Insights into the adaptive history of African human populations from whole-genome sequence data

Sandra Walsh Capdevila

TESI DOCTORAL UPF / 2019

DIRECTORS DE LA TESI

Dr. Jaume Bertranpetit i Dr. Hafid Laayouni

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

upf. Universitat Pompeu Fabra Barcelona

Agraïments

Aquesta tesi no hagués estat possible sense l'enorme dedicació i esforç personal que la meva mare i el meu pare m'han dedicat al llarg dels meus 28 anys de vida. Gràcies per donar-me la oportunitat per créixer, per donar-me la llibertat i la confiança per desenvolupar-me sense prejudicis. Per a mi això té un valor incalculable, aquesta tesi va per vosaltres.

Abstract

Africa is the origin source of modern humans. Despite that African populations harbor the highest levels of genetic diversity worldwide, they remain underrepresented in genetic studies. Therefore, in order to fully understand modern human evolutionary history it is fundamental to include more African populations in genetic studies. The work in this thesis is a small contribution to the study of African evolutionary history. In particular we have focused on two different locations of Africa, eastern and southern Africa. We have tried to unravel candidates of positive (or adaptive) selection through the analysis of whole-genome sequences of five Ethiopian populations and one KhoeSan population. Moreover, we have tried to fill the gap between genotype and phenotype of a candidate of adaptive selection in an Ethiopian population.

Resum

Àfrica és la font d'origen dels humans moderns. Malgrat que les poblacions Africanes són les que contenen la major diversitat genètica al món, estan molt poc representades en estudis genètics. Així doncs, per poder plenament entendre la història evolutiva humana és fonamental incloure més poblacions Africanes en estudis genètics. Aquesta tesi és una petita contribució en l'estudi de la història evolutiva humana a l'Àfrica. Ens hem centrat en dos localitzacions diferents, a l'est i al sud de l'Àfrica. Hem intentat dilucidar les possibles senyals de selecció positiva (o adaptativa) a través de l'anàlisi de seqüències completes de genomes de cinc poblacions d'Etiòpia i una KhoeSan. A més a més, en l'última part de la tesi s'ha intentat entendre a nivell funcional la relació entre el genotip i el fenotip d'un candidat de selecció adaptativa descobert en una població d'Etiòpia.

Preface

The genomic era has brought a large number of complete genomes available which provides a comprehensive view of human genetic diversity and the opportunity to rigorously study human evolutionary history. Now that scientific community has the tools, there is a need to expand the sampling towards underrepresented populations from Africa. Only then a complete picture of human genetic diversity will be provided.

This thesis focuses on the study of African populations through whole-genome sequence analysis. In particular, the study of the detection, through computational methods, of the footprints of adaptive selection in the genomes of Ethiopian and KhoeSan populations. Moreover, a final study in this thesis tries to understand the link between genotype and phenotype that ultimately validates the effect of adaptive selection.

CONTENTS

I. INTRODUCTION

1. Africa, the cradle of Humankind

1.1. Archaeological and linguistic record

The archaeological, paleoanthropological and linguistic records provide relevant information to understand modern human origins and diversity. These fields of study have been crucial setting the grounds of our origins that later have been studied by human population genetics and human population genomics.

The oldest anatomically modern human (AMH) remains are found in Africa. Specifically in the Omo Kibish and Herto sites in Ethiopia where crania have been dated to 154,000-195,000 years ago (1,2) and are considered fully AMH (see Figure 1). Moreover, additional hominin remains have been found in Ngaloba (Tanzania) dating back to 120,000 years ago (3,4), indicating a putative East African AMH origin.

However, fossils with archaic and modern morphological features are found in other locations of Africa such as the south of Africa (Florisbad, South Africa) dated to 259,000 years ago (5) and from North Africa (Jebel Irhoud, Morocco) dated to 300,000 years ago (6). Hence, this morphological diversity between 300,000-100,000 years ago suggests a pan-African origin of *Homo sapiens* involving expansions, dispersals, gene flow and extinctions of several populations across the African continent. Given the evidence, this is a more plausible scenario than a simple linear evolution of morphology from an archaic to a modern *Homo sapiens* in just a specific place.

Additionally, it has been proposed that archaic admixture could have played a major role in the origin of our species (7–9). This hypothesis is supported by the discovery of quite recent (14,000 to 22,000 years ago) fossils with archaic and modern morphological features in Central Africa (10) and West Africa (11) and the genetic evidence of admixture between modern humans and archaic hominins (12–14). Specifically in Africa, there is evidence of admixture, in the order of 3-5%, with an early divergent and currently extinct ghost modern human lineage (15). In fact it is

extremely difficult to pinpoint these admixture events just considering morphology.

Unluckily, the archaeological record is scarce and there is a dearth of specimens especially in Central and West Africa. This is why the models of the emergence of *Homo sapiens* are continuously reshaped by new findings.

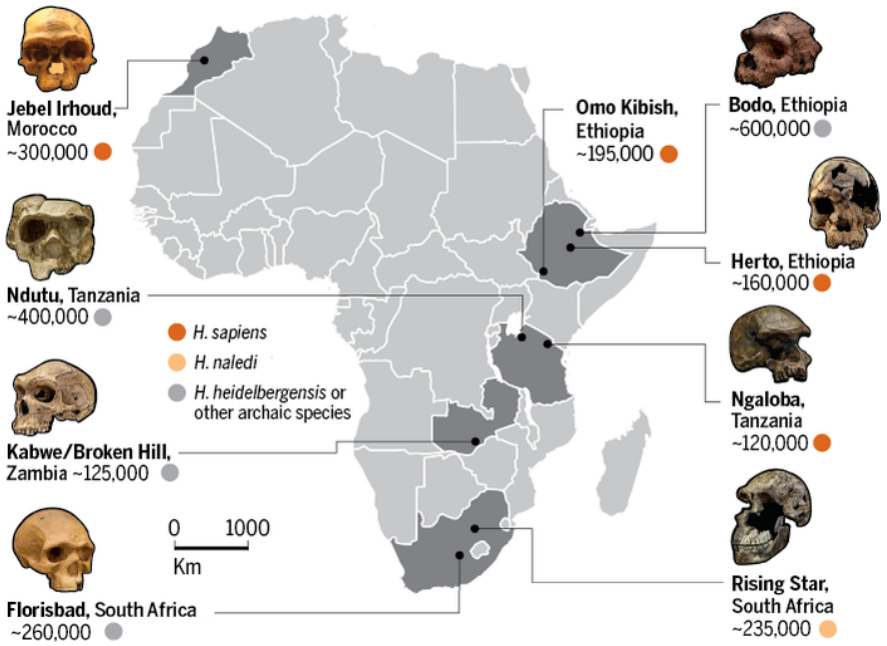


Figure 1: The human archaeological fossil record in Africa (16).

One of the most interesting and debated events in our evolutionary history is the Out of Africa (OoA), that is, the migration that spread modern humans initially to Eurasia and later to the whole world. Many open questions are still unanswered regarding the point of embarkation, the number and the migratory routes of the OoA. Until now, genetic evidence indicates that all living humans originated from a single OoA around 60,000 years ago (that will be discussed later in this chapter). But again, archaeological data suggests a more complicated scenario. In particular, the remains found in the Levant (Skhul, Qafzeh and Misliya Cave) dating to 120,000-177,000 years ago (17–19), and more recently the fossils from the Apidima cave (Greece) dated back to 200,000 years ago (20) and others that are

likely to be uncovered in the future, that could be indicators of another and earlier dispersal from Africa (see Figure 2). Additionally, presence of clearly modern humans pre-dating the 60,000 OoA has been recorded as far as in Southern China dated up to 120,000 years ago (21), Sumatra dated to 73,000 years ago (22), the Philippines dated back to 67,000 years ago (23) and even in Australia, dated back to 65,000 years (24). All these findings predating 60,000 years suggest that multiple dispersals of *Homo sapiens* outside Africa could have begun in the Late Pleistocene. Whether these dispersals resulted in the simple extinction of migrants, the ancestors of present day humans had contact or are directly descendants of these putative earlier dispersals is now being addressed in multiple genetic studies (discussed in later).

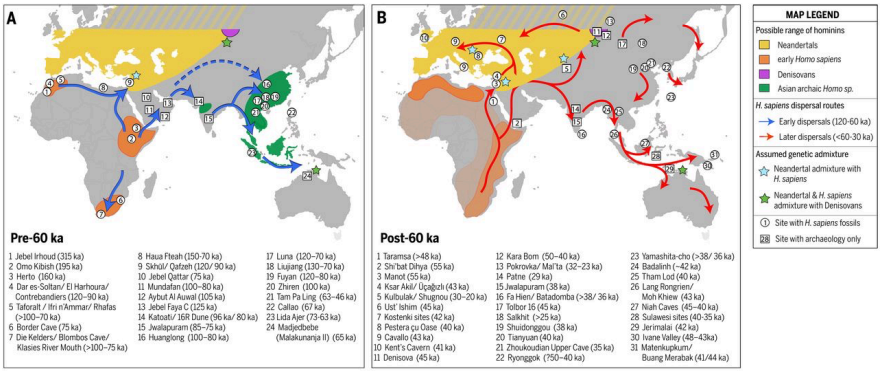


Figure 2: Distribution of key human archaeological sites with morphological modern humans and putative migratory routes **A**) Pre-dating 60,000 years **B**) Post 60,000 years (25).

Linguistics has also played a role in these discussions. Complex language is a singular human trait with a key role in our evolutionary history. It is thought that language enabled humans to acquire a more complex social organization and cooperation system (26), the ability of symbolic thought and abstract thinking (27). The work by Atkinson 2011 found the existence of a decline of phoneme diversity with increasing distance from Africa and pinpointed Africa as the origin source of all modern human languages. Although these conclusions should be taken carefully due to the limitations in the field, it is of special interest to notice the parallel mechanisms that shape genetic and linguistic diversity. In fact, among other factors, genetics sometimes correlate well with linguistic affiliation (29,30). An illustrative example is the

expansion of the Bantu languages from western to eastern and southern Africa that started 4,000 years ago, associated with a demic diffusion of people (31,32). But it is difficult to accept that linguistics may have a say in such ancient events like the OoA.

But there is a lot of interest to understand the distribution of spoken languages, their distribution in families and their relationship with past history of the populations. African languages constitute almost one third of the total worldwide languages, with more than 2000 ethnolinguistic groups and the presence of four main language families (Figure 3).

The Niger-Congo or Niger-Kordofanian language family is present throughout sub-Saharan Africa (sSA) and is the largest language family in the world, including a total of 1542 languages. The Bantu languages are the most represented subgroup of the Niger-Congo family comprising more than 500 languages. It is thought that the Bantu languages originated in the grasslands of Cameroon around 5000 years ago and its rapid and wide expansion throughout sSA during the last 4 millennia has brought interest in linguistics, archaeology and genetics.

The Afro-Asiatic language family (that includes the old family called Hamito-Semitic) comprises 377 languages divided in 6 subgroups (Berber, Chadic, Cushitic, Egyptian, Omotic and Semitic) and is predominantly spoken in north and eastern Africa and west Asia (33). Archaeological evidence suggests that the origin of Afro-Asiatic languages is in the Levant around 12,000 BP. Afro-Asiatic languages would have started to spread around 7000 BP through the Sinai Peninsula into Egypt and North Africa (Berber and Egyptian) and across the Red Sea into Sudan and Ethiopia (Chadic, Cushitic and Omotic). Much later, migrations from South Arabia (including the Arabic expansion) brought Semitic languages into the African continent (32).

The Nilo-Saharan language family is spoken in the greater Nile Basin and the central Sahara desert. There are around ten different groups of Nilo-Saharan languages although the classification is still under debate. The expansion of Nilo-Saharan languages is associated with cultures inhabiting the surroundings from the lake Turkana to the north of the Niger River around 10,000 BP (during

the wet or “Green Sahara” period with higher lake levels). The current geographic isolation of some Nilo-Saharan speakers might be explained by the shift to a dryer Sahara around 6000 BP (34).

The KhoeSan language family, denoted by the use of click-consonants, is mainly spoken in the south of Africa by a very reduced number of people. It is usually classified as a single language family but nowadays three different independent language families should be considered (Kx’a, Khoe-Kwadi and Tuu). Sandawe and Hadza, spoken in Tanzania, were also historically misclassified as KhoeSan because of the use of click consonants. Both are now considered language isolates (33,35).

Archaeological, fossil and linguistic records show the relevance of the African continent in the study of human history. This is why an interdisciplinary approach is crucial to define and test new hypothesis about the evolutionary history of our species. The following parts of this chapter will be focused on the current knowledge on African populations through population genetics.

