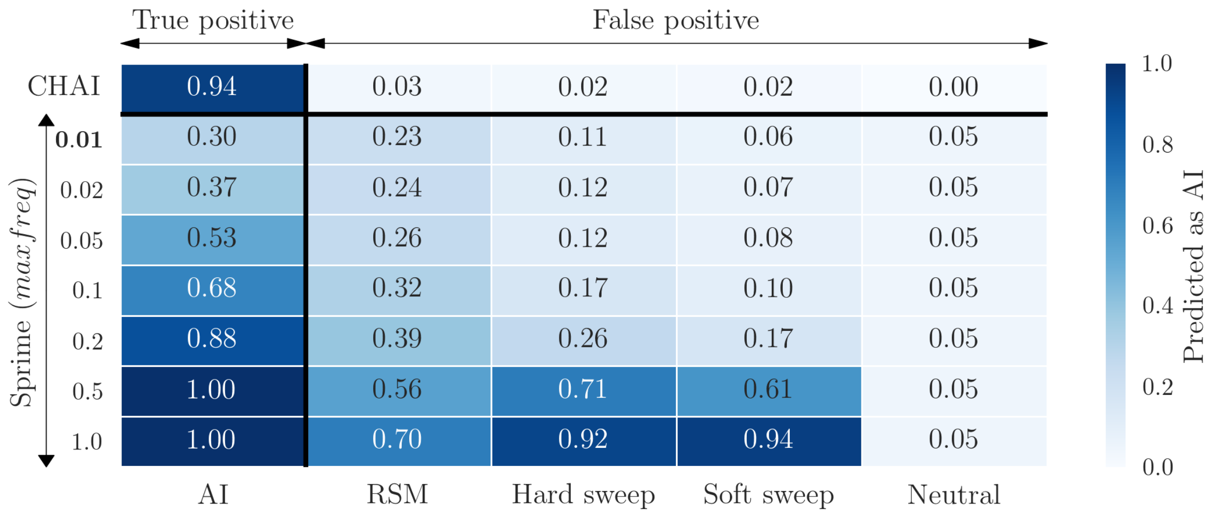


Figure 5.2: **CHAI performance**. We simulated 5 different scenarios including, adaptive introgression (AI), adaptation on recombination suppressor mechanism (RSM), hard sweep, soft sweep, and neutral. As Sprime is not specifically designed to capture adaptive introgression we assume that we know the region is under selection and simply use Sprime to capture introgression. Therefore, if Sprime maximum value in the 1 Mbp region around the selected locus is above a specific threshold (5% FDR in neutral scenario) we label the region as adaptive introgression. For each scenario we simulated 1000 replicates. As you can see, CHAI is very powerful in detecting AI and minimizing the false inference of other evolutionary scenarios, specially RSM. We tested Sprime for different $maxfreq \in \{0.01,\ 0.02,\ 0.05,\ 0.1,\ 0.2,\ 0.5,\ 1.0\}$, a parameter of Sprime method that specifies the maximum frequency of an introgressed variant in the outgroup with default value 0.01 (boldfaced).

***EPAS1* locus.** CHAI performance on simulated data is very promising (Figure 5.2). As an evaluation with real data (see Section B.5), we applied CHAI on *EPAS1* locus in Tibetans highlanders (TIB) [14]. We used 38 whole genome sequences of TIB provided by Lu et. al. [111]. Figures 5.3a,b show iSAFE and CHAI scores for all SNPs in 3 Mbp region around *EPAS1* locus in Tibetan highlanders. In this region, 173 SNPs (spanning ~300 kbp region chr2:46567916-46870805) have CHAI $>$ 0.98 and predicted to be adaptively introgressed. These SNPs include the five SNPs (turquoise color in Figures 5.3a-c) used in Huerta-Sánchez et al. 2014 [14], that led them to conclude this segment of genome is adaptively introgressed from an archaic hominin like Denisovans. Figure 5.3c shows the frequency of these SNPs in TIB population and also in EAS, EUR, and AFR super-populations of the 1000GP (left panel), and the state of the derived alleles in the Denisova and Altai, Vindija33.19, and Mez1 Neanderthals (right panel). The median frequency of these 173 SNPs is 0% for EUR and AFR, 2% for EAS, and 76% for TIB. The values on top of right panels in Figures 5.3c,d are the proportion of the derived alleles (excluding missed ones) in the corresponding archaic sample which can be considered as a measure of similarity. These 173 SNPs are 47% match to Denisova compared to 20%, 19%, and 18% match to Altai, Vindija33.19, and Mez1 Neanderthals, respectively. This observation is consistent with the claim that this haplotype is adaptively introgressed from an archaic hominin similar to Denisovans.

The iSAFE top candidate haplotype (set of fully-linked SNPs) have 41 fully-linked SNPs around *TMEM247* (chr2:46657114-46730100), 45 kbp down stream of the *EPAS1*. One of these 41 SNPs is missense (rs116983452; orange color in Figure 5.3). Figure 5.3d shows the frequency of these 41 SNPs in super-populations of the 1000GP (left panel), and the state of the derived alleles in the archaic samples (right panel). For this missense SNP, Denisova has the derived allele and all three Neanderthals have ancestral allele. The derived allele frequency of this SNP is 76% (TIB), 2% (EAS), and 0% (EUR and AFR). This observation suggests the missense SNP rs116983452 is a potential candidate for being the favored mutation driving this adaptive introgression and it is likely introgressed from an archaic hominin similar to Denisovans.

5.4 The story of *LCT* locus

The selective sweep in *LCT* locus in European populations is probably the most famous example of positive selective sweep. The signature of the selection in this region is very strong and extends to a few Mbp on each side of the sweep (See Figure 4.1). Almost all of the selection tests (see Figure 1.1) pick this region in European with a high confidence. The putative favored mutation rs4988235 is top iSAFE candidate in FIN population, joint with five other fully-linked SNPs rs182549, rs12465802, rs56369224, rs6730157, and rs7570971. These six SNPs span 779 kbp region chr2:135837906-136616754 surrounding the *LCT* gene. This SNPs represent the famous favored haplotype associated with the lactose persistence in European populations. CHAI scores for these SNPs is zero that suggests the selective sweep driven by these putative favored haplotype is not adaptively introgressed. Frequency of these SNPs are 59% in FIN and 0% in CHB populations. This putative favored haplotype in FIN population does not exist in CHB population and not surprisingly its corresponding iSAFE scores are zero in CHB (see Figure 5.5a). However, the iSAFE score in *LCT* region in CHB population shows signal of selection with a high confidence (see Figure 5.5b). Our analysis suggests *LCT* locus have been target of multiple selection events in human population. Some of these candidate favored haplotypes are adaptively introgressed. We elaborate on this below.

Distinguish multiple selection events at the same locus

iSAFE treat each SNP as a binary classifier, carriers and non-carriers of the derived allele, and it represent how well the SNP can explain the pattern of haplotype homozygosity in its neighboring region. Carriers/non-carriers of the favored mutation are supposed to have high/low haplotype homozygosity, respectively. In Figure 5.4 we demonstrate that we cannot distinguish multiple selective sweeps in the same region when their genomic distance is very close (a few hundred kbp), just by looking at the iSAFE signal as a function of their genomic position. However, by looking at iSAFE signals in frequency domain we might be able to detect and distinguish multiple selective sweeps.

For each SNP, when the frequency distance to the favored mutation increases (LD to the favored mutation decreases), we expect a predictable decay in its iSAFE score (see the right panels in the upper part of the Figure 5.4). Therefore, hitchhikers of the favored mutation are expected to have a predictable distribution of the iSAFE score as a function of frequency distance to the favored mutation (gray shade in the right panels of the Figure 5.4). Sometimes, this pattern is violated when there are more than a single selective sweep in the same region and we utilize this pattern to distinguish different sweeps (see the right panels in the lower part of the Figure 5.4).

Multiple selection events at *LCT* locus

The putative favored haplotype in FIN population (represented by rs4988235) does not exist in CHB population and not surprisingly its corresponding iSAFE scores are zero in CHB. However, the iSAFE score in *LCT* region in CHB population shows signal of selection with a high confidence (Figures 5.5, 5.6). In Figure 5.4 we showed when two selections occur near each other sometimes their signals can be distinguished by their iSAFE scores in the frequency domain even when their signals as a function of chromosomal position are not separable. In Figure 5.5, we looked at the pattern of iSAFE signal in FIN, CHB, JPT, and combined FIN+CHB+JPT samples of 1000GP. The right panels of Figure 5.5 (iSAFE score in frequency domain), along with Figure 5.4, imply that pattern of signals in *LCT* locus is consistent with the pattern of multiple selections. The coloring of these SNPs are based on the significance of their iSAFE score (P value $<$ 1e-4) in FIN, CHB, and JPT populations. Following we summarize our observation and predictions based on analyzing the *LCT* locus in 1000GP population using our methods.

The blue (FIN only). The blue sweep, in Figure 5.5, is the well-characterized selective sweep associated with lactose persistence in European population and our analysis suggests it is not adaptively introgressed. The blue SNPs shows signal of selection in FIN population and combined FIN+CHB+JPT samples. The putative favored mutation (rs4988235; blue square in Figure 5.5) is ranked first by iSAFE in both (FIN and FIN+CHB+JPT). This blue signal disappears in East Asian populations (CHB and JPT). The CHAI score for blue-favored haplotype (represented by rs4988235) is zero. Therefore, the CHAI method suggests that the well-characterized selective sweep in European population (blue SNPs in Figure 5.5) is not an adaptive introgression.

The yellow (shared). The yellow sweep, in Figure 5.5, is shared between CHB, JPT, and FIN populations and the CHAI method suggests it is an adaptive introgression, perhaps from an unknown archaic hominin. Yellow SNPs have the dominant signal of selection in East Asian population (CHB and JPT) and this signal also exist in European (FIN) where these yellow SNPs are linked to, and dominated by, the blue sweep. As you can see, in the left panel of Figure 5.5

for combined FIN+CHB+JPT, the blue and yellow signals are clearly separated in frequency domain. This yellow-haplotype is the top iSAFE candidate in CHB and JPT populations and is predicted to be adaptive introgression (CHAI ≈ 1) in CHB, JPT, and FIN populations. The iSAFE candidate haplotype in CHB population includes 124 fully-linked SNPs that are also highly-linked in FIN population with significant iSAFE scores (Figures 5.5, 5.6). 71 (out of 124) of these SNPs are missed in archaic genome sequences (the Denisova and Altai, Vindija33.19, and Mez1 Neanderthals). Only 2 SNPs (out of remaining 53 SNPs, 4% match) are in derived state in the Mez1 Neanderthal sample and all other are in ancestral state in all other archaic samples (0% match). This observation suggests that this haplotype is not introgressed from archaic hominin like Neanderthals and Denisovans. However, as this haplotype also exists in AFR population in low frequency ($\sim 32\%$), it could be introgressed from an unknown archaic hominin, perhaps in Africa.

The red (CHB only). The most interesting result in this region is for the red SNPs that are specific to CHB population. Our analysis suggests the red sweep, in Figures 5.5, is an adaptive introgression in CHB population from a Neanderthal-like archaic hominin. We elaborate on this below.

Adaptive introgression from Neanderthals at *LCT* locus

We further investigate the red haplotype in Figure 5.6. The CHAI score predicts this haplotype as adaptive introgression (Figure 5.6a). The iSAFE signals as a function of chromosomal position, for yellow and red SNPs are not separable (Figure 5.6b). However, in the frequency domain the red and yellow signals are well-separated (Figure 5.6c) and the red haplotype shows signal of a new sweep in CHB population. The solid-box shows the red haplotype in Figure 5.6a-c. The shared sweep (yellow sweep in Figures 5.5, 5.6) is also demarcated by dashed-box. As we mentioned before, this shared-yellow sweep is predicted by CHAI to be an adaptive introgression but it does not match to any of the archaic genome sequences. However, it can be introgressed from an unknown archaic hominin.

In contrast, the red haplotype not only is predicted by CHAI method to be adaptively introgressed (Figure 5.6a), but also it is highly similar to the Vindija33.19 Neanderthal with 94% match. The upper panel of Figure 5.6d shows frequency of 149 SNPs on the red haplotype that have CHAI ≈ 1 only in CHB population (solid-box). The lower panel of Figure 5.6d shows the state of the alleles in three Neanderthals (Altai,Vindija33.19, and Mez1) and the Denisova. These SNPs have median frequency zero in FIN, EUR, and AFR populations and 0.01 and 0.07, 0.26 in SAS, JPT, and CHB, respectively. We showed this haplotype have a signal of new selection (red SNPs) in CHB population near an older sweep that is shared between European and East Asian populations (yellow SNPs). Interestingly, derived alleles of these SNPs match in 94%, 73%, 65%, and 41% of the cases, if not missed, with Vindija33.19, Altai, Mez1, and Denisova, respectively. Based on these observations, the red haplotype is a candidate for a recent adaptive introgression in CHB population on *LCT* locus from a Neanderthal-like archaic hominin.

To sum up, these observations suggest *LCT* locus seems to be target of multiple selection events in human population (see Figure 5.5). Some of these candidate favored haplotypes are adaptively introgressed. Comparison to the genome of archaic sample confirm that a haplotype (red haplotype in Figures 5.5, 5.6) is adaptively introgressed from an archaic hominin like Neanderthals into the CHB population.

5.5 Summary and discussion

Together with recent advancement in ancient genome sequencing technologies and availability of archaic hominin genome sequences, different methods have been proposed to infer introgressed segments of the genome. An introgressed haplotype at high frequency, relative to other populations, can be a signal of adaptive introgression. However, introgression increases LD and also distorts the pattern of allele frequency distribution. Therefore, using standard test of selection on region with introgression can easily lead to false inference of selection.

The CHAI method captures adaptively introgressed haplotypes without knowledge of archaic samples. Our results suggest that statistics obtained from the coalescent structure of a region under a selective sweep can detect favored haplotypes in human populations that are introgressed from archaic hominins without knowledge of archaic reference. CHAI is a supervised machine learning approach that is designed to minimize false inference of other types of selection, including hard/soft sweeps and adaptation on recombination suppressor mechanisms (RSM) like chromosomal inversion. All the data used in this Chapter are provided in Section B.5.

5.6 Acknowledgments

Chapter 5, in full, contains material from Ali Akbari and Vineet Bafna. "Detecting adaptive introgression in human populations without knowledge of the archaic samples", which is currently being prepared for submission for publication of the material. I was the primary investigator and author of this material.

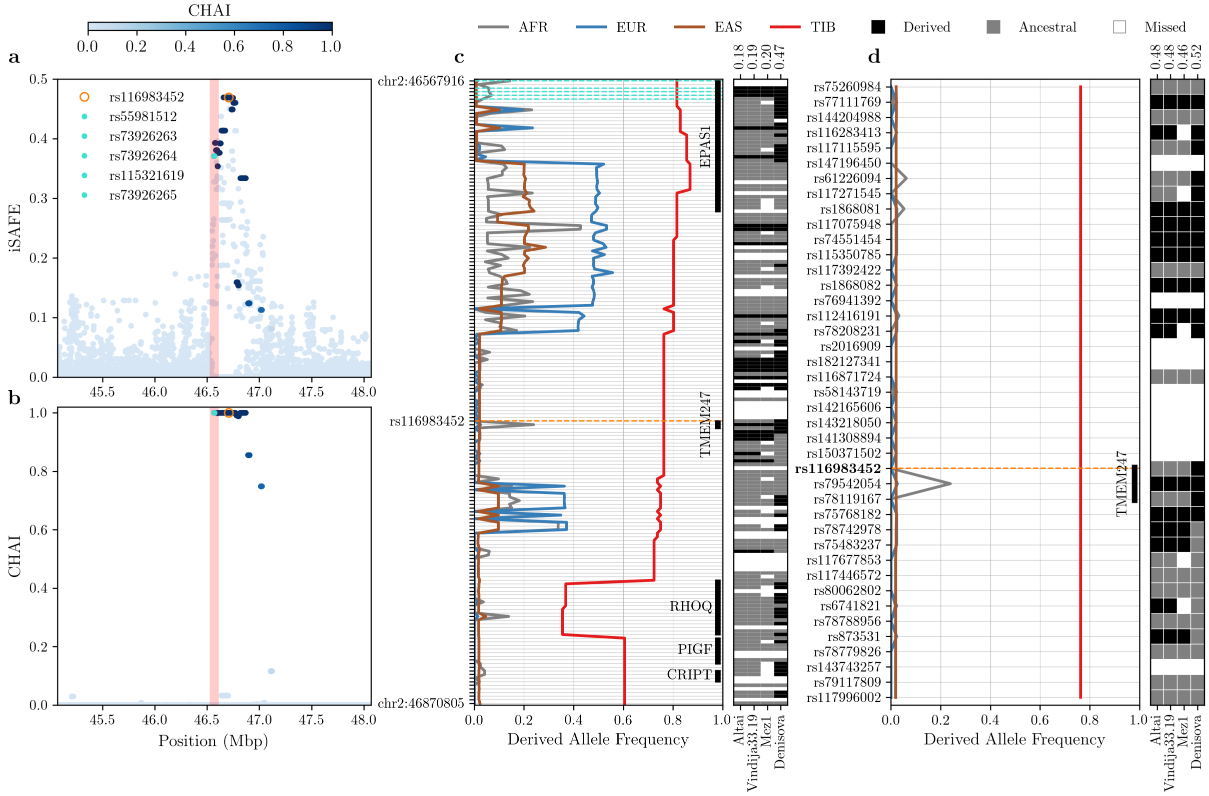


Figure 5.3: **Adaptive introgression at *EPAS1* locus in Tibetan highlanders**. **(a, b)** The iSAFE and CHAI scores for all SNPs in 3 Mbp region around *EPAS1* locus in Tibetan highlanders. The span of *EPAS1* gene (chr2:46525050-46611799) is shaded by red. **(c)** Frequency of 173 SNPs (chr2:46567916-46870805) with CHAI > 0.98 in Tibetan Highlanders (TIB, red) and three 1000GP super populations EAS (brown), EUR (blue), and AFR (gray). The state of the allele in three Neanderthals (Altai, Vindija33.19, and Mez1) and the Denisova are shown in the right panel. **(d)** The iSAFE top haplotype have 41 fully-linked SNPs around *TMEM247* (chr2:46657114-46730100), 45 kbp down stream of the *EPAS1*. One of these 41 SNPs is missense (rs116983452; orange color) and only exists in the Denisova. The values on top of right panels in **(c, d)** are the proportion of the derived alleles (excluding missed ones) in the corresponding archaic sample which can be considered as a measure of similarity. **(a-c)** Turquoise color represents the five SNPs used in Huerta-Sánchez et al. 2014 [14], that led them to conclude this segment of genome is adaptively introgressed from an archaic hominin like Denisovans.

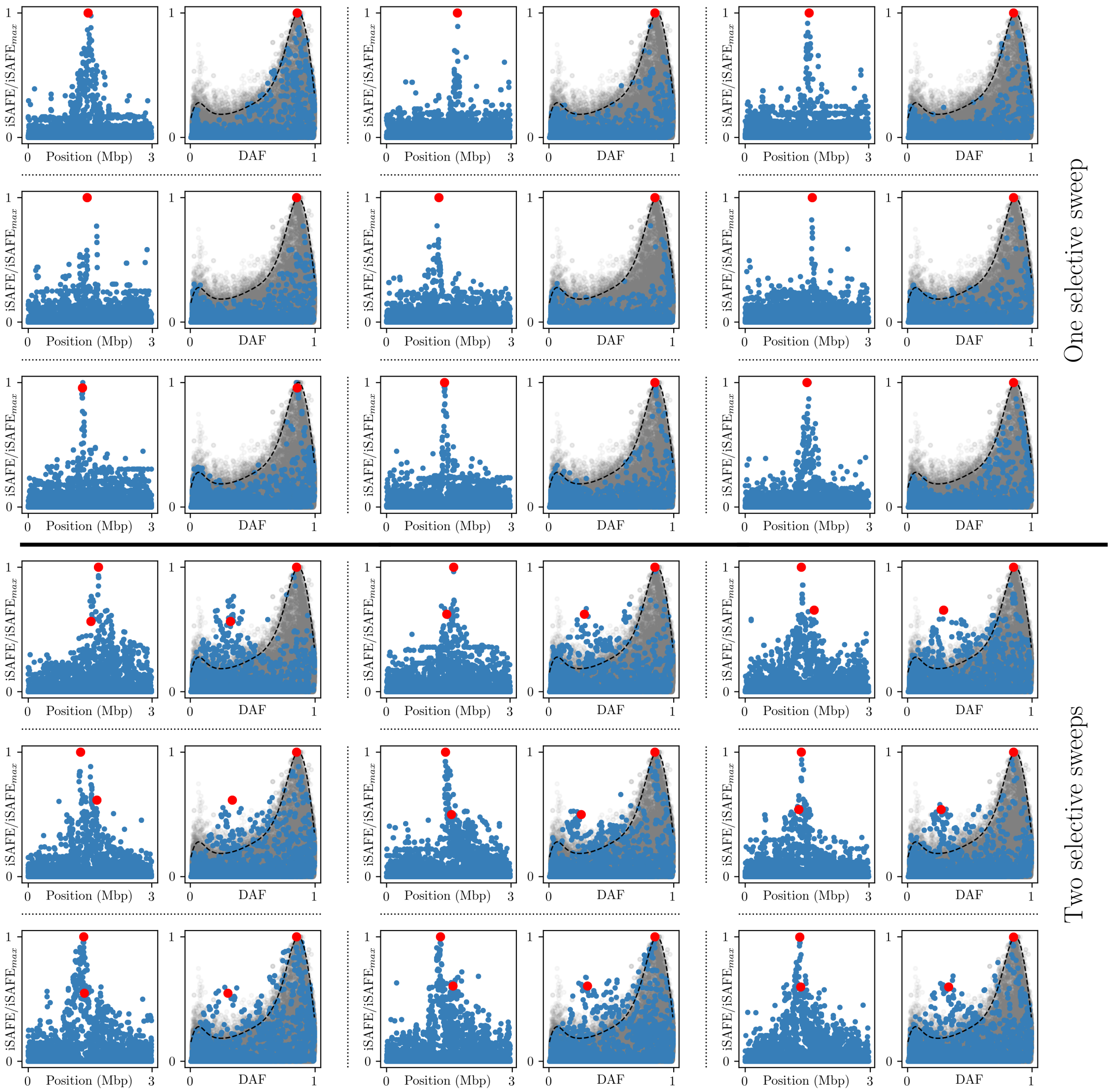


Figure 5.4: **Distinguish multiple selections**. When two selections occur near each other their signals interfere and usually not separable by looking at their signals as function of their chromosomal position. However, if these sweeps have difference in frequency, sometimes they can be distinguished by their iSAFE score in the frequency domain (regardless of their chromosomal position in the target region). Here we show a few examples of single selective sweep (top three rows) versus multiple selections (bottom three rows). Panels between dotted lines represent an independent simulation. The gray shade (in right panels) show the expected distribution of SNPs if there is only one selective sweep. The black-dashed curve represent the 99 percentile of the distribution, conditioned on derived allele frequency (DAF). These distributions are derived by 100 simulations to mimic the dominant sweep (sweep with highest iSAFE score). As you can see in these examples a significant deviation from the expected distribution (gray shade and black-dashed line) can be a signal of extra selective sweep, while in a single selective sweep SNPs do not deviate significantly from the expectation. The red circles are favored mutations.

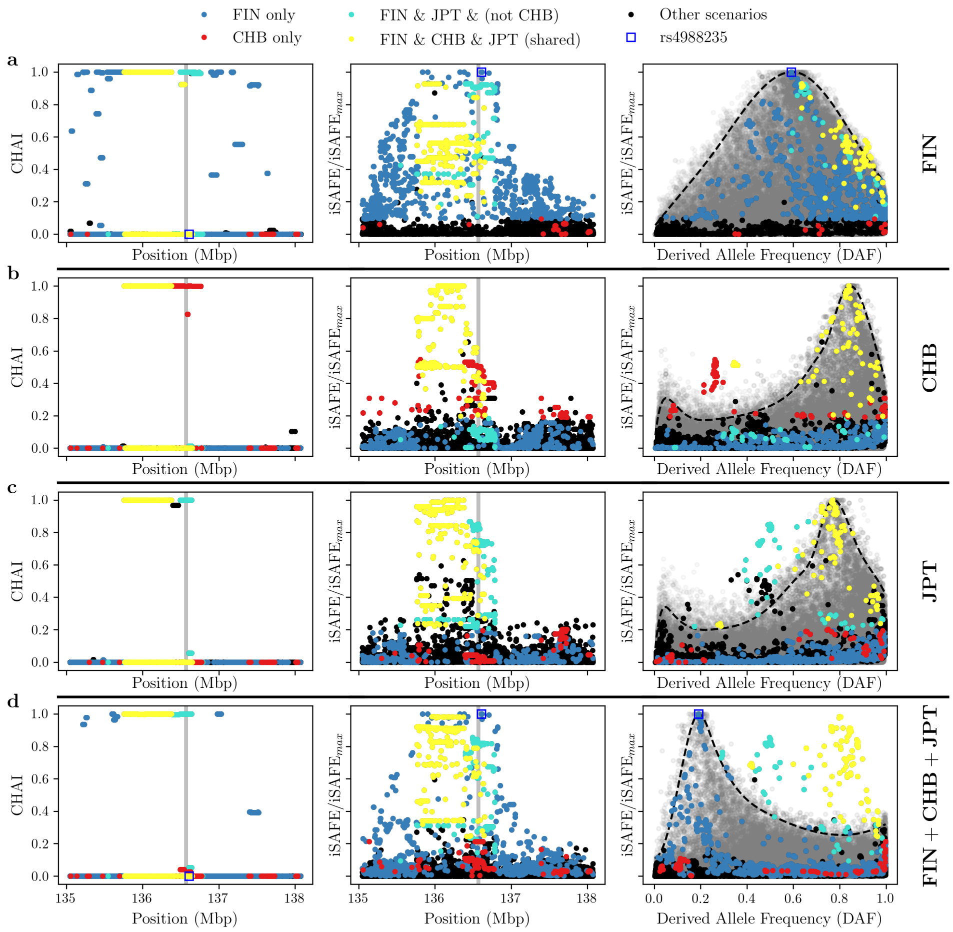


Figure 5.5: **Multiple selections at *LCT* Locus**. In Figure 5.4 we showed when two selection occur near each other sometimes their signals can be distinguished by their iSAFE score in the frequency domain even when their signals as a function of chromosomal position are not separable. Right panels, along with Figure 5.4, imply that this pattern of signals is consistent with the pattern of multiple selections on *LCT* locus. The coloring of these SNPs are based on the significance of their iSAFE score (*P* value < 1e-4) in FIN, CHB, and JPT populations. For example, the blue SNPs only shows a signal of selection in FIN population and they represent the well-characterized selective sweep associated with Lactose persistence in European population and this strong signal is observed in FIN population and combined FIN+CHB+JPT populations with the putative favored mutation (rs4988235; blue square) being ranked first by iSAFE score in both (FIN and FIN+CHB+JPT). This blue signal disappears in East Asian populations (CHB and JPT). Yellow SNPs have the dominant signal of selection in East Asian population (CHB and JPT) and this signal also exist in European (FIN) where these yellow SNPs are highly linked to, and dominated by, the blue SNPs. As you can see, in the bottom-right panel for combined FIN+CHB+JPT, these two signals (blue and yellow) are clearly separated in frequency domain. The red SNPs are specific to CHB population and we explain them in detail in Figure 5.6. These SNPs show signal of a new sweep in CHB population with a zero or close to zero frequency in non-Asian populations of 1000GP and surprisingly this haplotype not only predicted by our method to be adaptively introgressed but also it is highly similar to the Vindija33.19 Neanderthal (94% match). There is also a turquoise signal that is shared between JPT and FIN populations and dominated by the blue signal in European. The gray shade (in right panels) show the expected distribution of SNPs if there is only one selective sweep. The black-dashed curve represent the 99 percentile of the distribution, conditioned on derived allele frequency (DAF). These distributions are derived by 100 simulations to mimic the dominant sweep (sweep with highest iSAFE score).

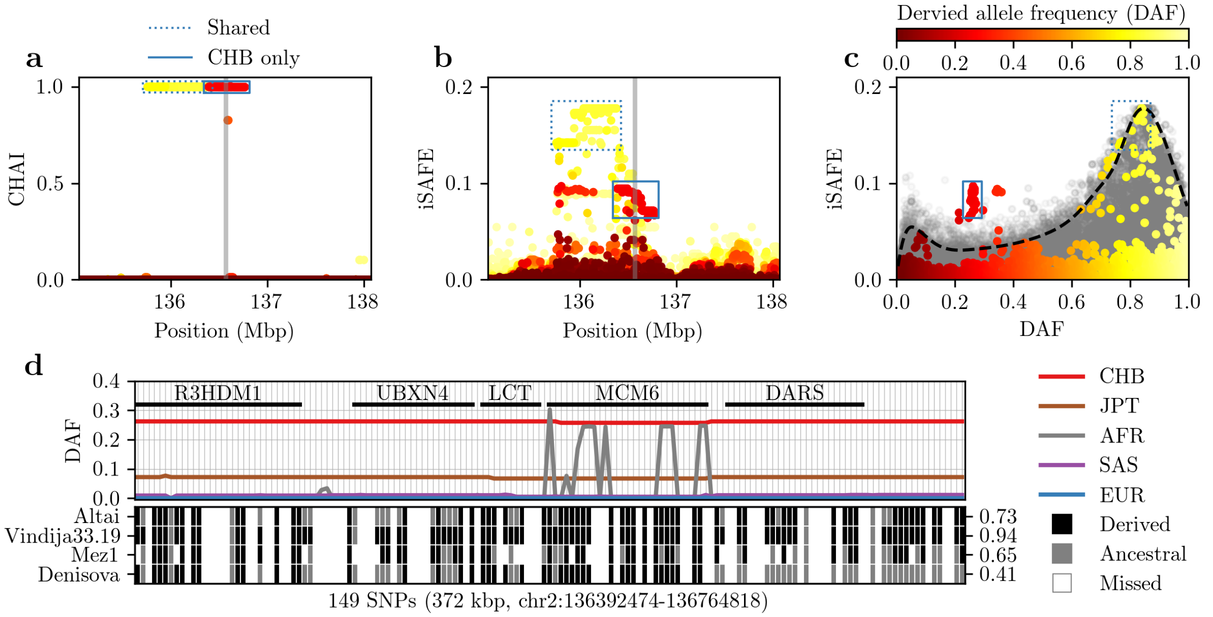


Figure 5.6: **Adaptive introgression at *LCT* locus**. **(a)** The CHAI and **(b)** iSAFE scores for all SNPs in 3 Mbp region around *LCT* locus in CHB population of 1000GP. The gray-shaded region is the location of *LCT* gene. **(c)** The iSAFE score as a function of derived allele frequency (DAF) suggests multiple selections are happening in this locus. The yellow signal represents a sweep that is shared between all non-African populations and the red signal is only in CHB population (see Figure 5.5). The gray shade in **(c)** shows the expected distribution of SNPs if there is only one selective sweep. The black-dashed curve represent the 99 percentile of the distribution, conditioned on derived allele frequency (DAF). These distributions are derived by 100 simulations to mimic the dominant sweep (sweep with highest iSAFE score). **(d)** Frequency of 149 SNPs that have high CHAI score only in CHB population (red SNPs). The state of the allele in three Neanderthals (Altai,Vindija33.19, and Mez1) and the Denisova are shown in the lower panel. These SNPs have median frequency zero in FIN, EUR, and AFR populations and 0.01 and 0.07, 0.26 in SAS, JPT, and CHB, respectively. Interestingly, derived alleles of these SNPs match in 94%, 73%, 65%, and 41% of the cases, if not missed, with Vindija33.19, Altai, Mez1, and Denisova, respectively. Based on these observations, this red haplotype is a candidate for a recent adaptive introgression in CHB population on *LCT* locus from a Neanderthal-like archaic hominin. Besides, *LCT* locus seems to be target of multiple selection events in human populations (see Figure 5.5). The values on the right y-axis of the lower panel in **(d)** are the proportion of the derived alleles (excluding missed ones) in the corresponding archaic sample which can be considered as a measure of similarity.