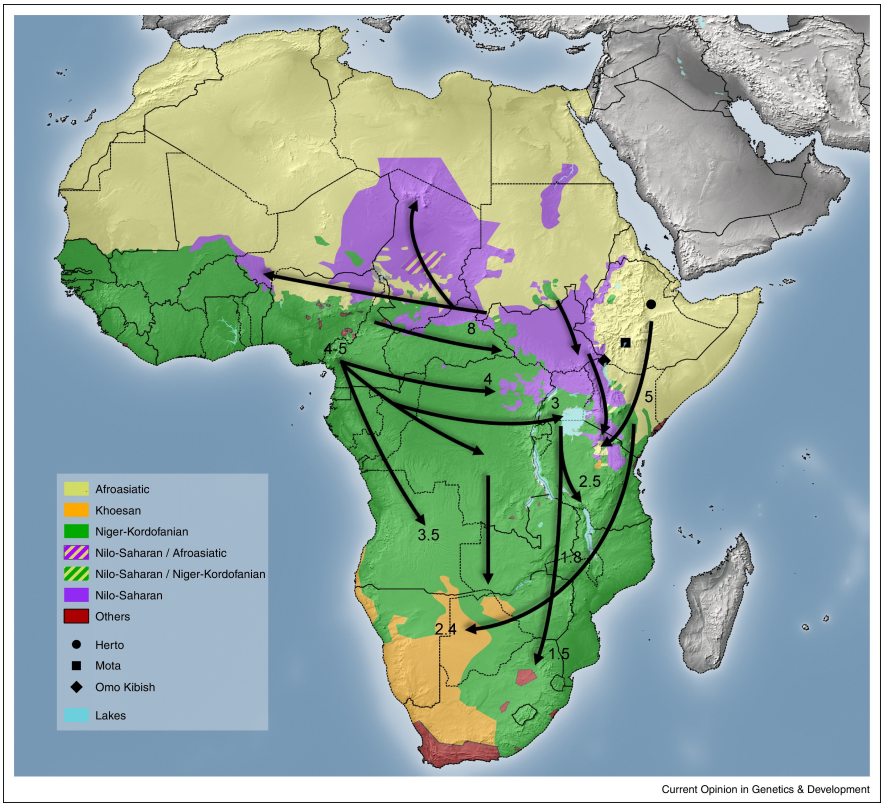


Figure 3: Distribution of the four main family languages in Africa and main migration routes of modern humans during the last 10000 years (36)**.**

1.2. Genetic history of Africa

Genetic studies have had an increasing important role in the exploration of human history during the last decades. One of the first scientists to use molecular genetics to unravel human history was Luca Cavalli-Sforza. In his book, *The History and Geography of Human Genes*, published in 1994, and thus before the advent of the molecular genetics revolution, he used a cross-disciplinary methodology including the historical record, archaeology, linguistics and added genetic data from classical genetic markers (e.g. blood group systems and other protein polymorphisms) to reconstruct human demographic history. At that time, even with the limitations of classical markers, he was already able to pinpoint Africa as the origin source of modern humans.

Shortly after, with the advent of the polymerase chain reaction (PCR) and sequencing technologies, uniparental markers (mitochondrial DNA and Y chromosome) started to be used for human population genetic inferences. Both mitochondrial DNA (mtDNA) and Y chromosome analysis supports an African modern human origin. Non-African populations carry only a subset of all the variation found in Africa and phylogenies also show that branches of the most ancestral lineages are found in Africa (37,38). In particular, the KhoeSan from southern Africa carry mtDNA haplogroups (L0d and L0k) with the deepest split among human populations (39). In addition, the haplogroups A and B of the non-recombining region of the Y chromosome, belong to the oldest branches and are also found in Africans (40). Despite the advantages of uniparental markers such as the easy reconstruction of trees because the absence of recombination, low price, and the abundance in the cell in the case of mtDNA, there are some limitations such as the strong influence of genetic drift, natural selection, sex-biased behaviours and the absence of recombination makes it not truly representative of the whole complexity of the autosomal genome (41).

Further development of sequencing, genotyping and next-generation sequencing (NGS) technologies together with powerful computation methods has provided with large sets of autosomal polymorphism data. This type of data has enabled researchers to test more complex models of human evolution. A great example of the use of genome-wide autosomal data is a seminal paper of African genetic history where 2432 African individuals from 121 African populations were genotyped (42). The authors found that Africans hold the highest levels of genetic diversity and that this diversity decreases with distance from Africa, mainly because the OoA founder effect. Moreover, they identified a complex population structure of 14 ancestral clusters with high levels of admixture reflecting migration events across the whole continent and a large and subdivided structure during the evolutionary history of these populations. Many subsequent studies also confirmed that the highest levels of genetic diversity are found in Africa, especially in hunter-gatherers from southern Africa (43), and an increased linkage disequilibrium with distance from Africa (44,45), pointing again to Africa as the birthplace of modern humans.

It is also known that one of the most significant events of African prehistory was the migration of sedentary populations with agricultural and herding practices from western Africa, the Bantu expansion. This dispersal of people began 4,000 BP in western Africa (Nigeria and Cameroon), first reaching the equatorial African rainforest (the "late-split" hypothesis) and from there splitting into two different waves that independently reached the south of Africa (31,46–48). The first wave represented by the ancestors of present-day Eastern Bantu speakers, reached East Africa around 3,000 BP and later on expanded to southern Africa around 1,300 BP. The second wave, represented by the ancestors of present-day Western Bantu speakers expanded through the Atlantic coast until the south of Africa (Figure 3). This expansion had a great demographic impact on the encountered indigenous hunter-gatherer populations that were probably replaced or assimilated and thus leaving the genomic footprint of the western African component that nowadays predominates throughout sSA (49,50).

The advent of ancient DNA (aDNA) sequencing has given the opportunity to sequence remains of ancient African modern humans, even if for most of Africa the climatic conditions are extremely poor for DNA preservation. But the cases in which it has been possible, it has given the opportunity to test more hypotheses and improve inferences about African prehistory using time-serial information. For instance, a recent study that sequenced 2,000 year old hunter-gatherers from southern African (ancestors of present day KhoeSan), captured the deepest split between human populations to have happened between 350,000 and 260,000 years ago (Figure 4) (51). Other cases will be discussed later.

In the next sections of this chapter we will mainly focus on the genetics of Eastern and Southern African populations, which are the main focus of this thesis.

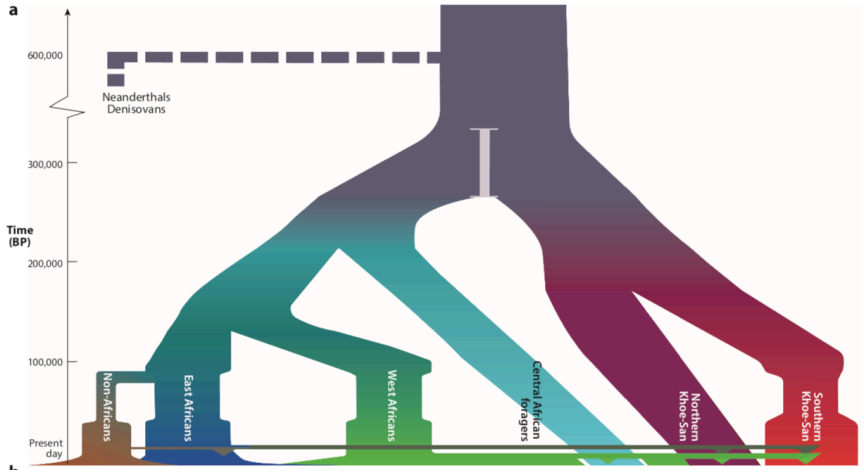


Figure 4: Simplified demographic model of modern human genetic history (52)

1.2.1. The genetics of East Africa

Eastern Africa is one of the key locations of human evolutionary history. Not only because its genetic and linguistic diversity, but also because two of the most influential old events of our history as species could have taken place there, the origin of our species (discussed earlier in this chapter) and the Out of Africa.

The expansion of modern humans out of Africa has always been intensively debated. The exact point of embarkation, the timing and number of dispersals or even the number of individuals involved, are the main points that are not yet fully clarified. Genetic studies agree that present day non-Africans derive from a single wave out of Africa (53,54), and with a potential genetic contribution from an earlier wave present in the archaeological record (55). Whether the point of embarkation was through the Bab el-Mandeb strait (56,57) or the corridor in the north through the Sinai, in present Egypt (58) remains an open question.

Little is known about the pre-farming societies inhabiting this area. Only a few aDNA studies have shed light on such prehistorical times. In fact, these studies show the past cline of geographically structured hunter-gatherer populations from Ethiopia (ancient 4,500 year old Mota remains) (59), to Tanzania (1,400 year old remains)

until reaching southern Africa (2,000 year old South African close to present-day KhoeSan) (60).

The majority of populations inhabiting what is nowadays Ethiopia are farmers speaking Afro-asiatic (Cushitic, Semitic and Omotic branches) and Nilo-Saharan languages. In general, the genetic structure found in Ethiopia correlates with linguistic affiliation. A key demographical event that changed the genetic landscape of eastern Africa was the "Back to Africa" migration of the Levantine farmers dated to around 3,000 years ago (49,60,61). This is especially noticeable in some Afro-asiatic speaking populations where almost half of their genome is of Levantine ancestry (62). In contrast, Nilotic herder populations have remained more isolated and received very little or none Eurasian gene flow.

The majority of studies on detecting positive selection in East African populations have focussed on lactase persistence and adaptation to high altitude environments. Several variants that confer lactase persistence have arisen independently in pastoralist populations, mainly in Europe and East African populations from Kenya and Tanzania (63,64). Several studies on high-altitude adaptation have focused on Afro-asiatic populations, mainly Amhara and Oromo that inhabit regions at 2,500 meters above the sea level and several genes and molecular pathways have been reported as putative candidates of positive selection (65–68).

1.2.2. The Nama: a KhoeSan population from Namibia

The indigenous populations from southern Africa, the KhoeSan, are believed to be the deepest extant human lineage (Figure 4) and hence making southern Africa a putative key point in modern human origins (45). Some of them still practice a hunter-gatherer mode of subsistence, which is the main mode that our species has used during its evolutionary history. Moreover, the KhoeSan language family (or families) has the greatest phonemic diversity (e.g. click consonants) worldwide, another indicator of the southern African origin of modern humans (28). The term KhoeSan (*khoe* means 'person' and *san* means 'bushmen' in Khoekhoe) reflects an ethnological division into two groups: the Khoekhoe and the San (35).

Studies of the genetic landscape of the KhoeSan indicate the presence of three different ancestry components. These components are: Northern component (mostly represented by the Ju|'Hoansi and the !Xun), Central component (e.g. G|ui, G||ana and Naro) and Southern or Circum-Kalahari component (e.g. the Nama, ≠Khomani and Karretjie). There is not a clear correlation between genetics, linguistics and modes of subsistence among the KhoeSan since the three genetic components contain different KhoeSan family languages, modes of subsistence and cultural practices (50,69,70).

Recent admixture events have been described in the KhoeSan, the main sources being from east Africa, Bantu speaking populations and European colonists. East African admixture, has been dated to 2000 years (49,71) and is found at higher levels among the Khoekhoe groups, up to a 23-30% of ancestry (45,49,51,60). This migration supports a demic diffusion of pastoralism from East Africa during the same period of time, it would have been a migration previous to the Bantu expansion and would be at the base of finding a proportion of Levant ancestry in the southern African populations (60,71,72). Recent studies have found that the most probable East African donor source population was an already admixed group with an average 30% of Levantine ancestry, such as the current Afro-asiatic populations from Ethiopia (51). As discussed earlier in this chapter, there is evidence that the Bantu-speaking groups reached the south of Africa through two different waves of the Bantu expansion by 1,300 BP (31,46–48). The Bantu expansion had a great impact on the populations of southern Africa at the linguistic, cultural and genetic level (49,50). Lastly, the most recent event of admixture described in the KhoeSan is the arrival of European colonists in the area some 300 years ago, with a bigger genetic impact that the strong ethnic separation would have envisaged.

The multiple episodes of gene flow in the KhoeSan that in some cases led to lifestyle transitions (e.g. to pastoralism) have also left recent footprints of adaptive selection in the KhoeSan genomes. Specific examples of this are the *LCT* (related to lactase persistence) and *SLC24A5* (related to skin pigmentation) gene loci. In both cases, there is evidence that migration introduced variation that was then positively selected in KhoeSan populations. In the case of *LCT,* the East African migration that brought pastoralism into southern

Africa also conferred lactase persistence in some Khoekhoe groups (64,72,73). The light skin associated pigmentation gene, *SLC24A5,* has also been investigated in the KhoeSan populations. In fact, the European light skin associated allele (rs1426654) is found at a higher allele frequency expected just by admixture, indicating an event of strong positive selection (74). Part of this thesis is dedicated to investigate these special cases of adaptation.

In addition, the influx of infectious diseases by the migrant groups has probably been an important source of selective pressures among the KhoeSan. Several epidemics such as the flu and smallpox have been documented to greatly affect KhoeSan populations leaving genomic footprints of adaptation in the extant populations (75).

1.2.3. Underrepresented African populations in genetic studies

Despite the diversity and complexity of African evolutionary history unravelled by many different fields of study, African populations have remained hugely underrepresented in human population genetic studies. This lack of African studies is particularly overwhelming in genome-wide association studies (GWAS) where there is a clear bias towards studies with European individuals (76). In 2017 only 0.57% of GWAS were done with participants of African ancestry (77). This deficit of African data produces gaps in the understanding of human evolutionary history, complex traits and disease susceptibility in different environmental contexts (78). However, an increasing number of studies and consortia have tried to put more attention on the African continent. Some examples are the African Genome Variation Project (79), H3Africa (80), the 1000 Genomes Project 3rd phase release (81) which included five populations of African ancestry, the Southern African Human Genome Programme (82).