Appendix A

Supplementary notes: iSAFE results on selective sweeps in human populations

A.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 [56, 88, 90, 94, 95, 97]. iSAFE results for these loci are summarized in **Figure 3b** and Figure **8** and Table 4.1.

We also examined 14 other loci reported to be under selection with one or more candidate favored mutations [32, 56, 79, 80, 82].

A.2 Pigmentation genes

SLC45A2/MATP. This region is involved in human pigmentation pathways and is a target of selective sweep in European population [79]. A nonsynonymous mutation rs16891982 is associated with light skin pigmentation and is believed to be the favored variant [56, 79]. This mutation is also ranked first by iSAFE out of $\sim$21,000 mutations (5 Mbp) in CEU population with a significant score (see **Figure 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 [80]. This mutation is ranked first by iSAFE in CHB+JPT (see **Figure 3c**, iSAFE = 0.24, $P = 1.4\text{e-}6$) out of $\sim$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 [79]. A second intronic variant rs10831496 in GRM5 gene, 396 kbp upstream of TYR, has been shown to have a strong association with skin color [83].

In contrast, iSAFE ranks mutation rs672144 as the top candidate for the favored variant region out of $\sim$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 **Figure 3c** and Figure **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 Figure **10**. We plotted the haplotypes carrying rs672144 and found (**Figure 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 (Figure **10c**), but not to each other. The EUR haplotypes can be partitioned into 4 clusters

(Figure **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 [56, 79, 112], 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 [79]. rs1667394 is also linked to blond hair and blue eyes [112]. Some other mutations, many fully linked, (rs4778138, rs4778241, rs7495174, rs1129038, rs916977) are also associated with blue eyes [112]. 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 Figure **11** and **Figure 3c**). This variant is close to fixation in all populations except for AFR, where $\nu = 20\%$ (see Figure **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 (Table A.1).

KITLG. This genomic region has been linked to skin pigmentation [113] 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 [80], with a candidate variant rs642742, that is associated with skin pigmentation [113].

iSAFE analysis identified the same mutations gaining the top rank in multiple populations (Figure **12**). Top rank mutations in EUR, SAS, EAS, and AMR populations are shown in Table A.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 [32]. 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" [114]. iSAFE suggests 10 strongly linked mutations located along a 9 kbp region located 84 kbp downstream of TRPV6 (see Figure A.5). These mutations are ranked in the top 10 in all non-African populations (Table A.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 (Figure **13**).

A.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 [56]. This mutation is ranked 12 by iSAFE in CHB+JPT (see Figure **9**, iSAFE $= 0.45$, $P < 1.34\text{e-}8$). 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 [32]. A non-synonymous mutation in this gene is associated with Alcohol dependence [115]. We tested this gene in CHB+JPT populations. iSAFE rank, in 2 Mbp around ADH1B gene, for the candidate mutation (rs1229984) is 8 (see Figure **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 [116].

A.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 [82]. 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 [117]. We tested this region in GBR population of 1000GP. This candidate mutation is jointly ranked first with two other mutations rs7137828, rs7310615 (see **Figure 3c**, iSAFE $= 0.27$, $P = 1.6\text{e-}7$). rs7137828 is an intronic mutation in ATXN2 that is associated with Primary Open Angle Glaucoma that is a leading cause of blindness worldwide [118]. The other first rank mutation (rs7310615) is associated with blood expression levels of SH2B3 [119]. Surprisingly, all of the top 10 mutations, ranked by iSAFE have a known association to a phenotype (Table A.4), and are highly linked (Figure A.6).

CYP1A2/CSK. We tested a 5 Mbp region around these genes in GBR population of 1000GP. The proposed mutation rs1378942 by [82] with frequency 0.69 in GBR population is ranked 89 by iSAFE (iSAFE = 0.13, $P = 7.0\text{e-}5$). The top-ranked mutation rs2470893 (**Figure 3c**, iSAFE = 0.16, $P = 2.7\text{e-}5$) is between CYP1A1 and CYP1A2 with frequency 0.40 in GBR and is associated with Caffeine metabolism [84]. 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 (Figure **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 [82] is ranked 4 ($P = 0.1$).

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

A.5 Other genes

PSCA. This gene has been reported as a target of selection in YRI population [32]. A $5'$UTR mutation rs2294008 proposed as a candidate favored mutation in this region that is associated with urinary bladder and gastric cancers [120, 121]. The signal of iSAFE in 5 Mbp around this gene in YRI population is weak (see Figure **9**, peak iSAFE $= 0.04$, $P = 2.4\text{e-}3$). 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 [121]. 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).

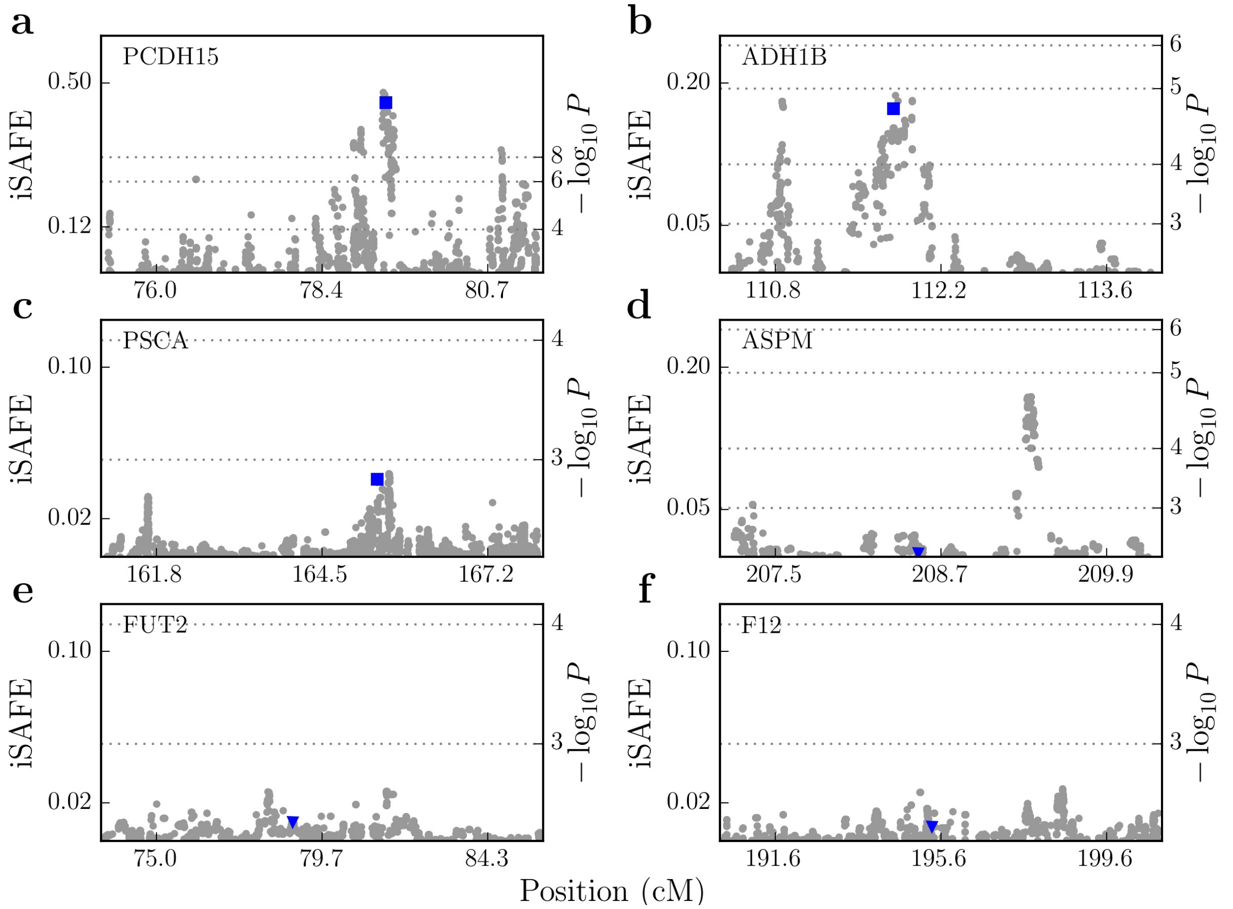


Figure A.1: **iSAFE on targets of selection.** iSAFE on 5 Mbp regions reported to be under selection. Putative favored mutation is shown in blue square when it is among iSAFE top rank mutations, and in blue triangle when the signal of selection is very weak (peak iSAFE $\ll 0.1$). The right axis is empirical P value (see Section B.4). (**a,b**) PCDH15 and ADH1B loci with 207 samples (414 haplotypes) from CHB+JPT populations. (**c**) PSCA locus with 108 samples (216 haplotypes) from YRI population. (**d,e,f**) ASPM, FUT2, and F12 loci with 91 samples (182 haplotypes) from GBR population.

ASPM. This gene is reported to be a target of weak selection in GBR population [32]. The signal in 2 Mbp around this gene is very weak (see Figure **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.

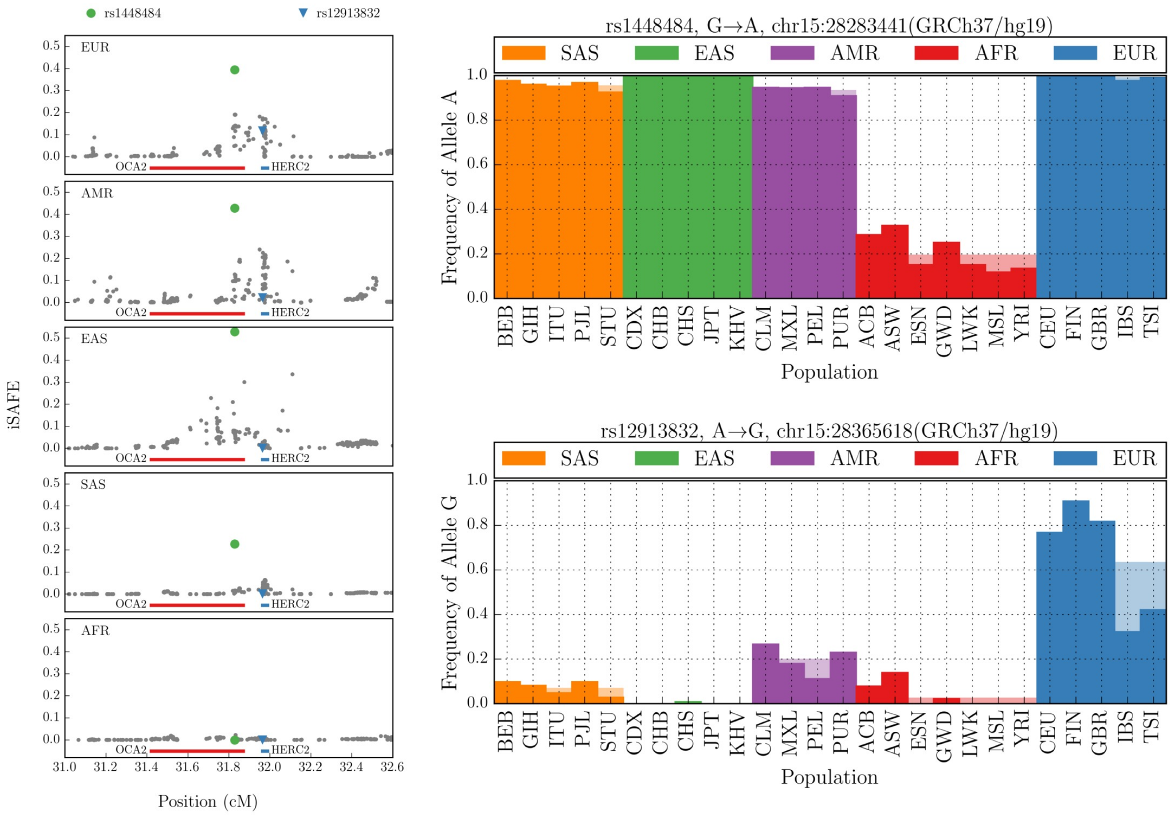


Figure A.2: **iSAFE on the OCA2-HERC2 locus.** The mutation rs1448484 is the iSAFE top rank mutation in all the population of 1000GP except African that does not show any signal of selection in this region. rs12913832 is a candidate favored mutation for the selection in European, proposed by Wilde et al. (2014) [79]. Table A.1 provides iSAFE rank of some other candidate mutations associated with pigmentation in this region (see Section A.2).

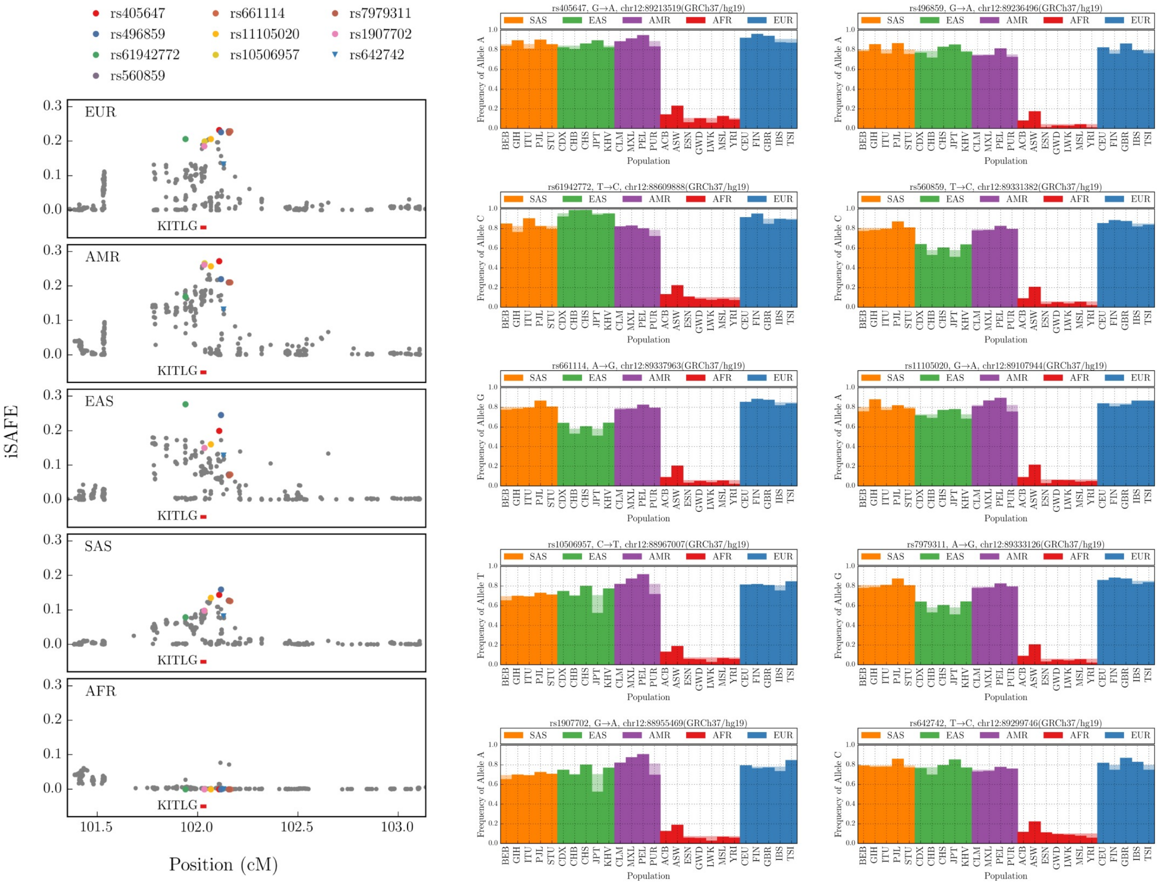


Figure A.3: **iSAFE on the KITLG locus.** iSAFE top rank mutations (circles) and candidate mutation rs642742 (blue triangle) proposed by Miller et al. (2007) [113]. See Section A.2 and Table A.2 for more details.

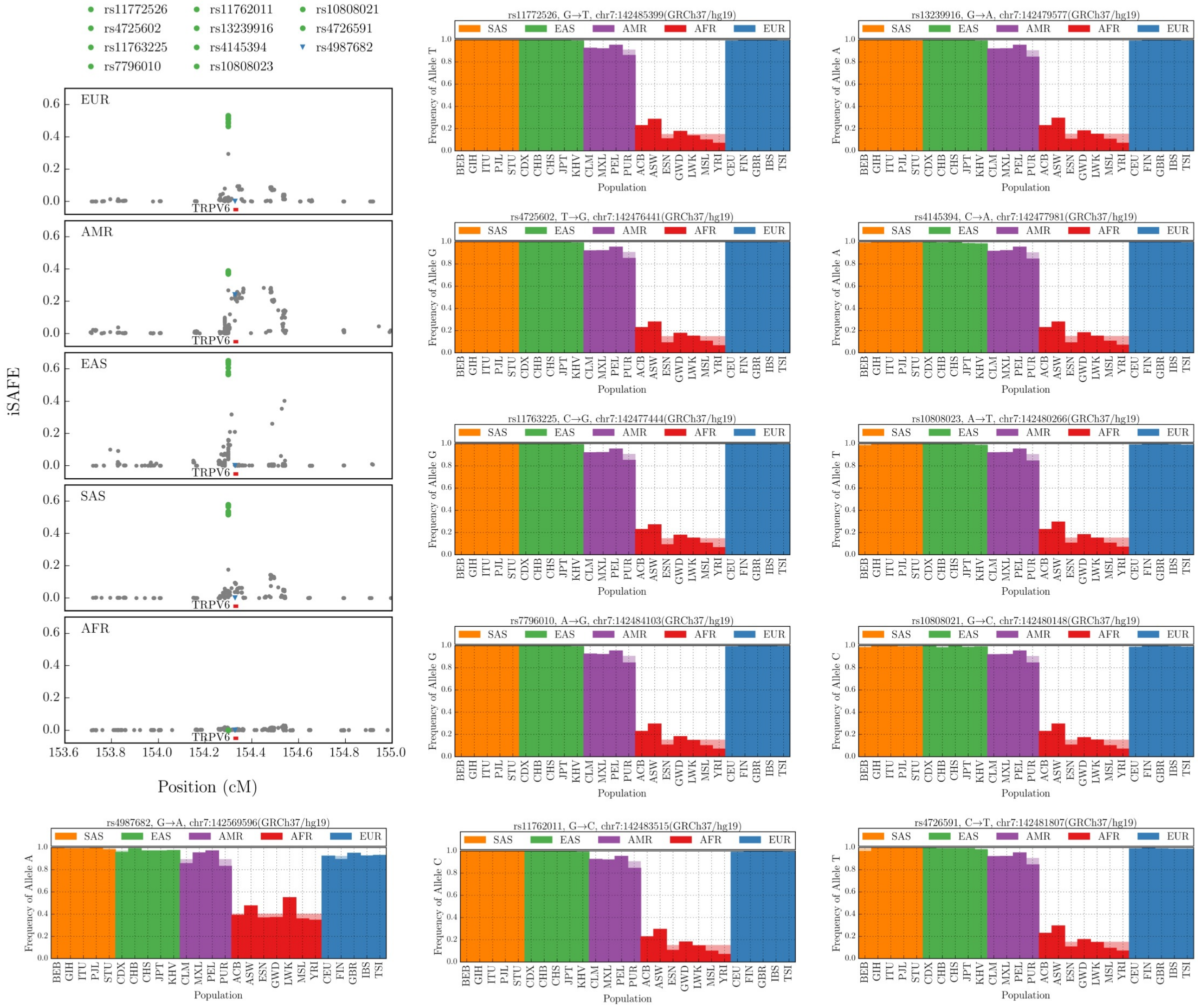


Figure A.4: **iSAFE on the TRPV6 locus.** 10 mutations (rs11772526, rs4725602, rs11763225, rs7796010, rs11762011, rs13239916, rs4145394, rs10808023, rs10808021, and rs4726591) are highly linked and are top 10 iSAFE candidate mutations in all the 1000GP populations except for AFR where there is no signals of selection. See Section A.2 and Table A.3 for more details.

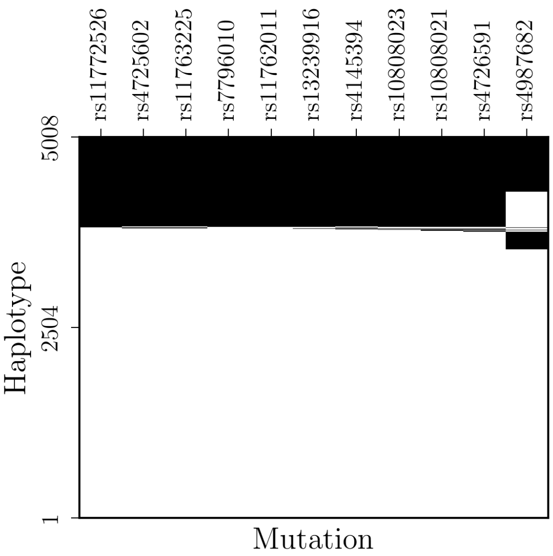


Figure A.5: **SNP matrix of TRPV6 top candidates.** Haplotypes of top 10 iSAFE mutations, and the proposed mutation (rs4987682) by [32], in 5 Mbp around TRPV6 in 2504 × 2 haplotypes of 1000GP are shown. These mutations are sorted by their iSAFE rank from left to right. iSAFE top 10 mutations span a 9 kbp region(chr7:142476441-142485399, GRCh37/hg19). White is derived and black is ancestral allele.

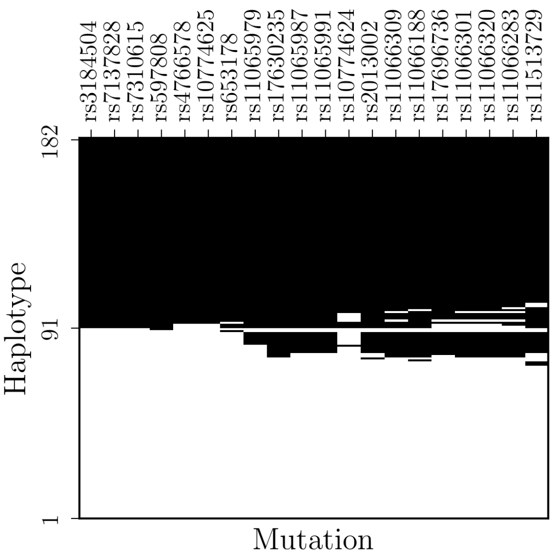


Figure A.6: **SNP matrix of ATXN2-SH2B3 top candidates.** Haplotypes of top 20 iSAFE mutations in 5 Mbp around ATXN2-SH2B3 in GBR population are shown. These mutations are sorted by their iSAFE rank from left to right. They span a 1.07 Mbp region around ATXN2-SH2B3 region (chr12:111833788-112906415, GRCh37/hg19). White is derived and black is ancestral allele. Most of these mutations are associated to a phenotype (see Table A.4).

Table A.1: **iSAFE rank of putative favored variants of OCA2-HERC2.** iSAFE rank of candidate mutations proposed by [79, 112] 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 Section B.4.

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

Table A.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 [113] is also reported in the last row). Frequency and iSAFE score for this region in all the 1000GP populations are provided in Figure**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

Table A.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

Table A.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 Section B.4.

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	[122]
rs7137828	1	2.2e-7	ATXN2	intron	0.5	Primary open-angle glaucoma	[118]
rs7310615	1	2.2e-7	SH2B3	intron	0.5	Fibrinogen levels	[119]
rs597808	4	2.7e-7	ATXN2	intron	0.49	Systemic lupus erythematosus	[123]
rs4766578	5	3.0e-7	ATXN2	intron	0.51	Vitiligo	[124]
rs10774625	5	3.0e-7	ATXN2	intron	0.51	Systemic lupus erythematosus, Retinal vascular caliber	[123]
rs653178	7	3.1e-7		regulatory	0.5	Blood pressure and hypertension, Myocardial infarction, & more	[122]
rs11065979	8	4.4e-7		intergenic	0.47	Cancer (pleiotropy)	[125]
rs17630235	9	4.6e-7	TRAFD1	downstream	0.43	Body mass index	[126]
rs11065987	10	4.9e-7		intergenic	0.45	Tetralogy of Fallot, Coronary artery disease, & more	[127]
rs11065991	10	4.9e-7	BRAP	intron	0.45		
rs10774624	12	5.2e-7	RP3-473L9.4	intron,nc	0.52	Rheumatoid arthritis	[128]
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	[129]
rs11066301	17	1.9e-6	PTPN11	intron	0.46	Hematological parameters	[130]
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		

Appendix B

Supplementary notes

B.1 Default simulation parameters

Neutral and sweep samples were generated with the simulator msms[131]. 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 [81, 132] and a recombination rate of approximately $r = 1.25 \times 10^{-8}$ per base pair per generation [133] 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 ν_0 was $1/N$ (hard sweep), and the frequency of the favored mutation (ν) 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.

B.2 A model of human demography

We simulated the demography of 1000GP AFR, EUR, and EAS populations with the parameters shown in Figure B.1, based on a popular demographic model of human population [134]. In the case of positive selection, the selection coefficient was set to $s = 0.05$, and the starting favored allele frequency $\nu_0 = 0.001$. The time of onset of selection was chosen at random (using the distribution in Figure B.1) 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.

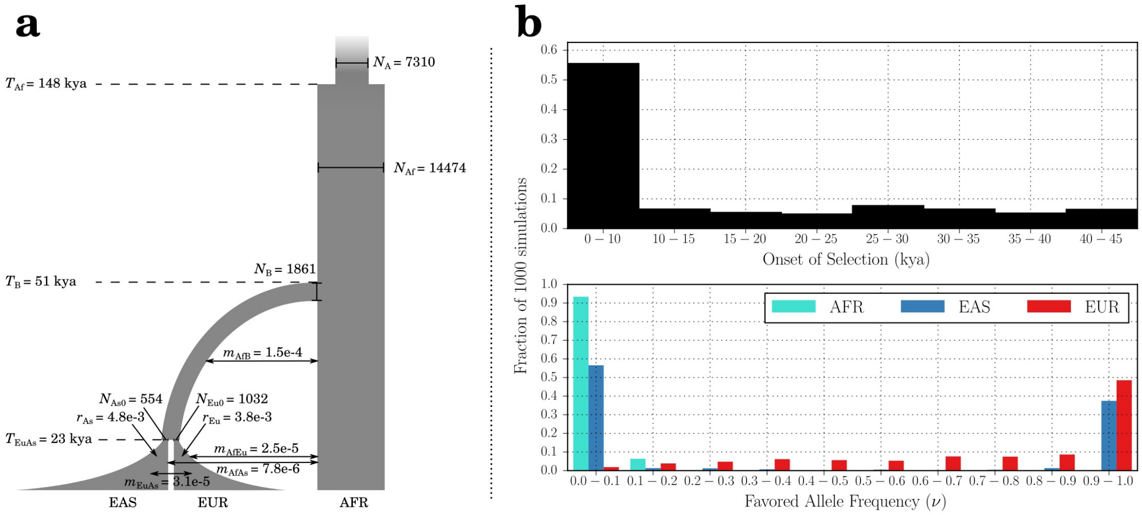


Figure B.1: **Simulation of selection on human demography.** (**a**) A model of human demography described by Figure 4 and Table 2 of Gravel et al. (2011) [134]. The model assumes an out-of-Africa split at time T_B , with a bottleneck that reduced the effective population from N_{Af} to N_B , allowing for migrations at rate m_{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 East Asian (EAS), European (EUR), and Africans (AFR) individuals were sampled. We used default values for simulation parameters not assigned. (**b**) We simulated 1000 selective sweeps on 5 Mbp region based on the model of human demography, and with selection coefficient $s = 0.05$ and starting favored allele frequency $\nu_0 = 0.001$. The selection happens in a random time, 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 (> 23 kya), both (EUR and EAS) are under selection.

B.3 Adding Outgroup Samples

Simulation of human population demography under neutral evolution (Figure **14**), shows $P(\text{MDDAF} > 0.78|D_T > 0.9) = 0.001$ (Figure **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 Figure **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 Figure **14**. The selection happens in a random time, with a distribution given in Figure **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 (Figure **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 Figure **3e**).

B.4 Computing empirical *P* values for iSAFE

We applied iSAFE on a neutrally evolving simulated population with window size of 5 Mbp, based on the European demography shown in Figure B.1. 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 1.34×10^{-8} for an iSAFE score of 0.304. Scores higher than this cutoff were considered to have $P < 1.34 \times 10^{-8}$.