2. Detecting signals of positive selection

Charles Darwin and Alfred Wallace introduced in 1858 the concept of natural selection (83). With this concept they provided a framework to understand how species evolve. The way we

understand today natural selection is as a simple and clear concept: heritable traits can increase or decrease a set of individuals fitness given a specific environment and the frequency of their associated genotype will in turn increase or decrease across generations. Positive (also called adaptive) selection is the force that drives the increase in frequency of beneficial heritable traits and negative (also called purifying) selection is the force that decreases deleterious heritable traits.

Natural selection is a force that shapes genetic diversity and can be classified into three major categories.

Positive selection: also called adaptive or Darwinian selection, drives the increase in frequency of advantageous alleles in a population. This event will cause a selective sweep, which is the typical pattern of decreased genetic variability around the beneficial allele due to genetic hitchhiking (84). Over time, recombination will break the beneficial haplotype and the footprints of positive selection will be less obvious. The majority of methods that detect positive selection look for patterns of selective sweeps (e.g. non-neutral distribution of the site frequency spectrum, unexpectedly long blocks of haplotypes and allele frequency differences between populations).

Purifying selection: also called negative or background selection, removes deleterious mutations from a population. It is considered the most common mode of selection in functional elements since new mutations are more likely to be harmful than beneficial.

Balancing selection: maintains genetic diversity, as opposed to positive or negative selection that drives allele fixation. It can be the consequence of an advantageous heterozygote state, a frequency dependent selection (the fitness depends on the allele frequency) or the variation of selection across time and space.

This thesis is focused on the detection and interpretation of signals of positive selection in whole genome sequences of African populations. In the following parts of this thesis we will focus on the statistical methods existing to detect positive selection, the confounding factors and how to minimize them.

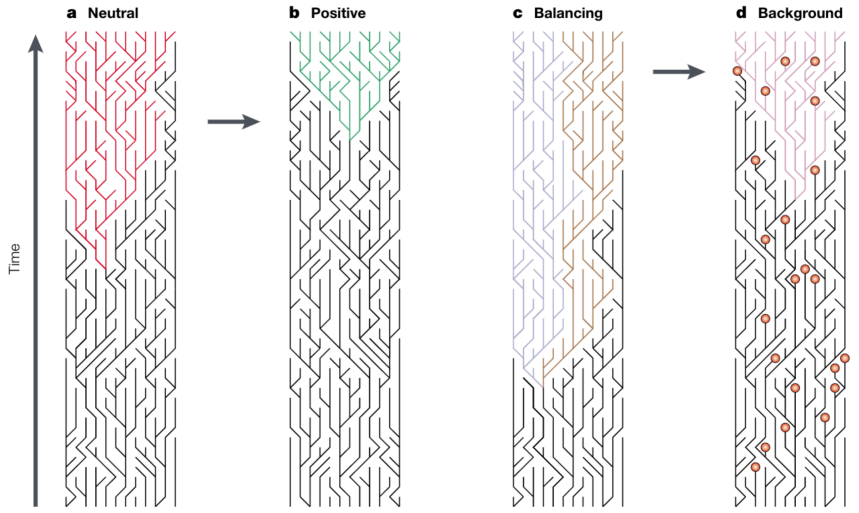


Figure 5: Gene genealogies under different neutral and natural selection scenarios. **a**) A genealogy of a neutral allele (red) drifts until fixation. **b**) Genealogy of an allele (in green) that reaches fixation much more quicker than in the neutral case after the beginning of positive selection (black arrow). **c**) Genealogy of two alleles (blue and brown) that are maintained in the population through the action of balancing selection. **d**) The genealogy of the allele (pink) drifts to fixation while background selection wipes out deleterious alleles (85).

2.1 Statistical methods to identify positive selection

Historically, two families of methods have been used to detect selection. Those based on divergence comparing different species, and those based on polymorphisms analysing populations. In this thesis we will only focus on polymorphism data.

Most methods that try to unravel the specific genetic adaptations in genomes using polymorphism data are focused on detecting the signatures of a hard sweep event, where selection acts on a newly arisen beneficial mutation (84). The specific properties of the genomic signatures of hard sweep model (skewed site frequency spectrum, extended homozygous haplotypes and high population differentiation) led to the development of three big families of methods (Figure 6). An important factor when choosing the appropriate method of analysis is the timeframe of study. Figure 7 indicates which are the time-depths for the detection of positive selection for each type of test.

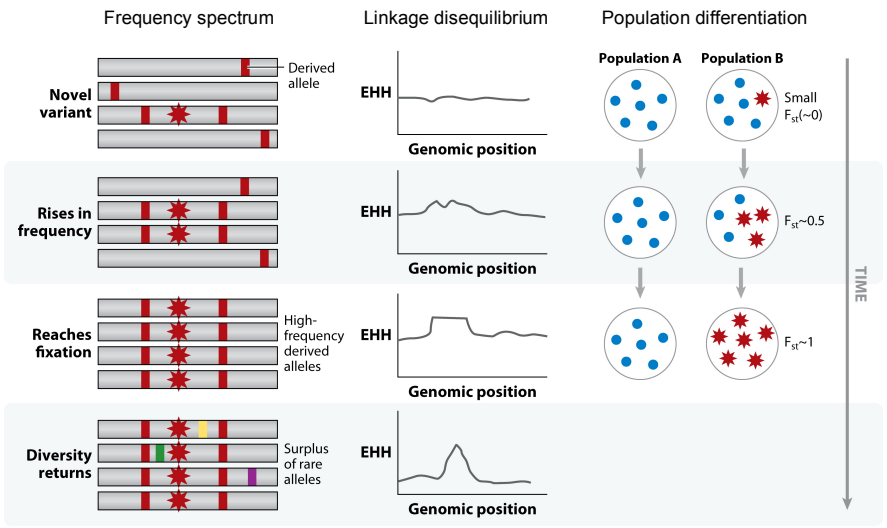


Figure 6: Detecting the footprints of positive selection in the genome. The site frequency spectrum, linkage disequilibrium and population differentiation-based tests (adapted from Vitti, Grossman, and Sabeti 2013).

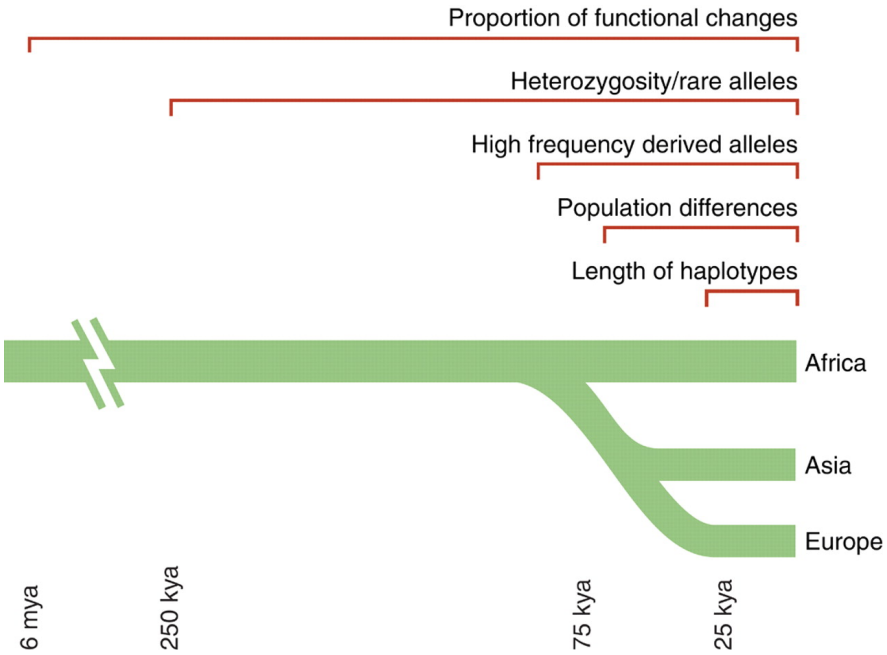


Figure 7: The different patterns left by positive selection in the genome have different time scales where they can be detected (87)**.**

Site Frequency Spectrum based tests: The Site Frequency Spectrum (SFS) is the allele frequency distribution composed by the numbers S_1, ..., S_{n-1}, where S_k is the number of SNPs present in *k* from *n* individuals. Usually it may consider the derived variants, thus having a direction. The increase in frequency of an adaptive allele will skew the SFS expected under neutrality such as there will be an increase of rare derived alleles and a decrease of intermediate allele frequencies. One of the widely used SFS based methods is Tajima's D (88) that tries to identify regions with an excess of rare alleles.

Haplotype based tests: the rapid increase of the adaptive allele will leave a pattern of high linkage disequilibrium generating long homozygous haplotypes. One of the firsts tests based on linkage disequilibrium is the Extended Haplotype Homozygosity (EHH, Pardis C. Sabeti et al. 2002), which calculates de decay of haplotype homozygosity from a core SNP (Figure 8). However, given that a low recombination rates could mimic the effect of a selective sweep, methods such as iHS (90) that compare the EHH of the haplotypes carrying the ancestral and the haplotypes carrying the derived allele have been developed to overcome this bias. Other flavours of these tests (XP-EHH and Rsb) are able to perform cross-population tests that contrasts patterns of EHH between populations (91,92). These tests are well suited for recent adaptive events since the patterns of linkage disequilibrium are erased by recombination over time.

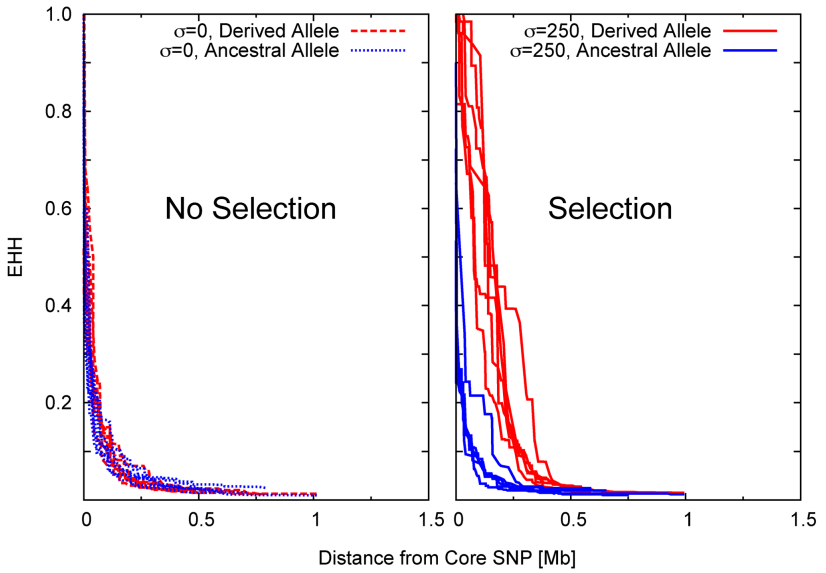


Figure 8: Decay of haplotype homozigosity from a core SNP in a neutral scenario (left) and a selection scenario (right). The ratio between the area below the curve of the ancestral and derived allele is at the base of iHS (90).

Population Differentiation: populations are subject to distinct environmental pressures, hence adaptive selection will change the allele frequencies from one population but not another. Therefore differences in allele frequencies between populations may be indicators of targets of positive selection. The population index, F_{st} (93), which is the proportion of genetic diversity due to allele frequency difference between populations, is the most commonly used test of this type. Other F_{st} derived tests such as the population branch statistic (PBS) and the locus-specific branch lengths (LSBL) were developed to infer in the direction of selection by adding a third population (94,95).

Composite and machine-learning methods: The composite methods usually combine multiple selection tests to increase the power of detection of positive selection. They can combine multiple types of methods such as XP-CLR that combines linkage disequilibrium and allele frequency information. There are even more complex composite methods that use machine-learning algorithms and extensive neutral and selection simulations to summarize multiple tests into a single score (96).

Although not considered in this work, several methods are trying to incorporate other types of sweep models than the classical hard sweep model. For example, empirical and theoretical studies indicate the importance of the soft sweep model (97–100). This type of selective sweep model considers that a neutral variant already segregating in the population becomes beneficial and rises in frequency; in this case it can also be named a sweep from standing variation. Another scenario could involve recurrent beneficial mutations in the same loci that rise in frequency together (101). The majority of methods focused on detecting soft sweeps are based on linkage-disequilibrium (102–104), but given the heterogeneous haplotypic background of a soft sweep, it is not an easy task to detect them. Additionally, another mode of selection is polygenic adaptation, which consists in hundreds of subtle allele frequency changes in multiple genes (105,106). The current methods to detect polygenic adaptation are still afflicted by a lack of statistical power and most of the methods rely on GWAS.

2.2 Confounding factors

This thesis mainly focuses on the detection of footprints of adaptive selection in human populations. It is therefore important to discuss some of the challenges that arise when conducting such studies. In this section we will briefly outline the confounding factors that can bias the detection of positive selection.

2.2.1 Background selection

Background selection, reduces the diversity nearby the deleterious alleles that are removed from the population. This pattern is similar to the one left by positive selection around a selected allele. Nonetheless, several studies argue that it is possible to exclude the confounding effect of background selection, specially using extended haplotype based methods such as iHS and XP-EHH (107,108).