B.5 Data availability

For all the following datasets, the genome build is GRCh37/hg19. We downloaded the phased haplotypes of the 1000GP [135](phase 3) dataset from ftp.1000genomes.ebi.ac.uk/vol1/ftp /release/20130502/. The ancestral allele dataset from Ensembl [136] (release 75) was downloaded from ftp.ensembl.org/pub/release-75/fasta/ancestral_alleles/. The physical position was converted into genetic position using the HapMap II [137] genetic map downloaded from ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20110106_recombination_hotspots/ . We used VCF files of archaic samples provided by Prüfer et. al. [138], available at cdna.eva.mpg.de /neandertal/Vindija/VCF/ . For analyzing the *EPAS1* region in Tibetans highlanders (TIB), we used 38 whole genome sequences of TIB provided by Lu et. al. [111]

Appendix C

Supplementary figures

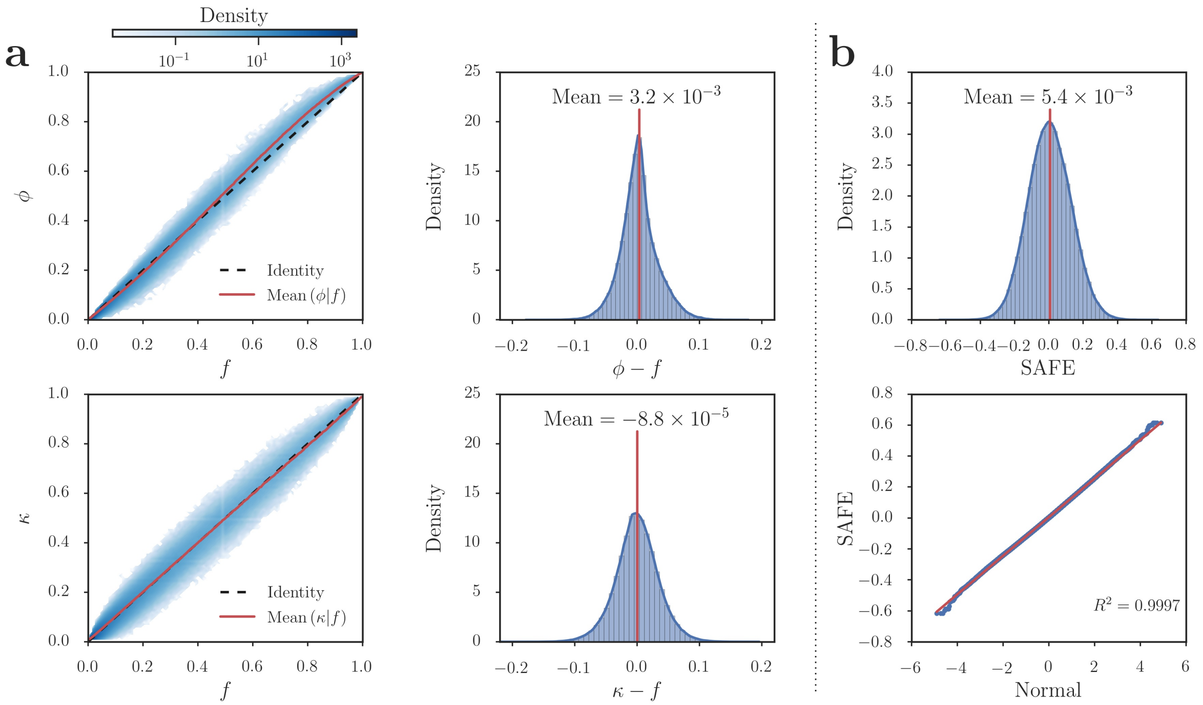


Figure C.1: **Empirical SAFE distribution.** (**a**) ϕ and κ as estimators of f. Empirical analysis, with 10,000 neutrally evolving population (about 3 million SNPs) with default parameter set, shows that ϕ and κ are (biased) estimators of allele frequency f ($f = i/n$ for all integers $i \in [1, n-1]$). (**b**) The top panel is the SAFE score Probability Density Function (PDF) of 10,000 neutrally evolving population (about 3 million SNPs with minor allele frequency > 0.05) with default parameter set. The bottom panel is Quantiles of the SAFE score against the quantiles of the Normal distribution. The coefficient of determination ($R^2 = 0.9997$) for the QQ-plot shows that Gaussian distribution is a good approximation to the SAFE score distribution.

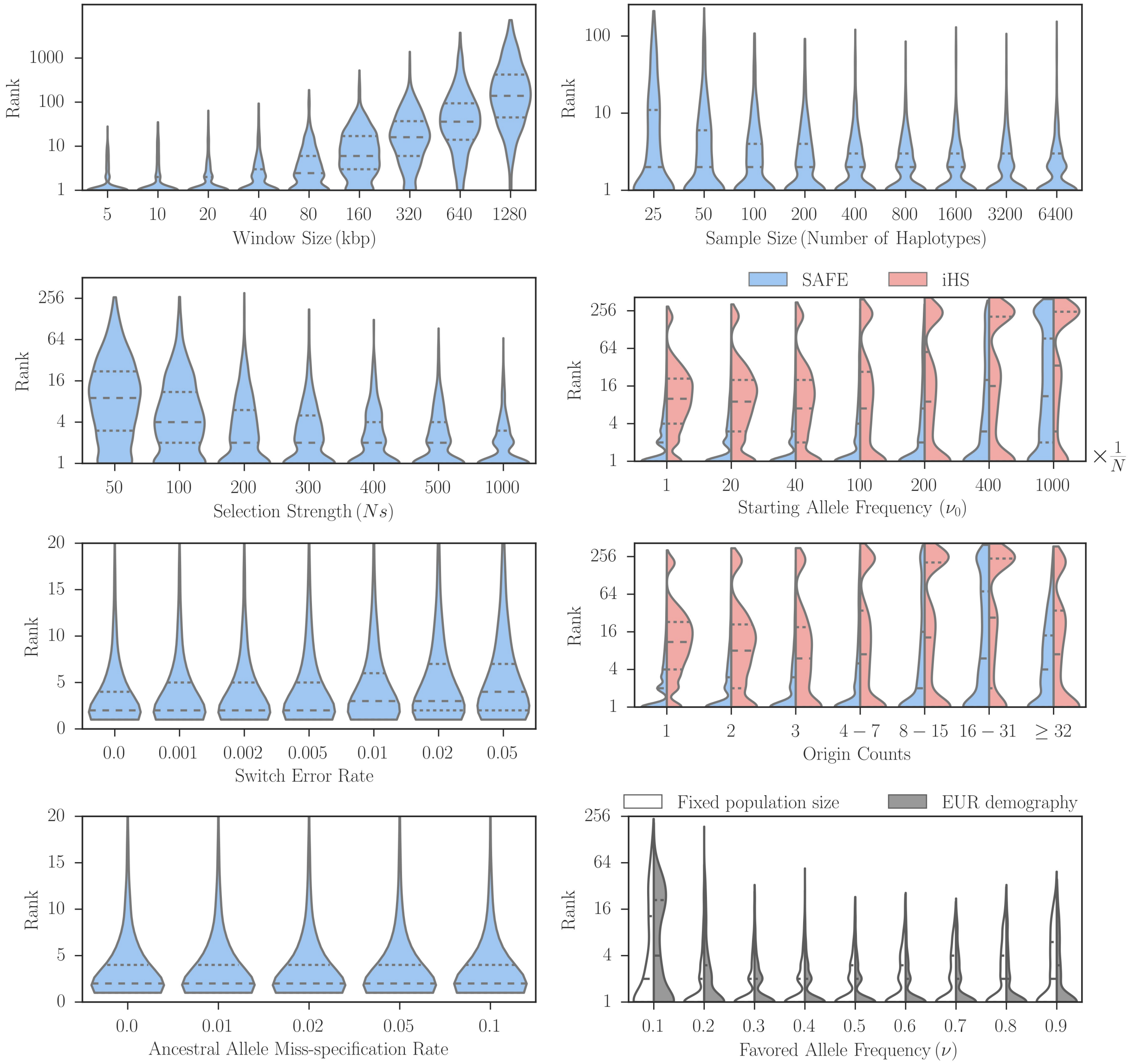


Figure C.2: **SAFE evaluation.** Performance of the safe score evaluated in different scenarios with 1000 simulations per bin. In each panel, we change one parameter and other parameters have their default values (see Section B.1). The fixed population size N = 20,000. The dashed (dotted) line represents median (quartile). In the bottom-right panel, white represents the result for a fixed size population model with default parameters and gray represents a model of human demography for EUR population (see Section B.2). The onset times of selection was post-bottleneck (23 kya-current) epochs.

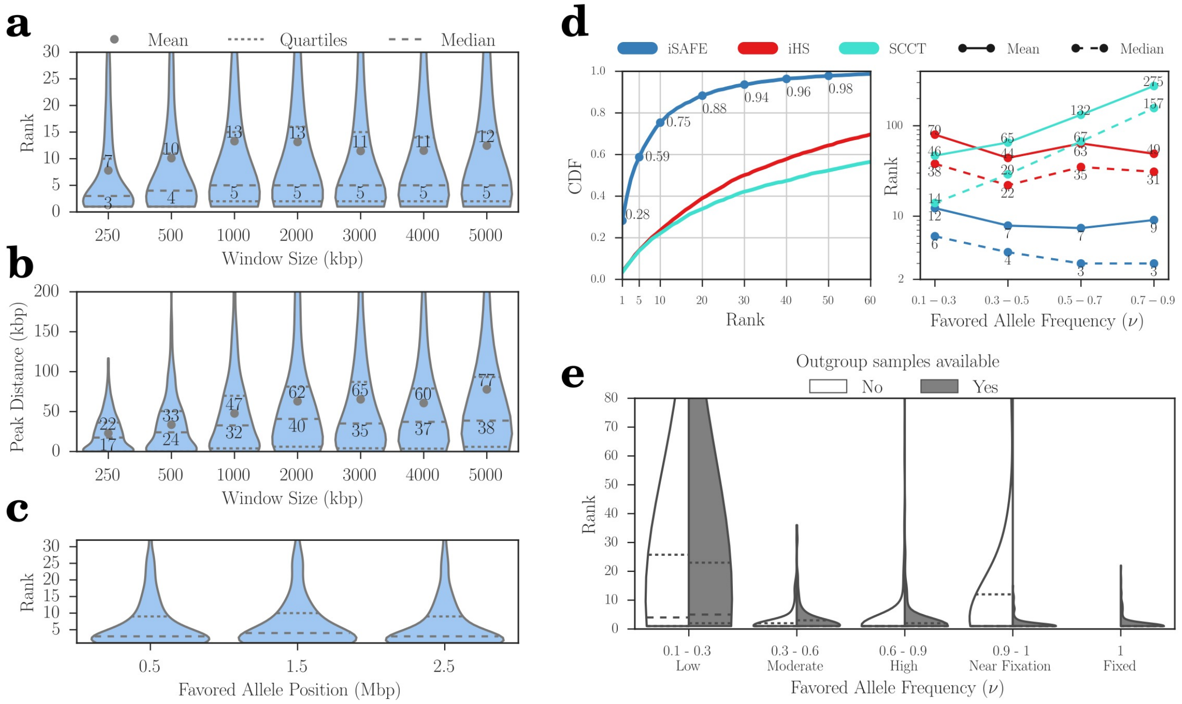


Figure C.3: **iSAFE evaluation.** (**a,b,c**) Performance of iSAFE measured by rank of the favored variant and the distance of the favored variant from the peak in 1000 simulations per bin. The dashed (dotted) line represents median (quartile). (**d**) Performance of iSAFE compared to iHS and SCCT measured by rank of the favored variant in 5000 simulations on 5 Mbp region around ongoing hard sweeps ($\nu_0 = 1/N$; $0.1 < \nu < 0.9$) with a fixed population size ($N = 20,000$) and default values for other simulation parameters. In the left panel, for any rank r on the x-axis, the y-intercept represents the proportion of samples where the favored allele had rank $\leq r$. In the right panel, solid (dashed) lines represent the mean (respectively, median) value of the favored allele rank. (**e**) iSAFE performance upon addition of outgroup samples. No deterioration is seen for low frequencies of the favored variant, but iSAFE performance improves dramatically when favored mutation is near fixation or fixed. The dashed (dotted) line represents median (quartile). This comparison is based on 1000 simulations of 5 Mbp genomic regions simulated using a model of human genome based on the human demography (Section B.2). The time of onset of selection was chosen at random (using the distribution in Figure B.1) after the out of Africa event, 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.

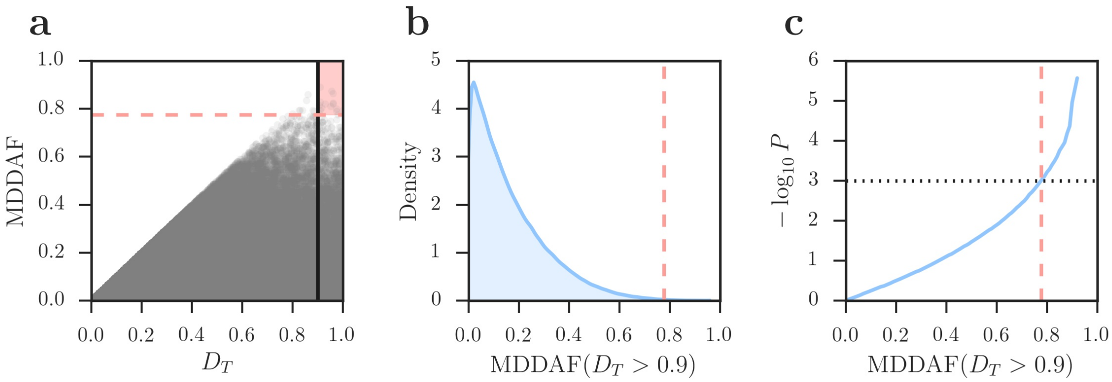


Figure C.4: **Maximum difference in derived allele frequency (MDDAF).** We simulated 25,000 instances of AFR, EUR, and EAS populations, based on a model of human demography (see Online Methods; Supplementary Figure 14). (**a**) The MDDAF score of mutations as a function of derived allele frequency in the target population D_T . (**b**) Distribution of the MDDAF score for mutations with $D_T > 0.9$. (c) Empirical P value of the MDDAF score for mutations with $D_T > 0.9$. The dashed-red lines represent the value 0.78, where MDDAF, given $D_T > 0.9$, has a P value less than 0.001.

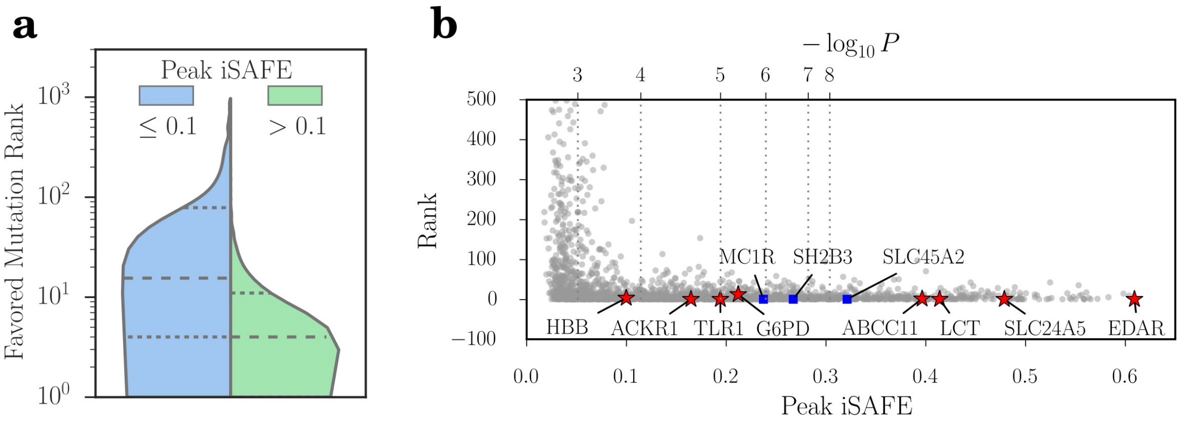


Figure C.5: **Peak iSAFE.** (**a**) Empirical analysis, with 5000 simulations on 5 Mbp region with a wide range of selection strength ($Ns \in$ [10, 50, 100, 200, 300, 400, 500, 1000]), shows difference in performance of iSAFE beyond a score threshold of 0.1 for peak value of iSAFE. (**b**) Rank of favored mutation as a function of peak iSAFE score (Bottom x-axis) or P value (top x-axis; see Section B.4) for the same data in part **a**. Each gray dot represents the favored mutation of a simulation using a wide range of selection coefficients. The performance deteriorates for iSAFE scores below 0.1. The dashed (dotted) line represents median (quartile).