2.2.2 Demography

The demographic history of a population shapes the patterns of polymorphism in their genomes. These patterns left by demography dynamics can mimic the patterns left by positive selection. It is therefore important to understand the main demographic mechanisms that could bias the detection of footprints of positive selection. The main actors in demography are outlined above.

Population structure: from a theoretical point of view, it is always assumed that all individuals in a population reproduce randomly and with the same probability. However, there are geographical, linguistic and cultural barriers that can affect the assumed random mating of individuals. In this case, we will find genetic substructures in the population with a higher genetic variability and an excess of variants at intermediate frequencies.

Migration: the movement of individuals from a population to another will introduce new variation in a population, increasing genetic variability and producing an excess of rare variants.

Population bottleneck: is the rapid decline of population size. This phenomenon will reduce genetic variability and increase linkage disequilibrium. It is common to find a recovery phase after a bottleneck, generating an excess of rare alleles. One of the best examples in humans of drastic population bottleneck is the OoA, which is at the base of the patterns of lower genetic variation and higher linkage disequilibrium found in present-day non-African populations.

Population expansion: is a sudden increase in size of a population, which generates an excess of rare variants and a lower genetic variability than expected.

2.3 Is it a true target of positive selection?

Given the confounding factors that could bias the detection of positive selection in a genome, assessing whether a particular region of the genome has been targeted by adaptive selection can be

challenging. There are two main approaches that can be implemented to minimize the effect of confounding factors.

2.3.1 The outlier approach

The outlier approach is based on the assumption that demography affects stochastically the entire genome while adaptive selection targets specific locations. Thus, one can simply build an empirical distribution of the positive selection statistics and consider the loci at the extreme tail of the distribution (the outliers) as putative targets of positive selection.
Although the wide use and simplicity of this methodology, there are some caveats worth to mention. First, the outlier approach requires defining the threshold used to consider a region significant. In other words, one should define which is the proportion of the genome that is considered under positive selection. Since there is no true estimate of this proportion, a 1% to 5% of most extreme scores are considered being under positive selection. Second, a particular region of the genome might randomly have extreme polymorphism patterns that mimic positive selection (e.g. a population bottleneck) and result in a false positive. Finally, another limitation is that this approach only identifies extreme cases of adaptive selection and many selected alleles affected by lower selective coefficients will be likely considered false negatives.

2.3.2 Population genetics simulations

Population genetics simulations enable to accurately reproduce the genomic properties of a population by simulating sequences using the parameters of a realistic demographical scenario (Figure 9).
The implementation of statistical tests of positive selection on simulated neutral genetic data allows investigating how would the statistical method behave under a neutral model without positive selection. The significance threshold is then estimated for a given false positive rate from the neutral data and can be applied to the empirical data. This methodology provides a perfect null hypothesis, which takes into account putative real data biases, to assess statistical significance of the targets of positive selection in empirical data.
Moreover, simulations that incorporate selective events with different attributes (e.g. combination of different selective

coefficients, timings and durations of the selective event) can be used to evaluate the accuracy of the significance threshold and the power of the statistic.

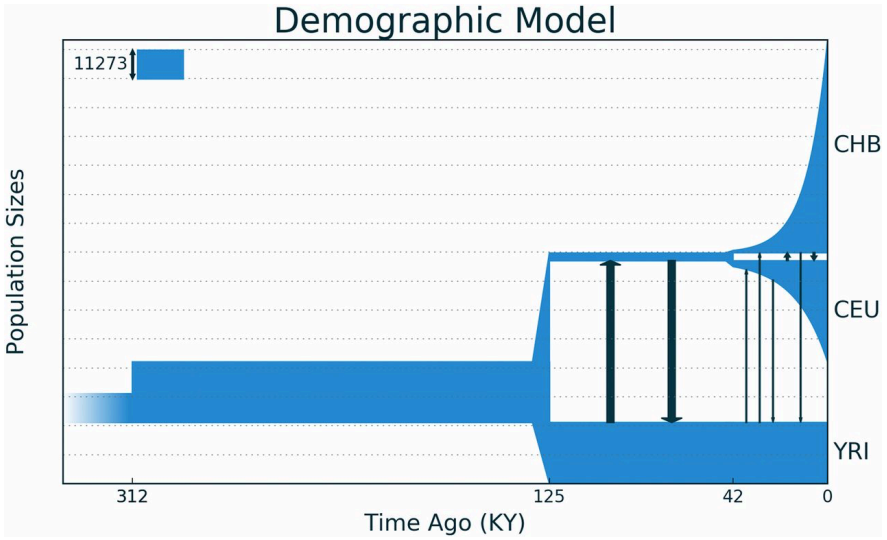


Figure 9: An estimated human three-population demographic model. Population sizes are indicated in blue, migrations are indicated in black arrows. The represented populations are Asian (CHB), European (CEU) and African (YRI) (109).

2.4 From genotype to phenotype

The ultimate goal of the study of positive selection signals in populations is to link an advantageous variant to a phenotype. Specifically, the goal is to understand the functional consequences of the adaptive allele from the most basic molecular level (intermediate phenotypes or endophenotypes) to the final adaptive trait. However, the majority of studies do not provide with a fully experimental follow-up to uncover the relationship between genotype and phenotype. There are typically four stages to be followed when trying to unravel the connections between genotype, phenotype and fitness:

Identification of candidate adaptive regions: in this step, there is a first implementation of statistical methods to detect adaptive selection in empirical data. Afterwards, and as mentioned in section 2.3, two methodologies can be used to assess the statistical significance of candidate loci under positive selection.

Identification of the functional variant: after identifying candidate loci, we must try to pinpoint the specific variants that drive adaptation. Bioinformatics tools to identify potential changes in protein function and comparison of allele frequencies between populations are commonly used to identify causal variants.

Exploring the functional consequences of the adaptive allele: this can include a broad range of experimental methodologies, from *in vitro* experiments, to model organisms and genotype-phenotype association studies in humans. The type of experimental analysis used will depend on the characteristics of the variant. For example, in the case of variants that cause nonsynonymous substitutions, transfection studies can explore the electrophysiological differences between cells carrying selected and non-selected alleles.

Understanding the link between the adaptive genotype and reproductive fitness: even if a functional effect is found at the molecular level, in the majority of the cases, there is still not enough evidence to unravel the specific selective force that drove the functional differences that in turn increased the reproductive fitness of carriers.

Thus, even if the huge power of high throughput whole-genome sequencing and the powerful computational methods that enable scientists to pinpoint regions of the genome that have been targeted by positive selection, there is still a considerable gap to fill between genotype and phenotype. That is why interdisciplinary approaches integrating different layers of *omics* data (transcriptomics, proteomics, metabolomics, epigenomics, interactomics etc…) across time and space, genome-wide association studies of different populations and more functional analyses are some of the keys to better understand human adaptation.

The work in this thesis represents a small contribution in the study of African human evolutionary history. First, the study of Ethiopian populations will shed some light into the potential adaptations of Afro-asiatic and Nilotic populations taking into account the recent admixture. Secondly, we will focus on the Nama, a semi-nomadic pastoralist KhoeSan population. Finally, the third study is focused on the functional consequences of a specific selection signal in the Gumuz population.

II. RESULTS

1. POSITIVE SELECTION IN ADMIXED POPULATIONS FROM ETHIOPIA

Sandra Walsh[1], Luca Pagani[2,3], Yali Xue[4], Hafid Laayouni[1,5], Chris Tyler-Smith[4], Jaume Bertranpetit[1]

1. Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Dr. Aiguader, 88. 08003 Barcelona, Catalonia, Spain.
2. Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu 51010, Estonia;
3. Department of Biology, University of Padova, Padova 35131, Italy.
4. The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK.
5. Bioinformatics Studies, ESCI-UPF, Barcelona, Catalonia, Spain.

Co-last authors CTS (cts@sanger.ac.uk) and JB (jaume.bertranpetit@upf.edu)

Keywords
Positive selection, selective sweeps, Ethiopia, Admixture, West Asia

Abstract

Background
Humans everywhere have adapted to their environments, including by genetic changes over many generations: the process of positive selection. Positive selection has, however, been under-studied in African populations, despite their diversity and importance for understanding human history.
Results
Here, we have used 119 available whole-genome sequences from five Ethiopian populations (Amhara, Oromo, Somali, Wolayta and Gumuz) to investigate the modes and targets of positive selection in this part of the world. The site frequency spectrum-based test SFselect was applied to identify a wide range of events of selection (old and recent), and the haplotype-based statistic iHS to detect more recent events, in each case with evaluation of the significance of candidate signals by extensive simulations. Additional insights were provided by considering admixture proportions and functional categories of genes. We identified both individual loci that are likely targets of classic sweeps and groups of genes that may have experienced polygenic adaptation. We found population-specific as well as shared signals of selection, with folate metabolism and the related UV response and skin pigmentation standing out as a shared pathway, perhaps as a response to the high levels of UV irradiation, and in addition strong signals in genes such as *IFNA*, *MRC1*, immunoglobulins and T-cell receptors which contribute to defend against pathogens.
Conclusions
Signals of both ancient and more recent positive selection could be detected in Ethiopian populations, revealing novel adaptations in East Africa, and abundant targets for functional follow-up.

Introduction

Genetic and archaeological data demonstrate that Africa is the origin of anatomically modern humans (6,52,110,111), and that populations outside Africa derive from an Out-of-Africa (OoA) migration some 60,000 years ago (58,112–115). African populations are genetically more diverse, holding the highest amount of genetic variation, low linkage disequilibrium (LD), and deep population structure (62,79,116,117). They also carry high cultural and phenotypic diversity, speak almost one-third of the world's languages (http://www.ethnologue.com), live in a wide variety of environments including deserts, tropical rainforests and mountain highlands, and follow many subsistence strategies, including pastoralism, agriculture and hunter-gathering (118). Surprisingly, however, African populations are underrepresented in big genetic projects such as the HGDP (119), 1000 Genomes Project (81) and HapMap (120). Consequently, not only is our understanding of the evolutionary processes that shape human diversity and adaptation limited, but also medical studies are prone to falter when African populations are included, due either to the fact that the single nucleotide polymorphisms (SNP) used are ascertained mainly in Eurasian populations, or to the lower LD found in all African populations (79). Additional African-specific studies are needed to counterbalance this historical bias (79,121).

Ethiopian populations lie geographically near a possible embarkation point of the OoA migration (56,57), exhibit high linguistic diversity encompassing three branches of the Afroasiatic language family (Omotic, Semitic, Cushitic) and also the Nilotic language family, and inhabit environments from lowland to highland. Previous genotyping studies have found a strong match between linguistic and genetic structures, and revealed admixture between Ethiopian (principally Afroasiatic) and OoA populations (most likely from West Asia) around 3,000 years ago, contributing about half of the ancestry of some present-day populations in what has been called a "back to Africa" migration (49,61,62). A 4,500-year old ancient Ethiopian fossil, Mota, does not show this West Asian backflow (59), and provides direct insights into the earlier genetic make-up.

Because African populations have adapted to a variety of environments and subsistence strategies, it is crucial to conduct natural selection studies in order to observe how selective pressures

shaped their genomes and understand both our evolutionary history as a species and the population-specific local adaptations to these circumstances. Given the diverse features of African populations, we could expect to find a considerable number of signals of local adaptation. A number of approaches have been established to detect positive selection (122–124), and a few signals of adaptive selection in Africans have been reported. Some of the most well-known cases involve malaria resistance, driven by genes such as Glucose-6-phosphate dehydrogenase (*G6PD*) and the Duffy antigen protein (125). There is also evidence of high-altitude adaptation in Ethiopians living in the highlands, as well of recent positive selection for lactase persistence in eastern African pastoralists (64,67,115,126). However, genome-wide analyses of adaptive selection footprints have often reported fewer signals in Africans than in OoA populations (96,127,128), or failed to find adaptive selection in Africans, some arguing that neutral simulations demonstrate that the tails of the empirical distributions contain mainly false positive signals (129), meaning that demographic events (bottlenecks, population structure and expansions), rather than selection, dominate the results (130). Thus African populations offer a challenge in recognizing events of adaptive selection in the genome.

In addition, the power and false-positive rates of positive selection tests in recently-admixed populations have only been addressed in a few studies. In a study of African-Americans using real and simulated genetic data, recent admixture did not result in an increase of false positive rates for site frequency spectrum-based tests, but in general the power decreased (131). In contrast, in some cases when the selective pressure was very strong, studying the admixed population could provide more power to detect selection than the ancestral population because the signature of derived alleles around the fixed selected site was lost in the ancestral population, but admixture made them polymorphic again producing a signature that is easier to detect. Studies with Latin American (132), Tibetan (133), Malagasy (134) and South Asian (135) populations have found potential admixture-mediated adaptive regions using this methodology, although some controversy exists since another Latin American study did not find evidence of directional selection after admixture (136).

These examples reveal that detecting positive selection is far from trivial. Most positive selection tests assume a simple model of

a hard sweep, where a mutation arises and spreads rapidly in a population until fixation, carrying the adjacent neutral variation with it (86). However, the relative importance of hard versus soft sweeps in explaining the adaptation of different human populations is debated, and all forms of selection need to be considered.