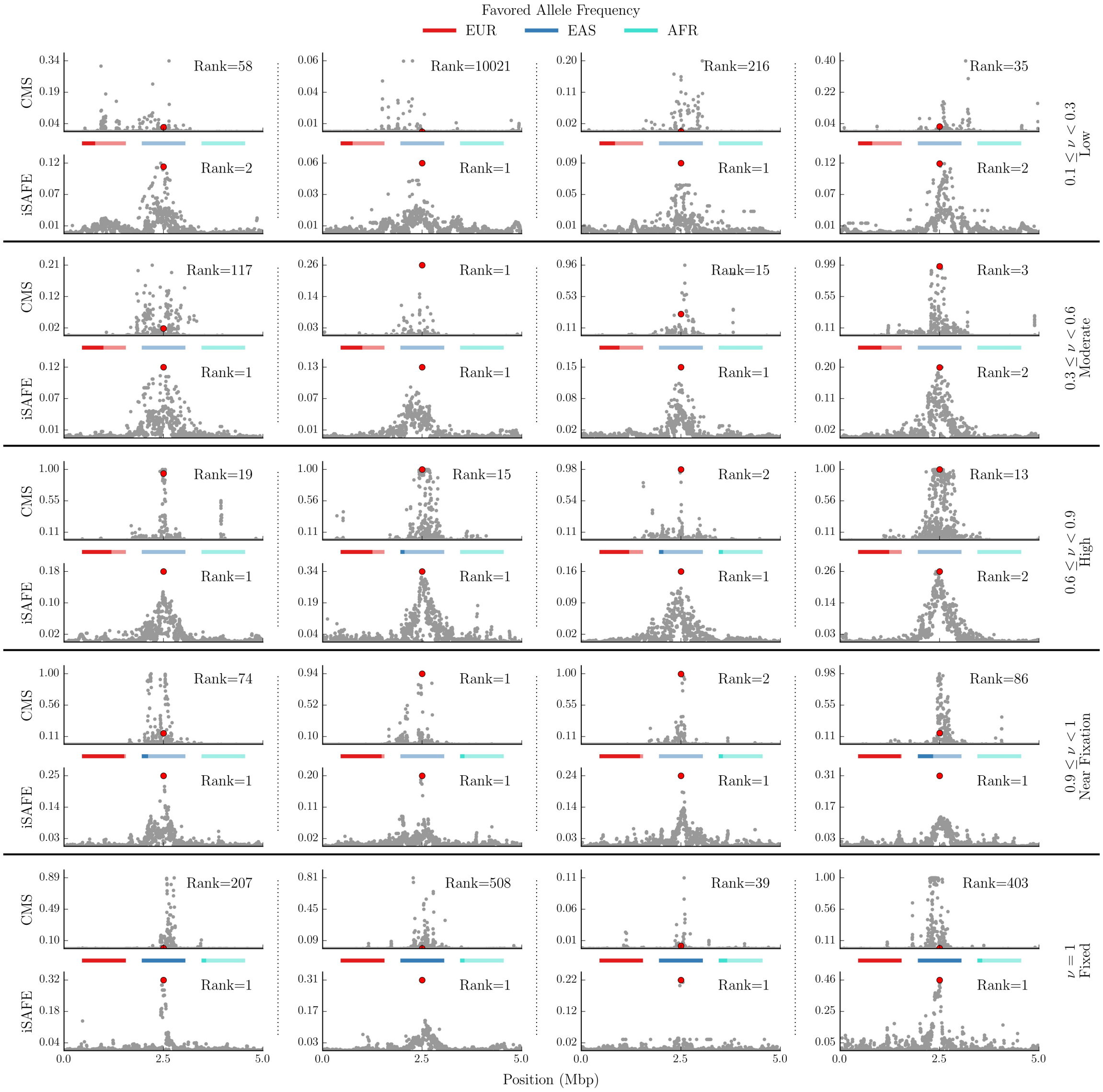


Figure C.6: **Demo I: iSAFE versus CMS in a model of human demography.** Comparing iSAFE and CMS signals in a model of human demography (see Section B.2). Solid-horizontal lines separate replicates based on the favored allele frequency (ν) in EUR as the target population, and dotted-vertical lines separate different replicates. The rank of the favored mutation (solid-red circle) for each test is shown on the top-right corner.

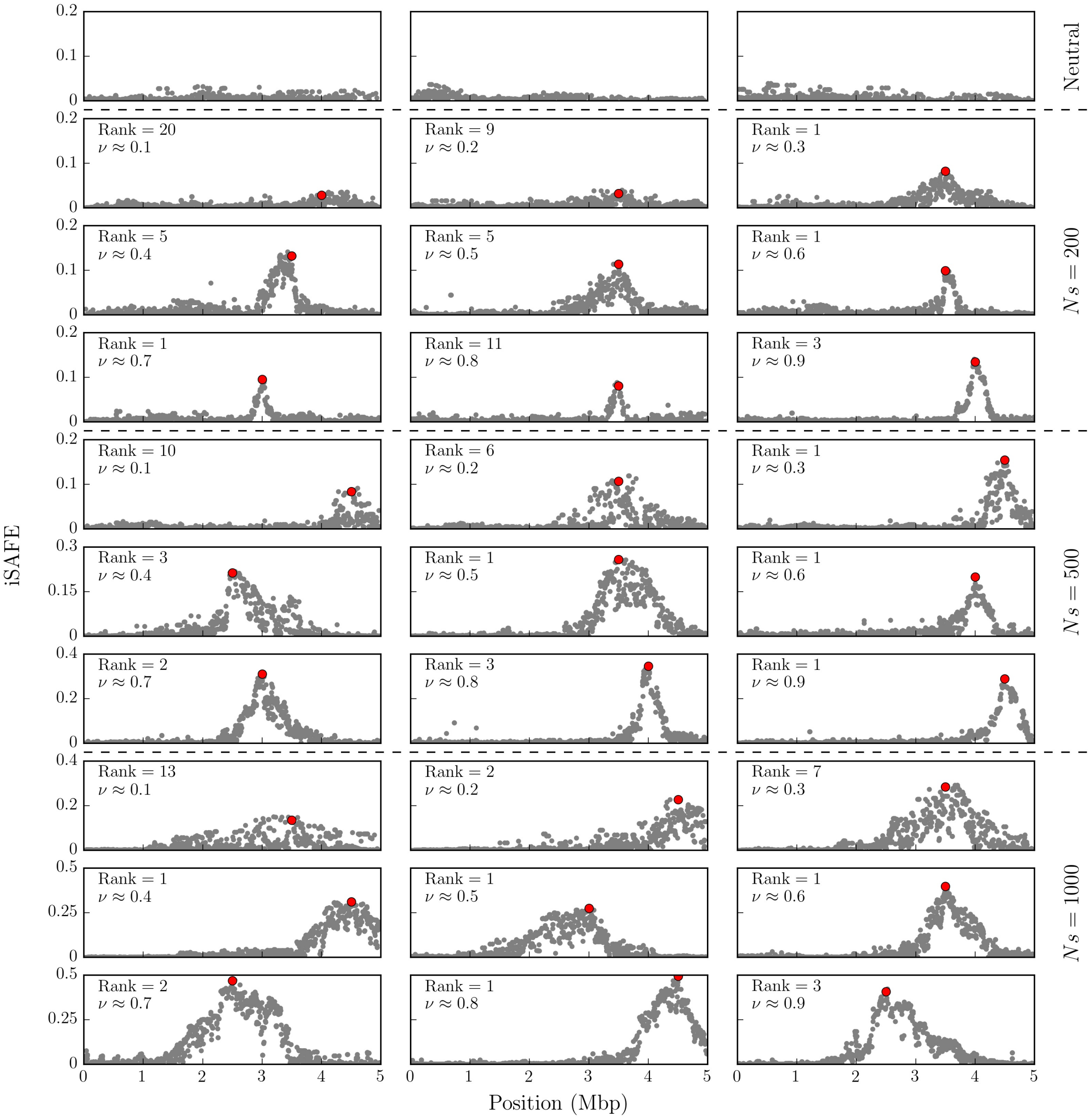


Figure C.7: **Demo II: iSAFE without outgroup samples.** iSAFE on ongoing hard selective sweeps ($\nu_0 = 1/N$) with different favored allele frequency (ν) in 5 Mbp region. The position of the favored mutation selected from range [2.5 Mbp, 5 Mbp]. Other simulation parameters are the default values for fixed population size (see Section B.1) and outgroup samples are not available.

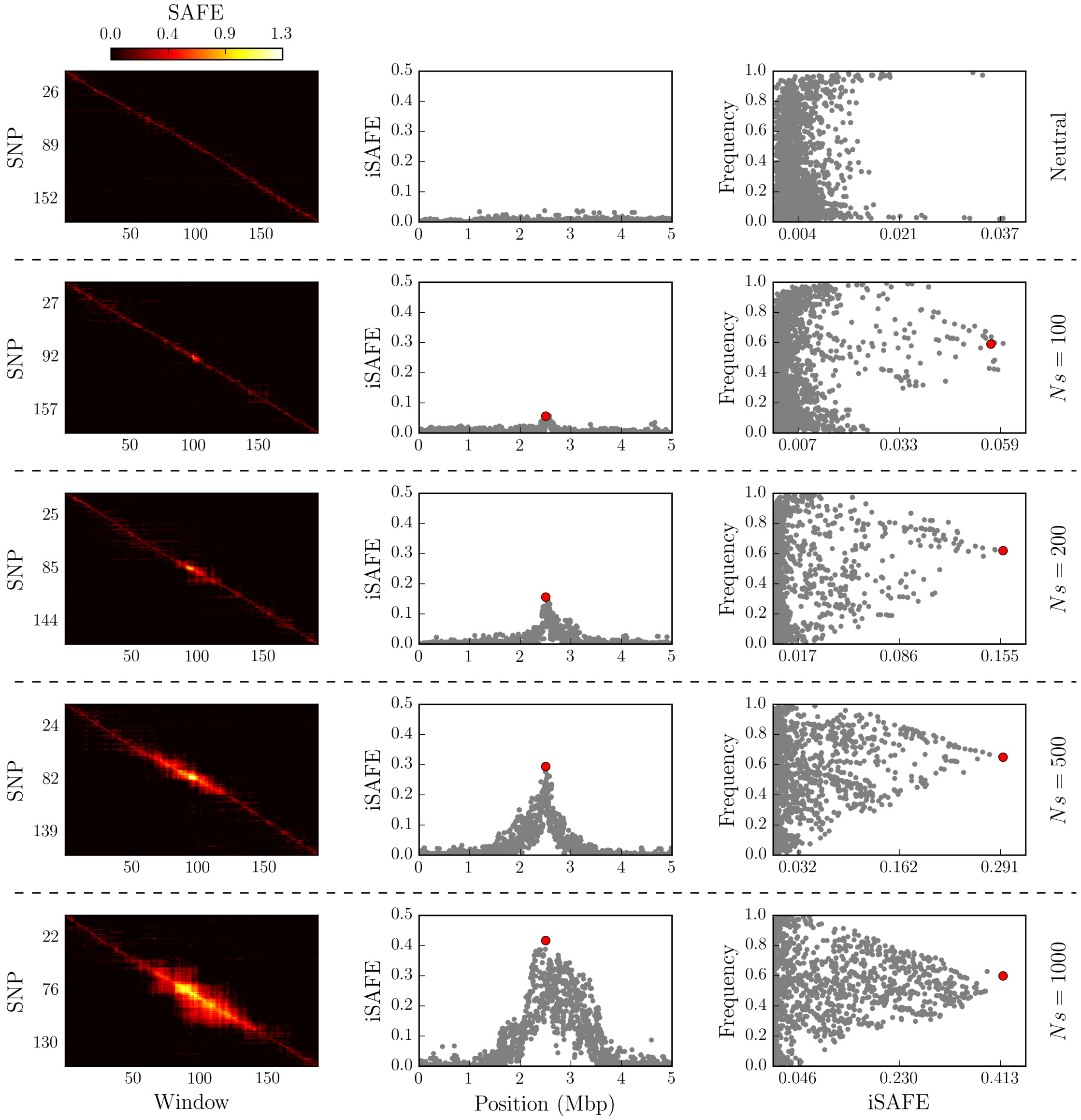


Figure C.8: **Demo III: iSAFE and selection strength.** iSAFE on 5 Mbp region with different selection strength, $Ns \in [0, 100, 200, 500, 1000]$. Left panels shows the $\Psi_{e,w}$ matrix. Middle panel shows the iSAFE score as a function of the variant position. Right panel show the derived allele frequency as a function of iSAFE score.

Bibliography

[1] Ronen, R., Tesler, G., Akbari, A., Zakov, S., Rosenberg, N. A. & Bafna, V. Predicting carriers of ongoing selective sweeps without knowledge of the favored allele. *PLoS Genet* **11**, e1005527 (2015).

[2] Akbari, A., Vitti, J. J., Iranmehr, A., Bakhtiari, M., Sabeti, P. C., Mirarab, S. & Bafna, V. Identifying the favored mutation in a positive selective sweep. *Nature Methods* **15**, 279–282 (2018).

[3] Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A. & Deloukas, P. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* **39**, 31–40 (2007).

[4] Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).

[5] Poznik, G. D., Henn, B. M., Yee, M. C., Sliwerska, E., Euskirchen, G. M., Lin, A. A., Snyder, M., Quintana-Murci, L., Kidd, J. M., Underhill, P. A. & Bustamante, C. D. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562–565 (2013).

[6] Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N. & Reich, D. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354 (2014).

[7] Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493 (2011).

[8] Fan, S., Hansen, M. E., Lo, Y. & Tishkoff, S. A. Going global by adapting local: A review of recent human adaptation. *Science* **354**, 54–59 (2016).

[9] Beall, C. M. Adaptation to high altitude: phenotypes and genotypes. *Annual Review of Anthropology* **43**, 251–272 (2014).

[10] Stobdan, T., Akbari, A., Azad, P., Zhou, D., Poulsen, O., Appenzeller, O., Gonzales,

G. F., Telenti, A., Wong, E. H. M., Saini, S., Kirkness, E. F., Venter, J. C., Bafna, V. & Haddad, G. G. New Insights into the Genetic Basis of Monges Disease and Adaptation to High-Altitude. *Molecular biology and evolution* **34**, 3154–3168 (2017).

[11] Azad, P., Stobdan, T., Zhou, D., Hartley, I., Akbari, A., Bafna, V. & Haddad, G. G. High-altitude adaptation in humans: From genomics to integrative physiology. *Journal of Molecular Medicine* **95**, 1269–1282 (2017).

[12] Udpa, N., Ronen, R., Zhou, D., Liang, J., Stobdan, T., Appenzeller, O., Yin, Y., Du, Y., Guo, L., Cao, R., Wang, Y., Jin, X., Huang, C., Jia, W., Cao, D., Guo, G., Claydon, V. E., Hainsworth, R., Gamboa, J. L., Zibenigus, M., Zenebe, G., Xue, J., Liu, S., Frazer, K. A., Li, Y., Bafna, V. & Haddad, G. G. Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. *Genome Biol.* **15**, R36 (2014).

[13] Iranmehr, A., Stobdan, T., Zhou, D., Poulsen, O., Strohl, K. P., Aldashev, A., Telenti, A., Wong, E. H., Kirkness, E. F., Venter, J. C., Bafna, V. & Haddad, G. G. Novel insight into the genetic basis of high-altitude pulmonary hypertension in Kyrgyz highlanders. *European Journal of Human Genetics* (2018).

[14] Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., Wang, W., Zhang, X., Xu, X., Yang, H., Li, Y., Wang, J., Wang, J. & Nielsen, R. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194 (2014).

[15] Kimura, M. *The neutral theory of molecular evolution* (Cambridge University Press, 1983).

[16] Nordborg, M. Coalescent Theory. In Balding, D. J., Bishop, M. & Cannings, C. (eds.) *Handbook of statistical genetics*, chap. 25, 843–877 (John Wiley & Sons, Ltd, 2008), third edn.

[17] Kingman, J. F. C. The coalescent. *Stochastic processes and their applications* **13**, 235–248 (1982).

[18] Kingman, J. F. C. On the genealogy of large populations. *Journal of Applied Probability* **19**, 27–43 (1982).

[19] Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).

[20] Slatkin, M. Linkage disequilibriumunderstanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**, 477 (2008).

[21] Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).

[22] Kaplan, N. L., Hudson, R. R. & Langley, C. H. The "hitchhiking effect" revisited. *Genetics* **123**, 887–899 (1989).

[23] Kim, Y. & Stephan, W. Selective sweeps in the presence of interference among partially linked loci. *Genetics* **164**, 389–398 (2003).

[24] Hermisson, J. & Pennings, P. S. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335–2352 (2005).

[25] Pennings, P. S. & Hermisson, J. Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.* **23**, 1076–1084 (2006).

[26] Pennings, P. S. & Hermisson, J. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS genetics* **2**, e186 (2006).

[27] Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol. (Amst.)* **28**, 659–669 (2013).

[28] Schrider, D. R., Mendes, F. K., Hahn, M. W. & Kern, A. D. Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics* **200**, 267–284 (2015).

[29] Schrider, D. R. & Kern, A. D. Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular biology and evolution* **34**, 1863–1877 (2017).

[30] Ferrer-Admetlla, A., Liang, M., Korneliussen, T. & Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular biology and evolution* **31**, 1275–1291 (2014).

[31] Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS genetics* **11**, e1005004 (2015).

[32] 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).

[33] Wilson, B. A., Petrov, D. A. & Messer, P. W. Soft selective sweeps in complex demographic scenarios. *Genetics* **198**, 669–684 (2014).

[34] Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., De La Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson,

P. L., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. & Pääbo, S. A draft sequence of the Neandertal genome. *science* **328**, 710–722 (2010).

[35] Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M. V., Derevianko, A. P., Hublin, J. J., Kelso, J., Slatkin, M. & Pääbo, S. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053 (2010).

[36] Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., De Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., Briggs, A. W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M. F., Shunkov, M. V., Derevianko, A. P., Patterson, N., Andrés, A. M., Eichler, E. E., Slatkin, M., Reich, D., Kelso, J. & Pääbo, S. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).

[37] Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J. C., Vohr, S. H., Green, R. E., Hellmann, I., Johnson, P. L., Blanche, H., Cann, H., Kitzman, J. O., Shendure, J., Eichler, E. E., Lein, E. S., Bakken, T. E., Golovanova, L. V., Doronichev, V. B., Shunkov, M. V., Derevianko, A. P., Viola, B., Slatkin, M., Reich, D., Kelso, J. & Pääbo, S. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43 (2014).

[38] Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., Viola, B., Prüfer, K., Meyer, M., Kelso, J., Reich, D. & Pääbo, S. An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216 (2015).

[39] Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics* **16**, 359 (2015).

[40] Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current Biology* **26**, 1241–1247 (2016).

[41] Racimo, F., Marnetto, D. & Huerta-Sanchez, E. Signatures of archaic adaptive introgression in present-day human populations. *Molecular biology and evolution* **34**, 296–317 (2016).

[42] Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).

[43] Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413 (2000).

[44] Pavlidis, P., Živković, D., Stamatakis, A. & Alachiotis, N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular biology and evolution* **30**, 2224–2234 (2013).

[45] Ronen, R., Udpa, N., Halperin, E. & Bafna, V. Learning natural selection from the site frequency spectrum. *Genetics* **195**, 181–193 (2013).

[46] DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I. & Nielsen, R. SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics* **32**, 1895–1897 (2016).

[47] Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72 (2006).

[48] Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. & Lander, E. S. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832 (2002).

[49] Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

[50] Nielsen, R. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005).

[51] Kim, Y. & Nielsen, R. Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513–1524 (2004).

[52] Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., Akey, J. M. & Jones, K. W. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human genomics* **1**, 274 (2004).

[53] Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G. & Bustamante, C. Genomic scans for selective sweeps using SNP data. *Genome research* **15**, 1566–1575 (2005).

[54] Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).

[55] Haasl, R. J. & Payseur, B. A. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Molecular ecology* **25**, 5–23 (2016).

[56] Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G.,

Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E. S., Schaffner, S. F. & Sabeti, P. C. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science* **327**, 883–886 (2010).

[57] Schrider, D. R. & Kern, A. D. S/HIC: robust identification of soft and hard sweeps using machine learning. *PLoS genetics* **12**, e1005928 (2016).

[58] Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I. & Pritchard, J. K. Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).

[59] Wang, M., Huang, X., Li, R., Xu, H., Jin, L. & He, Y. Detecting recent positive selection with high accuracy and reliability by conditional coalescent tree. *Molecular biology and evolution* **31**, 3068–3080 (2014).

[60] Pybus, M., Luisi, P., Dall'Olio, G. M., Uzkudun, M., Laayouni, H., Bertranpetit, J. & Engelken, J. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics* **31**, 3946–3952 (2015).

[61] Fu, W. & Akey, J. M. Selection and adaptation in the human genome. *Annu Rev Genomics Hum Genet* **14**, 467–489 (2013).

[62] Lachance, J. & Tishkoff, S. A. Population Genomics of Human Adaptation. *Annu Rev Ecol Evol Syst* **44**, 123–143 (2013).

[63] Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting natural selection in genomic data. *Annual review of genetics* **47**, 97–120 (2013).

[64] Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W. & Pritchard, J. K. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009).

[65] Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10**, e1004412 (2014).

[66] Jeong, C. & Di Rienzo, A. Adaptations to local environments in modern human populations. *Curr. Opin. Genet. Dev.* **29C**, 1–8 (2014).

[67] Tekola-Ayele, F., Adeyemo, A., Chen, G., Hailu, E., Aseffa, A., Davey, G., Newport, M. J. & Rotimi, C. N. Novel genomic signals of recent selection in an Ethiopian population. *European Journal of Human Genetics* **23**, 1085 (2015).

[68] Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., Zou, J., Shan, Y., Li, S., Yang, Q., Asan, A., Ni, P., Tian,

G., Xu, J., Liu, X., Jiang, T., Wu, R., Zhou, G., Tang, M., Qin, J., Wang, T., Feng, S., Li, G., Huasang, H., Luosang, J., Wang, W., Chen, F., Wang, Y., Zheng, X., Li, Z., Bianba, Z., Yang, G., Wang, X., Tang, S., Gao, G., Chen, Y., Luo, Z., Gusang, L., Cao, Z., Zhang, Q., Ouyang, W., Ren, X., Liang, H., Zheng, H., Huang, Y., Li, J., Bolund, L., Kristiansen, K., Li, Y., Zhang, Y., Zhang, X., Li, R., Li, S., Yang, H., Nielsen, R., Wang, J. & Wang, J. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* **329**, 75–78 (2010).

[69] Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., Prchal, J. T. & Ge, R. Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**, 72–75 (2010).

[70] Scheinfeldt, L. B., Soi, S., Thompson, S., Ranciaro, A., Woldemeskel, D., Beggs, W., Lambert, C., Jarvis, J. P., Abate, D., Belay, G. & Tishkoff, S. A. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biol.* **13**, R1 (2012).

[71] Alkorta-Aranburu, G., Beall, C. M., Witonsky, D. B., Gebremedhin, A., Pritchard, J. K. & Di Rienzo, A. The genetic architecture of adaptations to high altitude in Ethiopia. *PLoS Genet.* **8**, e1003110 (2012).

[72] Huerta-Sanchez, E., Degiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T., Cardona, A., Montgomery, H. E., Cavalleri, G. L., Robbins, P. A., Weale, M. E., Bradman, N., Bekele, E., Kivisild, T., Tyler-Smith, C. & Nielsen, R. Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. *Mol. Biol. Evol.* **30**, 1877–1888 (2013).

[73] Zhou, D., Udpa, N., Ronen, R., Stobdan, T., Liang, J., Appenzeller, O., Zhao, H. W., Yin, Y., Du, Y., Guo, L., Cao, R., Wang, Y., Jin, X., Huang, C., Jia, W., Cao, D., Guo, G., Gamboa, J. L., Villafuerte, F., Callacondo, D., Xue, J., Liu, S., Frazer, K. A., Li, Y., Bafna, V. & Haddad, G. G. Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in Andean highlanders. *Am. J. Hum. Genet.* **93**, 452–462 (2013).

[74] Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J. & Ayala, F. J. Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster. *Genetics* **136**, 1329–1340 (1994).

[75] Depaulis, F., Mousset, S. & Veuille, M. Haplotype tests using coalescent simulations conditional on the number of segregating sites. *Mol. Biol. Evol.* **18**, 1136–1138 (2001).

[76] Innan, H., Zhang, K., Marjoram, P., Tavare, S. & Rosenberg, N. A. Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics* **169**, 1763–1777 (2005).

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

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

[79] Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., Hollfelder, N., Potekhina, I. D., Schier, W., Thomas, M. G. & Burger, J. 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).

[80] Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W. & Pritchard, J. K. The role of geography in human adaptation. *PLoS Genet* **5**, e1000500 (2009).

[81] Campbell, C. D., Chong, J. X., Malig, M., Ko, A., Dumont, B. L., Han, L., Vives, L., O'Roak, B. J., Sudmant, P. H., Shendure, J., Abney, M., Ober, C. & Eichler, E. E. Estimating the human mutation rate using autozygosity in a founder population. *Nature genetics* **44**, 1277–1281 (2012).

[82] 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).

[83] Beleza, S., Johnson, N. A., Candille, S. I., Absher, D. M., Coram, M. A., Lopes, J., Campos, J., Araújo, I. I., Anderson, T. M., Vilhjálmsson, B. J., Nordborg, M., Correia e Silva, A., Shriver, M. D., Rocha, J., Barsh, G. S. & Tang, H. Genetic Architecture of Skin and Eye Color in an African-European Admixed Population. *PLoS Genetics* **9** (2013).

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

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

[86] Fujimoto, A., Ohashi, J., Nishida, N., Miyagawa, T., Morishita, Y., Tsunoda, T., Kimura, R. & Tokunaga, K. 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).

[87] Bryk, J., Hardouin, E., Pugach, I., Hughes, D., Strotmann, R., Stoneking, M. & Myles, S. Positive selection in East Asians for an EDAR allele that enhances NF-κB activation. *PLoS One* **3**, e2209 (2008).

[88] Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L. & Järvelä, I. Identification of a variant associated with adult-type hypolactasia. *Nature genetics* **30**, 233–237 (2002).

[89] 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).

[90] Heffelfinger, C., Pakstis, A. J., Speed, W. C., Clark, A. P., Haigh, E., Fang, R., Furtado, M. R., Kidd, K. K. & Snyder, M. P. Haplotype structure and positive selection at TLR1. *European Journal of Human Genetics* **22**, 551–557 (2014).

[91] Wong, S. H., Gochhait, S., Malhotra, D., Pettersson, F. H., Teo, Y. Y., Khor, C. C., Rautanen, A., Chapman, S. J., Mills, T. C., Srivastava, A., Rudko, A., Freidin, M. B., Puzyrev, V. P., Ali, S., Aggarwal, S., Chopra, R., Reddy, B. S. N., Garg, V. K., Roy, S., Meisner, S., Hazra, S. K., Saha, B., Floyd, S., Keating, B. J., Kim, C., Fairfax, B. P., Knight, J. C., Hill, P. C., Adegbola, R. A., Hakonarson, H., Fine, P. E. M., Pitchappan, R. M., Bamezai, R. N. K., Hill, A. V. S. & Vannberg, F. O. Leprosy and the adaptation of human toll-like receptor 1. *PLoS Pathog* **6**, e1000979 (2010).

[92] McManus, K. F., Taravella, A. M., Henn, B. M., Bustamante, C. D., Sikora, M. & Cornejo, O. E. Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. *PLoS genetics* **13**, e1006560 (2017).

[93] 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).

[94] 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).

[95] Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T., Altshuler, D. & Lander, E. Positive natural selection in the human lineage. *science* **312**, 1614–1620 (2006).

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

[97] Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., Piro, A., Stoneking, M., Tagarelli, A., Tagarelli, G., Touma, E. H., Williams, S. M. & Clark, A. G. Haplotype diversity and linkage disequilibrium at human G6PD: Recent origin of alleles that confer malarial resistance. *Science* **293**, 455–462 (2001).

[98] Hu, C.-J., Wang, L.-Y., Chodosh, L. A., Keith, B. & Simon, M. C. Differential roles of hypoxia-inducible factor 1α (HIF-1α) and HIF-2α in hypoxic gene regulation. *Molecular and cellular biology* **23**, 9361–9374 (2003).

[99] Seguin-Orlando, A., Korneliussen, T. S., Sikora, M., Malaspinas, A. S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., Goebel, T., Westaway,

M., Lambert, D., Khartanovich, V., Wall, J. D., Nigst, P. R., Foley, R. A., Lahr, M. M., Nielsen, R., Orlando, L. & Willerslev, E. Genomic structure in Europeans dating back at least 36,200 years. *Science* **346**, 1113–1118 (2014).

[100] Durvasula, A. & Sankararaman, S. Recovering signals of ghost archaic admixture in the genomes of present-day Africans. *bioRxiv* 285734 (2018).

[101] Wolf, A. B. & Akey, J. M. Outstanding questions in the study of archaic hominin admixture. *PLoS genetics* **14**, e1007349 (2018).

[102] Lachance, J., Vernot, B., Elbers, C. C., Ferwerda, B., Froment, A., Bodo, J. M., Lema, G., Fu, W., Nyambo, T. B., Rebbeck, T. R., Zhang, K., Akey, J. M. & Tishkoff, S. A. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457–469 (2012).