Here, we analyse previously-generated whole-genome sequences of 120 Ethiopians from five different populations (9) covering a wide geographical range and belonging to four different linguistic groups (Nilotic, Omotic, Cushitic and Semitic) (Table 1 and Supplementary Figure 1 in Additional File 1). We provide new information about the adaptive processes that these populations have undergone by first detecting the regions of the genome that have been selected, and then interpreting the biological meaning and context of these adaptations.

Results

SFselect selection analysis

In order to detect selective sweeps, we first analysed the data using the site frequency spectrum-based test SFselect (137) to identify old events of selection in the five Ethiopian populations (Supplementary Figure 8, Additional File 1). This approach generates a score for each 30 kilobase (kb) window in the genome. We assessed the statistical significance of the scores by defining a critical value of the test, after performing extensive neutral simulations (see Methods), as corresponding to the 99.99^{th} percentile of the neutral distribution (see Methods); the threshold is different for each population. Our simulations were based on a three-population demographic model representing Africans, Europeans and Asians (109), adding an admixture event between Africans and Europeans. We calculated two different thresholds, one for an unadmixed African population (here, the Gumuz) and the second for an admixed African population (here, the four Afroasiatic populations) (Supplementary Figure 6 in Additional File 1 and Supplementary Table 1 in Additional file 2). After applying the relevant threshold to each of the five populations, we obtained windows considered as putative candidates for adaptive selection. The number of significant windows is shown in Table 2a.

To interpret these windows, we annotated the protein-coding genes that intersected them (Supplementary Table 4, Additional file 2). Many of the signals were shared between Afroasiatic

populations (Table 2a and Figure 4a), as expected from their genetic similarity and shared environment (see Additional File 3 and Supplementary Figures 2 and 3 in Additional File 1 for a short demographic analysis of the studied populations). The Amhara and Oromo populations shared the highest number of signals (79), whereas the Gumuz shared the least (from 37 to 41). We found many examples of shared signals of selection by all five East African populations (Supplementary Table 2 in Additional file 2 and Figure 4). We discuss illustrative examples of shared and population-specific signals here, and further examples in the Additional File 3.

One of the top-scoring windows in all populations (Amhara 4.8, Oromo 4.7, Somali 5.2, Wolayta 5.1, Gumuz 3.7) contains genes including *FOLR1* and *FOLR2* (Figure 1a), members of the folic acid receptor family. Members of this gene family bind folic acid and its reduced derivatives, and transport 5-methyltetrahydrofolate into cells. The gene products are secreted proteins that either anchor to membranes via a glycosyl-phosphatidylinositol linkage or exist in a soluble form. Mutations in these genes have been associated with neurodegeneration due to cerebral folate transport deficiency; supplementation of folic acid is usually recommended for pregnant women to avoid neural tube defects during foetal development (138). Folate is also essential in DNA synthesis, survival and growth of the malaria parasite, so antifolate antimalarial drugs are widely used in the treatment of malaria (139). To our knowledge, this is the first study that finds this gene cluster to be under selection. The fact that we found this window under selection in all populations, together with the important functions of these genes especially during development, indicates that these genes have probably played a pivotal role during the evolutionary history of East Africans and possibly in general within the human species. We discovered, and discuss below, other selection signals related directly or indirectly to folic acid metabolism.

Another example of a top-scoring window among all five populations does not directly overlap with any gene, but the very strong signal lies downstream of the gene *ZNF473* (Figure 1b and Supplementary Table 2 in Additional file 2). This is an interesting region since it has been described as under long-term balancing selection in African populations, and that has been recently targeted by positive selection in Eurasian populations (140).

A signal shared among the Amhara, Somali and Wolayta populations contains *MRC1* (Supplementary Table 4a, c, d in Additional file 2 and Figure 1c). *MRC1* (also known as *CD206*) encodes a mannose receptor that is part of the C-type lectin superfamily and plays important roles in both adaptive and innate immune systems such as clearance of endogenous molecules and antigen presentation. MRC1 is an endocytic receptor that can bind to numerous endogenous and exogenous molecules and is mainly expressed in macrophages, dendritic cells and nonvascular epithelium. Numerous studies have shown that the C-type lectin-like domain (CTLD) of MRC1 can bind to viruses (HIV, Dengue, HBV), fungi (*Candida albicans*) and bacteria (*Mycobacterium tuberculosis*) (141–143). It has also been shown that MRC1 can internalize antigens that can then be processed for cross-presentation in antigen-presenting cells (144) and that MRC1 directly interacts with and inhibits CD45 on the T-cell surface resulting in impaired cytotoxic activity of T-cells and antigen-specific T-cell tolerance (145). This inhibitory effect of T cells by MRC1 has been proposed as a possible therapeutic strategy to downregulate the excessive immune response of autoimmune diseases. In fact, variants in *MRC1* have been associated with asthma and sarcoidosis (146,147). In addition, variants of *MRC1* have been associated with susceptibility to leprosy in Vietnamese and Brazilian patients and to pulmonary tuberculosis in Chinese patients (148,149). This example introduces a second recurring theme, of selection on defence-related genes, which will be encountered further below.

We also detected population-specific signals of positive selection, and a particularly strong signal was found in the Amhara population, where the 30 kb window containing *IFNA14*, *IFNA16* and *IFNA17* showed a very high and statistically significant SFselect score of 3.9. These genes are members of the Interferon Alpha gene family (Figure 4d and Supplementary Table 4a in Additional file 2); Interferon Alpha is produced in virus-infected leukocytes and has antiviral activity. It has been shown *in vitro* that *IFNA17* is three times more efficient against Hepatitis C than *IFNA2A*, which is the most effective current treatment (150). Moreover, polymorphisms in *IFNA17* have been associated with a 3.6-fold increased risk for Crimean-Congo Haemorrhagic Fever development (151). These interferon genes provide further examples of selection on likely defence against pathogens. Other

categories of population-specific signal are discussed in the Additional File 3.

iHS selection analysis captures recent events of selection

We next used the linkage-disequilibrium-based test iHS (89) to detect recent events of selection in the five Ethiopian populations (Supplementary Figure 9 in Additional File 1). We analysed mean iHS scores in windows of 30 kb that passed the critical value defined after performing extensive neutral simulations (see Methods). We set a restrictive threshold at the 99.99^{th} percentile of the neutral distribution and only signals that passed this threshold were considered as candidates for adaptive selection (Supplementary Table 1 in Additional file 2 and Supplementary Figure 7 in Additional File 1).

The number of significant windows per population was low, and the five populations shared some windows under recent positive selection (Table 2b, Supplementary Table 3 in Additional file 2 and Figure 5a). Amhara and Oromo shared the highest number of windows (12 each) with Wolayta, while the lowest number of window shared (four each) was between Somali and Wolayta, and Somali and Gumuz. The Gumuz in general shared the lowest number of windows with the rest of the populations. The lower numbers of significant windows and shared windows from the iHS analysis compared with the SFselect analysis could be because the populations split quite recently, so there has been little time for selection signals to build up.

Although the intersection of signals between all five populations is modest, we do find some strong shared signals (Supplementary Table 3 in Additional file 2). *OTOA* shows high and significant mean iHS scores in all populations (except Wolayta, which is close) and variants with significant p-values. Specifically, one of the top variants in all populations (rs370153558) show p-values of 1.8×10^{-6}, 5×10^{-8}, 3.4×10^{-5}, 1×10^{-6}, 9×10^{-8} for Amhara, Oromo, Somali, Wolayta and Gumuz respectively (Figure 2a). All the variants in the *OTOA* gene found under strong selection lie in the intron 21. The protein encoded by *OTOA* (otoancorin) is expressed on the apical surface of epithelial cells in the sensory organs of the inner ear. Mutations in *OTOA* have been found in Palestinians and Pakistanis to be causative for autosomal recessive

deafness 22 (152,153). Hearing is a rapidly-evolving phenotype in humans (154) and thus the likely target of selection here.

Several examples of selection signals were related to UV protection or skin pigmentation, and thus indirectly to the folic acid metabolism discussed above. One of these was *UVRAG*, where we find a signal in all Afroasiatic populations. The strongest signal is found in the Oromo and Somali with mean iHS scores of 2.6 and 3.4 in the region and specific intronic variants such as rs10899132 (Somali $p=10^{-6}$, Oromo $p=10^{-4}$) are found although no clear functional predictions are yet described (Supplementary Table 5 in Additional file 2 and Figure 2b). The Amhara also show significant mean iHS scores in other windows containing *UVRAG* (mean iHS score of 2.65 and rs7117696 variant with $p=5x10^{-5}$). The lack of a signal from SFselect in this region of the genome also supports the idea of recent selection. *UVRAG* plays an essential role in protecting cells from UV-induced DNA damage by activating the nucleotide excision repair pathway (155). In addition, it acts as an autophagic tumour suppressor that is mutated in common human cancers (156).

A second candidate, shared between Wolayta and Gumuz, is *BNC2* with mean iHS scores per window 2.79 and 2.86 (Figure 2c) and specific variants such as rs113571602 with significant p-values ($4x10^{-7}$ and $2x10^{-7}$ respectively). Again, all the highest scoring variants fall in an intron of *BNC2*, pointing towards a putative regulatory change of gene expression. This gene codes for a DNA-binding zinc-finger protein that acts as an mRNA-processing enzyme and a transcription factor (157). It is expressed in melanocytes and keratinocytes and variants have been associated with skin colour, where higher expression levels correspond to darker skin (158). Interestingly, *BNC2* has been found to lie in an adaptive introgressed region from Neanderthals to Europeans (159), but the signal of selection in Ethiopia lies outside the reported introgressed region.

The third example in this category is found in a region containing *ZRANB3* with statistically significant mean iHS scores of 3.32 and variants with significant p-values such as rs11892059 ($p=1.8x10^{-6}$) (Supplementary Table 5c in Additional file 2, Figure 2d). ZRANB3 is an annealing helicase, fork remodeller and structure-specific nuclease; its deficiency can cause genome instability and hypersensitivity to diverse DNA damaging agents such as UV radiation (160). This region has previously been

reported as a putative selection candidate in the Maasai population (102) but the authors did not link the signal to adaptation related to UV radiation because this variant is in linkage disequilibrium with the well-known lactase (*LCT*) gene which has many times been reported to be under selection in several populations (63,64). In Ethiopians, the signal is clearly in *ZRANB3* and not in *LCT*. *ZRANB3* has also been found under selection among black Tibetan wild boars, providing more evidence for its important function to maintain genomic stability against the high UV radiation found in the Tibetan Plateau (161).

Many examples of selection signals related to defence were also found. Among these was a window in the Wolayta showing a statistically significant mean iHS of 3.14 and many variants with $p<10^{-5}$ in the upstream region of *IFNL1* and the intergenic location between *IFNL1* and *IFNL2*. Signals in the 99.9^{th} percentile are also found in the Amhara and Gumuz (Supplementary Table 5d in Additional file 2 and Figure 2e). The Interferon-λ family or type III IFNs has three members (*IFNL1*, *IFNL2*, *IFNL3*). These genes play a critical role in antiviral, antiproliferative, antitumor and immune responses (162). These responses often overlap with IFN-α functions such as MHC class I antigen expression and induction of antiviral cascades. Some of the antiviral activities of INFLs target hepatitis B and C virus, cytomegalovirus, influenza A virus, coronaviruses, encephalomyocarditis virus, intestinal infection viruses (noroviruses and rotaviruses) and human immunodeficiency virus (163). Clinical trials against hepatitis C virus have tested PEGylated IFNL1 and showed a better or equal effectiveness than PEGylated IFN-α with less extrahepatic adverse effects (164). Since humans are very frequently exposed to viruses of low pathogenicity, and IFN-λ mostly targets mucosal epithelial cells, the function of type III IFNs could be to protect from infections without triggering the severe inflammation and tissue damage that type I IFNs often produce in the long term (165). Additional signals related to defence in *TPCN1, CHUK, THEMIS* and *TRAV* are discussed in the Additional File 3.

Finally, a signal specific to Gumuz was found in *PKD2L1* (Figure 2f), with high iHS scoring variants such as rs74154621 and rs74154622 (iHS score $p<10^{-8}$) that are both whole blood eQTLs with a normalized effect size of -0.669 and $p<10^{-4}$ according to the GTEx portal. We also find several non-synonymous changes with a

high derived allele frequency in the Gumuz (rs17112895 and rs7909153 both at a frequency of 0.70). *PKD2L1* belongs to the TRPP subfamily of ion channels that are characterised by large extracellular domains (166). Several studies with mice have identified *PKD2L1* as a candidate for sour taste in mammals (167). It has been shown in mice that it can form complexes with other proteins of this family such as PKD1L3 or PKD1L1. The PKD2L1/PKD1L3 complex is expressed in a subset of taste receptor cells in specific taste areas (168). In humans, two patients with sour taste ageusia have been reported and neither had detectable *PKD2L1* transcripts (169). Sour taste is one of the five basic tastes. And although other tastes have a clear evolutionary purpose (sweet indicates carbohydrate rich food, salty taste sodium, bitter potentially poisonous and umami protein rich), sour tasting remains unexplored in humans. One of the main hypotheses of the evolutionary sour tasting function is that it could warn against the acidic ingestion of rotten or immature fruit (170). Further signals of selection from other functional categories such as skin pigmentation (*BLOC1S2*) and one in an RNA gene (*NSUN3*), were also detected (Additional File 3).

Effect of admixture on detecting ancient and recent selection

The power and false-positive rates of positive selection tests in admixed populations have only been addressed in a few studies (115,131–134). To provide further support for our selection analyses, we have investigated whether similar results could be obtained without the West Asian ancestry genetic component among the Afroasiatic populations. For that purpose, we masked the West Asian component from our data, keeping only the East African component (see Methods). Given that, on average, almost half of the genome was masked by this procedure, we merged all four Afroasiatic populations in a single meta-population and re-ran the positive selection tests used previously. PCA of the retained East African component confirmed the high similarity between the East African component of the Afroasiatic populations, supporting the combined meta-analysis of all the individuals (Supplementary Figure 7 in Additional File 1).

The comparisons between the top 20 signals of the SFselect analysis between each single population and the merged East African component show a high similarity between the two analyses (Figure 4b). In contrast, the overlaps of the iHS analyses were not

as strong (Figure 5b). This last result could be due because of the breaking down of the Ethiopian haplotypes by the ancestry switches that occurred after the West Asian admixture in the area or because of the nature of iHS that detects recent selection more likely to be specific to each population.

Enrichment of West Asian ancestry in windows under selection

The masked West Asian component measures the proportion of West Asian ancestry in each population (Table 3). The Amhara and Oromo populations have the highest amount (54% and 51%, respectively), Wolayta and Somali show 43% and 44%, respectively, while in contrast the Gumuz show the low amount of 0.7%. These values agree with previous estimates (58). To detect for any enrichment of West Asian ancestry in windows under selection, the same calculation of the proportion of the West Asian component was performed among regions under positive selection (99.99^{th} percentile after neutral simulations and 1% extreme scores) for all populations, for both iHS and SFselect. This analysis revealed a general increase of West Asian ancestry among the regions putatively under selection found with both the SFselect and iHS tests, with similar percentages for the two tests (Table 3). A resampling analysis shows that the difference is highly significant ($p < 10^{-5}$, see Methods): there is thus an overall enrichment of West Asian ancestry in regions under selection. It is worth mentioning that this enrichment of West Asian ancestry is not a source of false positive signals in our analysis given the results obtained when we analysed the effect of admixture on detecting adaptive selection (see above).

Unbalanced ancestry regions

Previous studies have used ancestral component proportions to detect regions with a strong ancestry imbalance that could potentially have positive or negative effects on the fitness of admixed populations (135). The admixture event between Ethiopian and West Asian populations is dated to 2500-3000 years ago (58,62), meaning that under a neutral model, we would expect the percentage of the West Asian ancestry component to be evenly distributed across the genome. Therefore we report regions with significant deviations from the expected distribution of West Asian component in several populations that could be candidates of adaptation.

A good example is a long stretch of chromosome 17 with an extreme 95% of African ancestry spanning more than 0.5 Megabases (Mb) in all Afroasiatic populations. Moreover, in this region we find the *CRHR1* (Corticotrophin Releasing Hormone Receptor 1) gene with a high SFselect score of 3 (significant after simulations) in the Amhara, and 2.22 and 2.1 in the Somali and Gumuz, respectively (close to significance). There are other genes in the region such as *KANSL1* and *MAPT* (Figure 3a). We also find in all Afroasiatic populations, except Wolayta, an excess of African ancestry in windows under selection containing among other genes *FADS1* and *FADS2*, two enzymes that participate in the omega-3 and omega-6 biosynthesis and found to be under positive selection in other human populations (Figure 3b) (171,172). In Oromo and Wolayta, high African ancestry (77 and 81% respectively) and high iHS scores (top scoring SNPs with $p<10^{-4}$) were found around the immunoglobulin heavy variable 1-8 genes (*IGHV1-8*), central to defence (Figure 3c).

Signatures of polygenic adaptation through functional enrichment analysis

Functional enrichment analysis can be used to understand the biological functions of groups of genes, in this case those that have putatively been under positive selection. For this analysis, we listed the genes contained in windows with scores higher than the empirical 99.5 percentile, either for SFselect or mean iHS. We relaxed the thresholds of significance since we are trying to detect loci contributing to polygenic selection and a biological term was considered significant if the p-value after a Benjamini-Hochbert correction was below an alpha value of 0.05 (Table 4). Details of the significant terms and associated genes in each population, and selection tests, can be found in the Additional file 3. Many of the biological categories significant in all populations are related to immune responses and defence (Table 4). Folate metabolism is also a recurrent function found in many populations, as well as for calcium homeostasis related functions. Finally, muscle development function also appears in several populations.

All in all, the enrichment analysis reinforced our previous analyses of selection, again highlighting several of the main adaptations that Ethiopian populations have undergone.

Discussion

In this study, we have found new gene candidates under adaptive selection in populations from East Africa. We have been able, by performing extensive simulations, to assess the significance of our candidate adaptive selection signals. We have provided evidence for both old and recent selective sweeps, and both shared and population-specific signals of selection, while accounting for any effect of admixture. Our work has also highlighted the genetic similarity among Afroasiatic Ethiopian populations since many of the old signals of selection are shared between them.

Selection analysis in recently admixed populations is of special interest as the adaptation process may maintain pre-admixture adaptations or use one of the components as the genetic background for new adaptations. It is thus of interest to compare the more ancient (likely pre-admixture) and more recent (post-admixture and population specific) adaptations. The site-frequency-based test SFselect captures ancient and shared selection events before the gene flow from West Asia into Africa and thus before the admixture of the West Asian and the East African components. Conversely, iHS captures recent selection that probably happened after admixture, and that is often population-specific either for the Nilotic (Gumuz) or Afroasiatic populations (Amhara, Oromo, Wolayta, Somali).

We have found that folate metabolism appears to have been crucial for Ethiopian populations, a trait that is new as an adaptation (173). Specifically, we have identified the genes *FOLR1, FOLR2* and *DHFRL1* (see Additional File 3) as candidates of adaptive selection, while the functional enrichment analysis also highlighted folate metabolism as a main function potentially under selection, and many genes related to skin pigmentation or UV protection were picked out. Folate is crucial for DNA biosynthesis, methylation and repair and its deficiency can cause fatal birth defects and hence can directly affect reproductive success. Sufficient folate is associated with a 72% reduced risk of neural tube defects (138) and it is known that folate deficiency severely challenges the nucleotide excision repair mechanism needed to remove UV induced DNA photoproducts (174). Ethiopia experiences very high ultraviolet radiation, which has consequences that include severe DNA damage and impaired genome integrity. It has been hypothesized that under high UVB and UVA radiation, dark skin pigmentation has been selected in order to avoid folate photolysis (the “vitamin D-folate

hypothesis") (175,176). Among the recent selective sweeps were many on genes involved in UV radiation response and pigmentation. In the Afroasiatic populations, we have found a region containing the *UVRAG* gene that activates the nucleotide excision repair pathway when there is UV-induced damage in cells. We have also found as selection candidates *BNC2* (among Amhara, Oromo, Wolayta and Gumuz) whose high expression is associated with dark skin colour, and *BLOC1S2,* encoding a subunit of the complex BLOC-1 that produces strong pigmentation phenotypes in mice and Hermansky-Pudlak syndrome in humans and also many functional enrichment categories related to UV responses. In addition, we have found *ZRANB3* in the Somali population, where deficiency causes genome instability and hypersensitivity to DNA damaging including UV radiation and has been found to be under selection in black Tibetan wild boars. Thus, there is strong evidence in our study pointing towards folate and pigmentation related adaptations.

The environmental changes and migrations that humans have often experienced have made immunological adaptations a key process during human evolution. Our study gives further insights into these immune-related adaptations in East Africa where the major causes of death are due to infections (HIV, tuberculosis, malaria and other acute lower respiratory infections). For example, we have found in Amhara a region containing *IFNA* genes that encode for interferon alpha (pivotal for antiviral responses) and a region that the Amhara, Somali and Wolayta share in common containing *MRC1* (an endocytic receptor involved in adaptive and innate immune responses). Most importantly, we have found in the Gumuz population regions under potentially recent adaptive selection containing genes belonging to the immunoglobulin heavy constant and variable chains and to the T-cell receptor alpha variable locus.

Although we have been able to highlight potentially adaptive regions through computational methods and elaborated on the possible biological implications that could have been pivotal for adaptation in East Africa, this is just a first step towards a better understanding of human adaptive evolution and further functional studies are needed in order to confirm our findings. Our work provides the foundation for such studies.

Methods

Data

The dataset comprised five East African populations (Amhara, Oromo, Somali, Wolayta and Gumuz) with 24 individuals each from Pagani et al. 2015. One Wolayta individual was excluded from all subsequent analysis due to a high degree of relatedness (data not shown). Additional samples from the 1000 Genomes Project (81) and a set of 100 Egyptian samples also from Pagani et al. 2015 were included in PCA and ADMIXTURE analyses. The genome assembly of the data is GRCh37 (hg19). A summary of the dataset is shown in Table 1.

PCA and ADMIXTURE

The PCAs were performed with *smartpca* from the Eigensoft 6.0.1 software (177). All individuals from the dataset were used to perform the worldwide PCA. For the local PCA, all Ethiopians were included, plus a random subset of 24 YRI and 24 CEU. We applied a general filter requiring minor allele frequency higher than 0.05. The PCA of the West Asian masked samples was done with the lsqproject mode that is suitable when the samples have large amounts of missing data.

Population structure analysis was performed with the ADMIXTURE software (178) on a reduced set of 13 populations, 24 individuals per population (with the exception of 23 Wolayta). Variants were pruned using the PLINK software (179) with parameters --indep 50 5 2 to remove the effect of linkage disequilibrium.

SFselect and iHS

SFselect is a machine-learning site frequency spectrum-based method to detect adaptive selection in polymorphism data (137). The program was developed using supervised learning (support vector machines) trained with extensive forward population simulations. The authors previously simulated neutral populations and populations where a selected allele experienced 200 different combinations of the parameters s (selection coefficient) and τ (time under selection). SFselect shows high power to detect positive selection compared to other tests based on the site frequency spectrum. Our sample size of 48 chromosomes per population provides with enough accuracy to make inferences based on SFS

(180). In this study, we used the general support vector machine trained model of SFselect and applied the test by dividing the whole genome into 30 kb windows with 5 kb overlap between windows.

We used the linkage disequilibrium-based test iHS (89) to detect recent events of selection in the five Ethiopian populations. The sample size per population is of 48 chromosomes, which according to (181) provides with enough power to detect signals of positive selection (minimum of 40 chromosomes is recommended). We used the physical positions to calculate iHS since there is no specific genetic map for these populations. We used the software rehh 2.0 (182) to calculate iHS for all the variants with a minor allele frequency higher than 0.05 and excluded a variant from the calculation if a 20 kb gap was found when calculating EHHs, as they may produce biases. In addition to the iHS score per SNP, we also calculated the mean iHS score (average of iHS scores across SNPs), and the maximum iHS value and $-\log_{10}$(p-value) of a SNP in each 30 kb window; these windows were the same as in the SFselect analysis.

To annotate the protein-coding gene content of windows, we used bedtools 2.24.0 (183) to intersect windows with the hg19 gene annotations from RefSeq. To annotate individual variants, ANNOVAR (184) was used.

Masking

Masking was performed as described previously (58). African and West Asian ancestries of the Ethiopian individuals were deconvoluted using PCAdmix on 20-SNP windows. The CEU and Gumuz populations were used as surrogate sources for the West Asian component and East African component respectively. The West Asian ancestry was subsequently masked.

After the masking procedure, the proportion of West Asian ancestry in a population was estimated by averaging the proportion of masked data across each SNP. For a specific 30 kb window, the same calculation was done but only including SNPs falling in the window. Consecutive 30 kb windows under selection were merged when calculating the West Asian component proportions. A resampling analysis was used to test if the general increase of West Asian component ancestry among the significant 99.99^{th} percentile SFselect and iHS windows was significant. We sampled the number of selected windows 10^5 times from the genome-wide windows and calculated the mean West Asian component ancestry in each to

obtain a distribution of means. The values obtained for windows were compared with this distribution.

Simulations

To test whether demographic events could mimic the genomic patterns expected from adaptation, we performed extensive simulations using a simple demographic model that captures the key elements to define the critical values for each of the tests. We used the sequence simulator SLiM (185) to generate samples of the human neutral demography. A demographic model adapted from (109) was used, adding a simple model of admixture between a sub-Saharan population and an OoA like population (58) (as a proxy for the West Asians) 2,600 years ago (Supplementary Figure 4 in Additional File 1). For simplicity, the Amhara population was used as example to model the admixture event common to all Afroasiatic populations, using a West Asian admixture proportion of 0.54.

We next checked the validity of the model by comparing the derived Site Frequency Spectra (SFS) from the real and simulated data (Supplementary Figure 5 in Additional File 1). The main differences between real and simulated data were seen among the singletons: a deficit of singletons was observed in the real data due to the low coverage, but otherwise the differences are very small, meaning that our model fits our data well.