[103] Hsieh, P., Woerner, A. E., Wall, J. D., Lachance, J., Tishkoff, S. A., Gutenkunst, R. N. & Hammer, M. F. Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. *Genome research* (2016).

[104] Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* 1245938 (2014).

[105] Vernot, B., Tucci, S., Kelso, J., Schraiber, J. G., Wolf, A. B., Gittelman, R. M., Dannemann, M., Grote, S., McCoy, R. C., Norton, H., Scheinfeldt, L. B., Merriwether, D. A., Koki, G., Friedlaender, J. S., Wakefield, J., Pääbo, S. & Akey, J. M. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* aad9416 (2016).

[106] Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell* **173**, 53–61 (2018).

[107] Zhu, J. & Hastie, T. Kernel logistic regression and the import vector machine. In *Advances in neural information processing systems*, 1081–1088 (2002).

[108] Hastie, T., Friedman, J. & Tibshirani, R. Linear methods for classification. In *The Elements of Statistical Learning*, 79–113 (Springer, 2001).

[109] Hudson, R. R. The variance of coalescent time estimates from DNA sequences. *Journal of molecular evolution* **64**, 702–705 (2007).

[110] Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. & Feldman, M. W. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proceedings of the National Academy of Sciences* **97**, 7360–7365 (2000).

[111] Lu, D., Lou, H., Yuan, K., Wang, X., Wang, Y., Zhang, C., Lu, Y., Yang, X., Deng, L., Zhou, Y., Feng, Q., Hu, Y., Ding, Q., Yang, Y., Li, S., Jin, L., Guan, Y., Su, B., Kang, L. &

Xu, S. Ancestral origins and genetic history of Tibetan highlanders. *The American Journal of Human Genetics* **99**, 580–594 (2016).

[112] Donnelly, M. P., Paschou, P., Grigorenko, E., Gurwitz, D., Barta, C., Lu, R. B., Zhukova, O. V., Kim, J. J., Siniscalco, M., New, M., Li, H., Kajuna, S. L., Manolopoulos, V. G., Speed, W. C., Pakstis, A. J., Kidd, J. R. & Kidd, K. K. A global view of the OCA2-HERC2 region and pigmentation. *Human Genetics* **131**, 683–696 (2012).

[113] Miller, C. T., Beleza, S., Pollen, A. A., Schluter, D., Kittles, R. A., Shriver, M. D. & Kingsley, D. M. cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell* **131**, 1179–1189 (2007).

[114] Suzuki, Y., Pasch, A., Bonny, O., Mohaupt, M. G., Hediger, M. A. & Frey, F. J. 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).

[115] Park, B. L., Kim, J. W., Cheong, H. S., Kim, L. H., Lee, B. C., Seo, C. H., Kang, T. C., Nam, Y. W., Kim, G. B., Shin, H. D. & Choi, I. G. Extended genetic effects of ADH cluster genes on the risk of alcohol dependence: From GWAS to replication. *Human Genetics* **132**, 657–668 (2013).

[116] Oze, I., Matsuo, K., Suzuki, T., Kawase, T., Watanabe, M., Hiraki, A., Ito, H., Hosono, S., Ozawa, T., Hatooka, S., Yatabe, Y., Hasegawa, Y., Shinoda, M., Kiura, K., Tajima, K., Tanimoto, M. & Tanaka, H. Impact of multiple alcohol dehydrogenase gene polymorphisms on risk of upper aerodigestive tract cancers in a Japanese population. *Cancer Epidemiology Biomarkers and Prevention* **18**, 3097–3102 (2009).

[117] Ehret, G. B. *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).

[118] Bailey, J. N., Loomis, S. J., Kang, J. H., Allingham, R. R., Gharahkhani, P., Khor, C. C., Burdon, K. P., Aschard, H., Chasman, D. I., Igo, R. P., Hysi, P. G., Glastonbury, C. A., Ashley-Koch, A., Brilliant, M., Brown, A. A., Budenz, D. L., Buil, A., Cheng, C. Y., Choi, H., Christen, W. G., Curhan, G., De Vivo, I., Fingert, J. H., Foster, P. J., Fuchs, C., Gaasterland, D., Gaasterland, T., Hewitt, A. W., Hu, F., Hunter, D. J., Khawaja, A. P., Lee, R. K., Li, Z., Lichter, P. R., Mackey, D. A., McGuffin, P., Mitchell, P., Moroi, S. E., Perera, S. A., Pepper, K. W., Qi, Q., Realini, T., Richards, J. E., Ridker, P. M., Rimm, E., Ritch, R., Ritchie, M., Schuman, J. S., Scott, W. K., Singh, K., Sit, A. J., Song, Y. E., Tamimi, R. M., Topouzis, F., Viswanathan, A. C., Verma, S. S., Vollrath, D., Wang, J. J., Weisschuh, N., Wissinger, B., Wollstein, G., Wong, T. Y., Yaspan, B. L., Zack, D. J., Zhang, K., Weinreb, R. N., Pericak-Vance, M. A., Small, K., Hammond, C. J., Aung, T., Liu, Y., Vithana, E. N., MacGregor, S., Craig, J. E., Kraft, P., Howell, G., Hauser, M. A., Pasquale, L. R., Haines, J. L. & Wiggs, J. L. Genome-wide association analysis identifies TXNRD2, ATXN2 and FOXC1 as susceptibility loci for primary open-angle glaucoma. *Nature genetics* (2016).

[119] 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).

[120] Wu, X., Ye, Y., Kiemeney, L. A., Sulem, P., Rafnar, T., Matullo, G., Seminara, D., Yoshida, T., Saeki, N., Andrew, A. S., Dinney, C. P., Czerniak, B., Zhang, Z. F., Kiltie, A. E., Bishop, D. T., Vineis, P., Porru, S., Buntinx, F., Kellen, E., Zeegers, M. P., Kumar, R., Rudnai, P., Gurzau, E., Koppova, K., Mayordomo, J. I., Sanchez, M., Saez, B., Lindblom, A., De Verdier, P., Steineck, G., Mills, G. B., Schned, A., Guarrera, S., Polidoro, S., Chang, S. C., Lin, J., Chang, D. W., Hale, K. S., Majewski, T., Grossman, H. B., Thorlacius, S., Thorsteinsdottir, U., Aben, K. K., Witjes, J. A., Stefansson, K., Amos, C. I., Karagas, M. R. & Gu, J. Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nature genetics* **41**, 991–995 (2009).

[121] Sakamoto, H., Yoshimura, K., Saeki, N., Katai, H., Shimoda, T., Matsuno, Y., Saito, D., Sugimura, H., Tanioka, F., Kato, S., Matsukura, N., Matsuda, N., Nakamura, T., Hyodo, I., Nishina, T., Yasui, W., Hirose, H., Hayashi, M., Toshiro, E., Ohnami, S., Sekine, A., Sato, Y., Totsuka, H., Ando, M., Takemura, R., Takahashi, Y., Ohdaira, M., Aoki, K., Honmyo, I., Chiku, S., Aoyagi, K., Sasaki, H., Ohnami, S., Yanagihara, K., Yoon, K. A., Kook, M. C., Lee, Y. S., Park, S. R., Kim, C. G., Choi, I. J., Yoshida, T., Nakamura, Y. & Hirohashi, S. Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nature genetics* **40**, 730–740 (2008).

[122] Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., Glazer, N. L., Morrison, A. C., Johnson, A. D., Aspelund, T., Aulchenko, Y., Lumley, T., Köttgen, A., Vasan, R. S., Rivadeneira, F., Eiriksdottir, G., Guo, X., Arking, D. E., Mitchell, G. F., Mattace-Raso, F. U., Smith, A. V., Taylor, K., Scharpf, R. B., Hwang, S. J., Sijbrands, E. J., Bis, J., Harris, T. B., Ganesh, S. K., O'Donnell, C. J., Hofman, A., Rotter, J. I., Coresh, J., Benjamin, E. J., Uitterlinden, A. G., Heiss, G., Fox, C. S., Witteman, J. C., Boerwinkle, E., Wang, T. J., Gudnason, V., Larson, M. G., Chakravarti, A., Psaty, B. M. & Van Duijn, C. M. Genome-wide association study of blood pressure and hypertension. *Nature genetics* **41**, 677–687 (2009).

[123] Bentham, J., Morris, D. L., Cunninghame Graham, D. S., Pinder, C. L., Tombleson, P., Behrens, T. W., Martín, J., Fairfax, B. P., Knight, J. C., Chen, L., Replogle, J., Syvänen, A.-C., Rönnblom, L., Graham, R. R., Wither, J. E., Rioux, J. D., Alarcón-Riquelme, M. E. & Vyse, T. J. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics* **47**, 1457–1464 (2015).

[124] Jin, Y., Birlea, S. A., Fain, P. R., Ferrara, T. M., Ben, S., Riccardi, S. L., Cole, J. B., Gowan, K., Holland, P. J., Bennett, D. C., Luiten, R. M., Wolkerstorfer, A., Van Der Veen, J. P., Hartmann, A., Eichner, S., Schuler, G., Van Geel, N., Lambert, J., Kemp, E. H., Gawkrodger, D. J., Weetman, A. P., Taïeb, A., Jouary, T., Ezzedine, K., Wallace, M. R., McCormack, W. T., Picardo, M., Leone, G., Overbeck, A., Silverberg, N. B. & Spritz,

R. A. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nature genetics* **44**, 676–680 (2012).

[125] 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).

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

[127] Cordell, H. J., Töpf, A., Mamasoula, C., Postma, A. V., Bentham, J., Zelenika, D., Heath, S., Blue, G., Cosgrove, C., Granados riveron, J., Darlay, R., Soemedi, R., Wilson, I. J., Ayers, K. L., Rahman, T. J., Hall, D., Mulder, B. J., Zwinderman, A. H., Van engelen, K., Brook, J. D., Setchfield, K., Bu'lock, F. A., Thornborough, C., O'sullivan, J., Stuart, A. G., Parsons, J., Bhattacharya, S., Winlaw, D., Mital, S., Gewillig, M., Breckpot, J., Devriendt, K., Moorman, A. F., Rauch, A., Lathrop, G. M., Keavney, B. D. & Goodship, J. A. Genome-wide association study identifies loci on 12q24 and 13q32 associated with tetralogy of Fallot. *Human molecular genetics* dds552 (2013).

[128] Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R. R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J. C., Carroll, R. J., Eyler, A. E., Greenberg, J. D., Kremer, J. M., Pappas, D. A., Jiang, L., Yin, J., Ye, L., Su, D. F., Yang, J., Xie, G., Keystone, E., Westra, H. J., Esko, T., Metspalu, A., Zhou, X., Gupta, N., Mirel, D., Stahl, E. A., Diogo, D., Cui, J., Liao, K., Guo, M. H., Myouzen, K., Kawaguchi, T., Coenen, M. J., Van Riel, P. L., Van De Laar, M. A., Guchelaar, H. J., Huizinga, T. W., Dieudé, P., Mariette, X., Bridges, S. L., Zhernakova, A., Toes, R. E., Tak, P. P., Miceli-Richard, C., Bang, S. Y., Lee, H. S., Martin, J., Gonzalez-Gay, M. A., Rodriguez-Rodriguez, L., Rantapää-Dahlqvist, S., Ärlestig, L., Choi, H. K., Kamatani, Y., Galan, P., Lathrop, M., Eyre, S., Bowes, J., Barton, A., De Vries, N., Moreland, L. W., Criswell, L. A., Karlson, E. W., Taniguchi, A., Yamada, R., Kubo, M., Liu, J. S., Bae, S. C., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P. K., Raychaudhuri, S., Stranger, B. E., De Jager, P. L., Franke, L., Visscher, P. M., Brown, M. A., Yamanaka, H., Mimori, T., Takahashi, A., Xu, H., Behrens, T. W., Siminovitch, K. A., Momohara, S., Matsuda, F., Yamamoto, K. & Plenge, R. M. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

[129] Dichgans, M., Malik, R., König, I. R., Rosand, J., Clarke, R., Gretarsdottir, S., Thorleifsson, G., Mitchell, B. D., Assimes, T. L., Levi, C., ODonnell, C. J., Fornage, M., Thorsteinsdottir, U., Psaty, B. M., Hengstenberg, C., Seshadri, S., Erdmann, J., Bis, J. C., Peters, A., Boncoraglio, G. B., März, W., Meschia, J. F., Kathiresan, S., Ikram, M. A., McPherson, R., Stefansson, K., Sudlow, C., Reilly, M. P., Thompson, J. R., Sharma, P., Hopewell, J. C., Chambers, J. C., Watkins, H., Rothwell, P. M., Roberts, R., Markus, H. S., Samani, N. J., Farrall, M. & Schunkert, H. Shared Genetic Susceptibility to Ischemic Stroke and Coronary Artery Disease. *Stroke* **45**, 24–36 (2014).

[130] Soranzo, N., Spector, T. D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., Salo, P., Voight, B. F., Burns, P., Laskowski, R. A., Xue, Y., Menzel, S., Altshuler, D., Bradley, J. R., Bumpstead, S., Burnett, M. S., Devaney, J., Döring, A., Elosua, R., Epstein, S. E., Erber, W., Falchi, M., Garner, S. F., Ghori, M. J., Goodall, A. H., Gwilliam, R., Hakonarson, H. H., Hall, A. S., Hammond, N., Hengstenberg, C., Illig, T., König, I. R., Knouff, C. W., McPherson, R., Melander, O., Mooser, V., Nauck, M., Nieminen, M. S., O'Donnell, C. J., Peltonen, L., Potter, S. C., Prokisch, H., Rader, D. J., Rice, C. M., Roberts, R., Salomaa, V., Sambrook, J., Schreiber, S., Schunkert, H., Schwartz, S. M., Serbanovic-Canic, J., Sinisalo, J., Siscovick, D. S., Stark, K., Surakka, I., Stephens, J., Thompson, J. R., Völker, U., Völzke, H., Watkins, N. A., Wells, G. A., Wichmann, H. E., Van Heel, D. A., Tyler-Smith, C., Thein, S. L., Kathiresan, S., Perola, M., Reilly, M. P., Stewart, A. F., Erdmann, J., Samani, N. J., Meisinger, C., Greinacher, A., Deloukas, P., Ouwehand, W. H. & Gieger, C. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genetics* **41**, 1182–1190 (2009).

[131] Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064–2065 (2010).

[132] Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).

[133] Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., Roskin, K. M., Chen, C.-F., Thomas, M. A., Haussler, D. & Jacob, H. J. Comparative recombination rates in the rat, mouse, and human genomes. *Genome research* **14**, 528–538 (2004).

[134] Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A. & Bustamante, C. D. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* **108**, 11983–11988 (2011).

[135] Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68 (2015).

[136] Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A. & Flicek, P. Ensembl 2018. *Nucleic acids research* **46**, D754–D761 (2017).

[137] Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851 (2007).

[138] Prüfer, K., De Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., Reher, D., Hopfe, C., Nagel, S., Maricic, T., Fu, Q., Theunert, C., Rogers, R., Skoglund, P., Chintalapati, M., Dannemann, M., Nelson, B. J., Key, F. M., Rudan, P., Kućan, Ž., Gušić, I., Golovanova, L. V., Doronichev, V. B., Patterson, N., Reich, D., Eichler, E. E., Slatkin, M., Schierup, M. H., Andrés, A. M., Kelso, J., Meyer, M. & Pääbo, S. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).