There is an increase of extreme SFselect and iHS scores in our real data (Supplementary Figure 6 in Additional File 1). The 99.99^{th} percentile SFselect score thresholds after the neutral simulations for the Gumuz and the Afroasiatic populations are 2.24 and 2.31, respectively (Supplementary Table 1 in Additional file 2). For iHS, we calculated after the neutral simulations the 99.99^{th} percentile of both the per SNP p-value distribution and the 30-kb window of the mean absolute iHS scores (for an easy comparison with SFselect). We found that for the SNP-based analysis, the 99.99^{th} percentiles per SNP were 3.88 and 3.62 for Gumuz and the Afroasiatic populations, respectively. The window analysis set the 99.99^{th} percentile thresholds at 2.54 and 2.53.

Functional Enrichment analysis

To understand the biological functions that may have been under positive selection, we used ClueGo (186), a Cytoscape (187) plug-in that integrates Gene Ontology, KEGG pathways and several other databases to map groups of genes to specific functions.

ClueGo enables visualisation in a functionally grouped annotation network, a pie graph showing the group leading terms (most significant term among a group) and a histogram with all significant terms after p-adjustment (<0.05 after Benjamini-Hochbert correction) and their number of genes from the analysed cluster found in our list of genes. In this case, we used the genes falling among the top 99.5 percentile of SFselect and mean iHS scores. All information about significant terms and associated genes for each population and selection tests can be found in the Additional file 3.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
The Ethiopian datasets analysed during the current study are available in the European Genome-phenome Archive (EGA) repository with accession number EGAS00001000238.

Competing interests
The authors declare that they have no competing interests

Funding
This study has been possible thanks to the F.P.I. grant BES-2014-068994 to SW, and grant BFU2016-77961-P (AEI/FEDER, UE) both awarded by the Agencia Estatal de Investigación (MINECO, Spain) and with the support of Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 702). Part of the "Unidad de Excelencia María de Maeztu", funded by the MINECO (ref: MDM-2014-0370). YX and CTS are supported by Wellcome Trust (098051), LP is supported by the European Union through the European Regional Development Fund Project No. 2014-2020.4.01.16-0024, MOBTT53.

Authors' contributions

CTS, YX, LP, JB, HL conceived the study. SW analysed and interpreted the data. CTS, YX, LP, JB, HL and SW wrote the manuscript. All authors approved the final manuscript.

Acknowledgments

The authors thank the comments of the two anonymous referees that have made possible the improvement of the manuscript.

References

1. McDougall I, Brown FH, Fleagle JG. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature. 2005 Feb 17;433(7027):733–6.
2. White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, et al. Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature. 2003 Jun 12;423(6941):742–7.
3. Grove M, Lamb H, Roberts H, Davies S, Marshall M, Bates R, et al. Climatic variability, plasticity, and dispersal: A case study from Lake Tana, Ethiopia. J Hum Evol. 2015 Oct 1;87:32–47.
4. Day MH, Leakey MD, Magori C. A new hominid fossil skull (L.H. 18) from the Ngaloba Beds, Laetoli, northern Tanzania. Nature. 1980;284(5751):55–6.
5. Grun R, Brink JS, Spooner NA, Taylor L, Stringer CB, Franciscus RG, et al. Direct dating of Florisbad hominid [2]. Vol. 382, Nature. 1996. p. 500–1.
6. Hublin J-J, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, et al. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. Nature. 2017 Jun 7;546(7657):289–92.
7. Stojanowski CM. Iwo Eleru's place among Late Pleistocene and Early Holocene populations of North and East Africa. J Hum Evol. 2014;
8. Stringer C. The origin and evolution of homo sapiens. Vol. 371, Philosophical Transactions of the Royal Society B: Biological Sciences. Royal Society of London; 2016.
9. Henn BM, Steele TE, Weaver TD. Clarifying distinct models of modern human origins in Africa [Internet]. Vol. 53, Current Opinion in Genetics and Development. Elsevier Current Trends; 2018. p. 148–56.
10. Crevecoeur I, Brooks A, Ribot I, Cornelissen E, Semal P. Late Stone Age human remains from Ishango (Democratic Republic of Congo): New insights on Late Pleistocene modern human diversity in Africa. J Hum Evol. 2016;
11. Harvati K, Stringer C, Grün R, Aubert M, Allsworth-Jones P, Folorunso CA. The later stone age calvaria from Iwo Eleru, Nigeria: Morphology and chronology. PLoS One. 2011 Sep 15;6(9).
12. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F,

Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science (80-). 2012 Oct 12;338(6104):222–6.
13. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the neandertal genome. Science (80-). 2010 May 7;328(5979):710–22.
14. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. Genetic evidence for archaic admixture in Africa. Proc Natl Acad Sci U S A. 2011 Sep 13;108(37):15123–8.
15. Lorente-Galdos B, Lao O, Serra-Vidal G, Santpere G, Kuderna LFK, Arauna LR, et al. Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. Genome Biol. 2019;20(1):77.
16. Gibbons A. World's oldest *Homo sapiens* fossils found in Morocco. Science (80-). 2017 Jun 7;
17. Stringer CB, Grün R, Schwarcz HP, Goldberg P. ESR dates for the hominid burial site of Es Skhul in Israel. Nature. 1989;338(6218):756–8.
18. Grün R, Stringer C, McDermott F, Nathan R, Porat N, Robertson S, et al. U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. J Hum Evol. 2005;49(3):316–34.
19. Hershkovitz I, Weber GW, Quam R, Duval M, Grün R, Kinsley L, et al. The earliest modern humans outside Africa. Science (80-). 2018 Jan 26;359(6374):456–9.
20. Harvati K, Röding C, Bosman AM, Karakostis FA, Grün R, Stringer C, et al. Apidima Cave fossils provide earliest evidence of Homo sapiens in Eurasia. Nature. 2019 Jul 25;
21. Liu W, Martinón-Torres M, Cai Y, Xing S, Tong H, Pei S, et al. The earliest unequivocally modern humans in southern China. Nature. 2015 Oct 29;526(7575):696–9.
22. Westaway KE, Louys J, Awe RD, Morwood MJ, Price GJ, Zhao JX, et al. An early modern human presence in Sumatra 73,000-63,000 years ago. Nature. 2017 Aug 17;548(7667):322–5.
23. Mijares AS, Détroit F, Piper P, Grün R, Bellwood P, Aubert M, et al. New evidence for a 67,000-year-old human presence at Callao Cave, Luzon, Philippines. J Hum Evol. 2010;59(1):123–32.
24. Clarkson C, Jacobs Z, Marwick B, Fullagar R, Wallis L,

Smith M, et al. Human occupation of northern Australia by 65,000 years ago. Nature. 2017 Jul 19;547(7663):306–10.
25. Bae CJ, Douka K, Petraglia MD. On the origin of modern humans: Asian perspectives. Vol. 358, Science. American Association for the Advancement of Science; 2017.
26. Smith EA. Communication and collective action: Language and the evolution of human cooperation. Vol. 31, Evolution and Human Behavior. 2010. p. 231–45.
27. Hauser MD, Yang C, Berwick RC, Tattersall I, Ryan MJ, Watumull J, et al. The mystery of language evolution. Vol. 5, Frontiers in Psychology. Frontiers Research Foundation; 2014.
28. Atkinson QD. Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science (80-). 2011 Apr 15;332(6027):346–9.
29. Pakendorf B. Coevolution of languages and genes. Vol. 29, Current Opinion in Genetics and Development. Elsevier Ltd; 2014. p. 39–44.
30. Cavalli-Sforza LL, Minch E, Mountain JL. Coevolution of genes and languages revisited. Proc Natl Acad Sci U S A. 1992;89(12):5620–4.
31. de Filippo C, Bostoen K, Stoneking M, Pakendorf B. Bringing together linguistic and genetic evidence to test the Bantu expansion. Proc R Soc B Biol Sci. 2012 Aug 22;279(1741):3256–63.
32. Diamond J, Bellwood P. Farmers and their languages: The first expansions. Vol. 300, Science. 2003. p. 597–603.
33. Lewis MP. Ethnologue: Languages of the world. SIL international DallasTX; 2009. 1248 p.
34. Shanahan TM, Mckay NP, Hughen KA, Overpeck JT, Otto-Bliesner B, Heil CW, et al. The time-transgressive termination of the African humid period. Nat Geosci. 2015 Feb 17;8(2):140–4.
35. Barnard A. Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan Peoples. Cambridge: Cambridge University Press; 1992.
36. Beltrame MH, Rubel MA, Tishkoff SA. Inferences of African evolutionary history from genomic data. Vol. 41, Current Opinion in Genetics and Development. Elsevier Ltd; 2016. p. 159–66.
37. Underhill PA, Kivisild T. Use of Y Chromosome and

Mitochondrial DNA Population Structure in Tracing Human Migrations. 2007;
38. Ingman M, Kaessmann H, Pääbo S, Gyllensten U. Mitochondrial genome variation and the origin of modern humans. Nature. 2000 Dec 7;408(6813):708–13.
39. Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, et al. The Dawn of Human Matrilineal Diversity. Am J Hum Genet. 2008 May 9;82(5):1130–40.
40. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet. 2016 Jun 1;48(6):593–9.
41. Jobling MA, Tyler-Smith C. Human Y-chromosome variation in the genome-sequencing era. Vol. 18, Nature Reviews Genetics. Nature Publishing Group; 2017. p. 485–97.
42. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. Science (80-). 2009 May 22;324(5930):1035–44.
43. Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, et al. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci. 2011 Mar 29;108(13):5154–62.
44. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008 Feb 21;451(7181):998–1003.
45. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Blum MGB, Soodyall H, et al. Genomic Variation in Seven Khoe-San. 2012;1187(October):374–9.
46. Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science (80-). 2017 May 5;356(6337):543–6.
47. Li S, Schlebusch C, Jakobsson M. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. Proc R Soc B Biol Sci. 2014 Oct 22;281(1793):20141448.
48. Semo A, Gayà-Vidal M, Fortes-Lima C, Alard B, Oliveira S, Almeida J, et al. Mozambican genetic variation provides new

insights into the Bantu expansion. bioRxiv. 2019 Jul 10;697474.
49. Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A. 2014 Feb 18;111(7):2632–7.
50. Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, van Helden PD, et al. Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. Genetics. 2016 Sep 1;204(1):303–14.
51. Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science (80-). 2017 Nov 3;358(6363):652–5.
52. Schlebusch CM, Jakobsson M. Tales of Human Migration, Admixture, and Selection in Africa. Annu Rev Genomics Hum Genet. 2018 Aug 31;19(1):annurev-genom-083117-021759.
53. Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. Vol. 538, Nature. Nature Publishing Group; 2016. p. 207–14.
54. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016 Oct 21;538(7624):201–6.
55. Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016 Jul 1;538(7624):238–42.
56. Melé M, Javed A, Pybus M, Zalloua P, Haber M, Comas D, et al. Recombination Gives a New Insight in the Effective Population Size and the History of the Old World Human Populations. Mol Biol Evol. 2012 Jan 1;29(1):25–30.
57. Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet. 1999 Dec 1;23(4):437–41.
58. Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the Route of Modern Humans out of

Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. Am J Hum Genet. 2015;96(6):986–91.
59. Gallego Llorente M, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, et al. Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science. 2015 Nov 13;350(6262):820–2.
60. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, et al. Reconstructing Prehistoric African Population Structure. Cell. 2017 Sep 21;171(1):59-71.e21.
61. Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, et al. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science. 2006 Dec 15;314(5806):1767–70.
62. Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, et al. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am J Hum Genet. 2012;91:83–96.
63. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. Identification of a variant associated with adult-type hypolactasia. Nat Genet. 2002 Feb 14;30(2):233–7.
64. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007 Jan 10;39(1):31–40.
65. Huerta-Sánchez E, DeGiorgio M, Pagani L, Tarekegn A, Ekong R, Antao T, et al. Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. Mol Biol Evol. 2013 Aug 1;30(8):1877–88.
66. Scheinfeldt LB, Soi S, Thompson S, Ranciaro A, Woldemeskel D, Beggs W, et al. Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biol. 2012 Jan 20;13(1):R1.
67. Udpa N, Ronen R, Zhou D, Liang J, Stobdan T, Appenzeller O, et al. Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. Genome Biol. 2014 Feb 20;15(2):R36.
68. Alkorta-Aranburu G, Beall CM, Witonsky DB, Gebremedhin A, Pritchard JK, Di Rienzo A. The Genetic Architecture of Adaptations to High Altitude in Ethiopia. Malik HS, editor.

PLoS Genet. 2012 Dec 6;8(12):e1003110.
69. Montinaro F, Busby GBJ, Gonzalez-Santos M, Oosthuitzen O, Oosthuitzen E, Anagnostou P, et al. Complex Ancient Genetic Structure and Cultural Transitions in Southern African Populations. Genetics. 2017;205(1):303–16.
70. Vicente M, Jakobsson M, Ebbesen P, Schlebusch CM. Genetic Affinities among Southern Africa Hunter-Gatherers and the Impact of Admixing Farmer and Herder Populations. Heyer E, editor. Mol Biol Evol. 2019 Sep 1;36(9):1849–61.
71. Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, et al. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc Natl Acad Sci. 2008 Aug 5;105(31):10693–8.
72. Breton G, Schlebusch CM, Lombard M, Sjödin P, Soodyall H, Jakobsson M. Lactase Persistence Alleles Reveal Partial East African Ancestry of Southern African Khoe Pastoralists. Curr Biol. 2014 Apr 14;24(8):852–8.
73. Macholdt E, Lede V, Barbieri C, Mpoloka SW, Chen H, Slatkin M, et al. Tracing pastoralist migrations to southern Africa with lactase persistence alleles. Curr Biol. 2014 Apr 14;24(8):875–9.
74. Lin M, Siford RL, Martin AR, Nakagome S, Möller M, Hoal EG, et al. Rapid evolution of a skin-lightening allele in southern African KhoeSan. Proc Natl Acad Sci U S A. 2018;115(52):13324–9.
75. Owers KA, Sjödin P, Schlebusch CM, Skoglund P, Soodyall H, Jakobsson M. Adaptation to infectious disease exposure in indigenous Southern African populations. Proc R Soc B Biol Sci. 2017;284(1852).
76. Bien SA, Wojcik GL, Hodonsky CJ, Gignoux CR, Cheng I, Matise TC, et al. The Future of Genomic Studies Must Be Globally Representative: Perspectives from PAGE. Annu Rev Genomics Hum Genet. 2019 Aug 31;20(1):181–200.
77. Mills MC, Rahal C. A scientometric review of genome-wide association studies. Commun Biol. 2019 Dec;2(1).
78. Martin AR, Teferra S, Möller M, Hoal EG, Daly MJ. The critical needs and challenges for genetic architecture studies in Africa. Vol. 53, Current Opinion in Genetics and Development. Elsevier Ltd; 2018. p. 113–20.
79. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome

Variation Project shapes medical genetics in Africa. Nature. 2015 Jan 3;517(7534):327–32.
80. Rotimi C, Abayomi A, Abimiku A, Adabayeri VM, Adebamowo C, Adebiyi E, et al. Research capacity. Enabling the genomic revolution in Africa. Vol. 344, Science. American Association for the Advancement of Science; 2014. p. 1346–8.
81. Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, et al. A global reference for human genetic variation. Nature. 2015 Oct 30;526(7571):68–74.
82. Choudhury A, Ramsay M, Hazelhurst S, Aron S, Bardien S, Botha G, et al. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. Nat Commun. 2017 Dec 12;8(1):2062.
83. Darwin C, Wallace A. On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. J Proc Linn Soc London Zool. 1858 Aug;3(9):45–62.
84. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23(1):23–35.
85. Bamshad M, Wooding SP. Signatures of natural selection in the human genome. Vol. 4, Nature Reviews Genetics. 2003. p. 99–111.
86. Vitti JJ, Grossman SR, Sabeti PC. Detecting Natural Selection in Genomic Data. 2013;
87. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Vol. 312, Science. 2006. p. 1614–20.
88. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3):585–95.
89. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002 Oct 9;419(6909):832–7.
90. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4(3):0446–58.
91. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations.

Nature. 2007 Oct 18;449(7164):913–8.
92. Tang K, Thornton KR, Stoneking M. A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. 2007 Jul;5(7):1587–602.
93. Weir BS, Clark C. Estimating F-Statistics for the Analysis of Population Structure. Vol. 38, Cockerham Source: Evolution. 1984.
94. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science (80-). 2010 Jul 2;329(5987):75–8.
95. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, et al. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics. 2004;1(4):274–86.
96. Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, et al. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics. 2015 Aug 26;31(24):btv493.
97. Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005 Apr;169(4):2335–52.
98. Innan H, Kim Y. Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci U S A. 2004 Jul 20;101(29):10667–72.
99. Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. Am J Hum Genet. 2000;66(5):1669–79.
100. Hermisson J, Pennings PS. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol Evol. 2017 Jun 1;8(6):700–16.
101. Pennings PS, Hermisson J. Soft sweeps III: The signature of positive selection from recurrent mutation. PLoS Genet. 2006 Dec;2(12):1998–2012.
102. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol. 2014;31(5):1275–91.
103. Schrider DR, Kern AD. S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. PLoS Genet. 2016 Mar 1;12(3).

104. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. PLoS Genet. 2015;11(2):1–32.
105. Pritchard JK, Pickrell JK, Coop G. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Curr Biol. 2010 Feb 23;20(4):R208–15.
106. Pritchard JK, Di Rienzo A. Adaptation - Not by sweeps alone. Vol. 11, Nature Reviews Genetics. 2010. p. 665–7.
107. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res. 2014;24(6):885–95.
108. Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Mol Biol Evol. 2014;31(7):1850–68.
109. Jouganous J, Long W, Ragsdale AP, Gravel S. Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. Genetics. 2017;206(3):1549–67.
110. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci. 2005 Nov 1;102(44):15942–7.
111. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017 Jan 19;541(7637):302–10.
112. Campbell MC, Tishkoff SA. The Evolution of Human Genetic and Phenotypic Variation in Africa. Curr Biol. 2010 Feb 23;20(4):R166–73.
113. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci. 2011;108(29):11983–8.
114. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014 Aug 22;46(8):919–25.
115. Huerta-Sánchez E, DeGiorgio M, Pagani L, Tarekegn A, Ekong R, Antao T, et al. Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. Mol Biol Evol. 2013;30(8):1877–88.

116. Kreager P, Winney B, Ulijaszek S, Capelli C. Population in the human sciences: Concepts, models, evidence. Population in the Human Sciences: Concepts, Models, Evidence. Oxford University Press; 2015. 1–640 p.
117. Dobon B, Hassan HY, Laayouni H, Luisi P, Ricaño-Ponce I, Zhernakova A, et al. The genetics of East African populations: A Nilo-Saharan component in the African genetic landscape. Sci Rep. 2015 Sep 28;5(1):9996.
118. Campbell MC, Tishkoff SA. African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. Annu Rev Genomics Hum Genet. 2008 Sep;9(1):403–33.
119. Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L, et al. A human genome diversity cell line panel. Science. 2002 Apr 12;296(5566):261–2.
120. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Zhang H, et al. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789–96.
121. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2019 Jan 19;51(1):30–5.
122. Fan S, Hansen MEB, Lo Y, Tishkoff SA. Going global by adapting local: A review of recent human adaptation [Internet]. Vol. 354, Science. 2016. p. 54–9.
123. Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci. 2014 Apr 1;111(13):4832–7.
124. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015 Dec 23;528(7583):499–503.
125. Kwiatkowski DP. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. Am J Hum Genet. 2005 Aug 1;77(2):171–92.
126. Scheinfeldt LB, Soi S, Thompson S, Ranciaro A, Woldemeskel D, Beggs W, et al. Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biol. 2012 Jan 20;13(1):R1.
127. Storz JF, Payseur BA, Nachman MW. Genome Scans of

DNA Variability in Humans Reveal Evidence for Selective Sweeps Outside of Africa. Mol Biol Evol. 2004 May 21;21(9):1800–11.
128. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, et al. The Role of Geography in Human Adaptation. Schierup MH, editor. PLoS Genet. 2009 Jun 5;5(6):e1000500.
129. Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. Limited Evidence for Classic Selective Sweeps in African Populations. Genetics. 2012 Nov 1;192(3):1049–64.
130. Hofer T, Ray N, Wegmann D, Excoffier L. Large Allele Frequency Differences between Human Continental Groups are more Likely to have Occurred by Drift During range Expansions than by Selection. Ann Hum Genet. 2009 Jan;73(1):95–108.
131. Lohmueller KE, Bustamante CD, Clark AG. Detecting directional selection in the presence of recent admixture in African-Americans. Genetics. 2011;187(3):823–35.
132. Jin W, Xu S, Wang H, Yu Y, Shen Y, Wu B, et al. Genome-wide detection of natural selection in African Americans pre- and post-admixture. Genome Res. 2012;22(3):519–27.
133. Jeong C, Alkorta-Aranburu G, Basnyat B, Neupane M, Witonsky DB, Pritchard JK, et al. Admixture facilitates genetic adaptations to high altitude in Tibet. Nat Commun. 2014 Feb 10;5:3281.
134. Pierron D, Heiske M, Razafindrazaka H, Pereda-Loth V, Sanchez J, Alva O, et al. Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. Nat Commun. 2018;9(1):1–9.
135. Yelmen B, Mondal M, Marnetto D, Pathak AK, Montinaro F, Gallego Romero I, et al. Ancestry-Specific Analyses Reveal Differential Demographic Histories and Opposite Selective Pressures in Modern South Asian Populations. Heyer E, editor. Mol Biol Evol. 2019 Aug 1;36(8):1628–42.
136. Bhatia G, Tandon A, Patterson N, Aldrich MC, Ambrosone CB, Amos C, et al. Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. Am J Hum Genet. 2014;95(4):437–44.
137. Ronen R, Udpa N, Halperin E, Bafna V. Learning Natural Selection from the Site Frequency Spectrum. 2013;230–3.

138. Prevention of neural tube defects: Results of the Medical Research Council Vitamin Study. Lancet. 1991 Jul 20;338(8760):131–7.
139. Metz J. Folic Acid Metabolism and Malaria. Food Nutr Bull. 2007 Dec 2;28(4_suppl4):S540–9.
140. De Filippo C, Key FM, Ghirotto S, Benazzo A, Meneu JR, Weihmann A, et al. Recent Selection Changes in Human Genes under Long-Term Balancing Selection. Mol Biol Evol. 2016;33(6):1435–47.
141. East L, Isacke CM. The mannose receptor family [Internet]. Vol. 1572, Biochimica et Biophysica Acta - General Subjects. Elsevier; 2002. p. 364–86.
142. Martinez-Pomares L. The mannose receptor. J Leukoc Biol. 2012 Dec;92(6):1177–86.
143. Gazi U, Martinez-Pomares L. Influence of the mannose receptor in host immune responses. Immunobiology. 2009 Jul 1;214(7):554–61.
144. Burgdorf S, Kautz A, Bohnert V, Knolle PA, Kurts C. Distinct Pathways of Antigen Uptake and Intracellular Routing in CD4 and CD8 T Cell Activation. Science (80-). 2007 Apr 27;316(5824):612–6.
145. Schuette V, Embgenbroich M, Ulas T, Welz M, Schulte-Schrepping J, Draffehn AM, et al. Mannose receptor induces T-cell tolerance via inhibition of CD45 and up-regulation of CTLA-4. Proc Natl Acad Sci. 2016 Sep 20;113(38):10649–54.
146. Hattori T, Konno S, Hizawa N, Isada A, Takahashi A, Shimizu K, et al. Genetic variants in the mannose receptor gene (MRC1) are associated with asthma in two independent populations. Immunogenetics. 2009 Dec 10;61(11–12):731–8.
147. Hattori T, Konno S, Takahashi A, Isada A, Shimizu K, Shimizu K, et al. Genetic variants in mannose receptor gene (MRC1) confer susceptibility to increased risk of sarcoidosis. BMC Med Genet. 2010 Dec 28;11(1):151.
148. Alter A, de Léséleuc L, Van Thuc N, Thai VH, Huong NT, Ba NN, et al. Genetic and functional analysis of common MRC1 exon 7 polymorphisms in leprosy susceptibility. Hum Genet. 2010 Mar 25;127(3):337–48.
149. Zhang X, Li X, Zhang W, Wei L, Jiang T, Chen Z, et al. The novel human MRC1 gene polymorphisms are associated with

susceptibility to pulmonary tuberculosis in Chinese Uygur and Kazak populations. Mol Biol Rep. 2013 Aug 8;40(8):5073–83.
150. Dubois A, François C, Descamps V, Fournier C, Wychowski C, Dubuisson J, et al. Enhanced anti-HCV activity of interferon alpha 17 subtype. Virol J. 2009 Jun 3;6(1):70.
151. Elaldi N, Yilmaz M, Bagci B, Yelkovan I, Bagci G, Gozel MG, et al. Relationship between *IFNA1* , *IFNA5* , *IFNA10* , and *IFNA17* gene polymorphisms and Crimean-Congo hemorrhagic fever prognosis in a Turkish population range. J Med Virol. 2016 Jul 1;88(7):1159–67.
152. Zwaenepoel I, Mustapha M, Leibovici M, Verpy E, Goodyear R, Liu XZ, et al. Otoancorin, an inner ear protein restricted to the interface between the apical surface of sensory epithelia and their overlying acellular gels, is defective in autosomal recessive deafness DFNB22. Proc Natl Acad Sci. 2002 Apr 30;99(9):6240–5.
153. Lee K, Chiu I, Santos-Cortez R, Basit S, Khan S, Azeem Z, et al. Novel *OTOA* mutations cause autosomal recessive non-syndromic hearing impairment in Pakistani families. Clin Genet. 2013 Sep;84(3):294–6.
154. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science. 2003 Dec 12;302(5652):1960–3.
155. Yang Y, Quach C, Liang C. Autophagy modulator plays a part in UV protection. Autophagy. 2016 Sep 20;12(9):1677–8.
156. He S, Zhao Z, Yang Y, O'Connell D, Zhang X, Oh S, et al. Truncating mutation in the autophagy gene UVRAG confers oncogenic properties and chemosensitivity in colorectal cancers. Nat Commun. 2015 Dec 3;6(1):7839.
157. Visser M, Palstra R-J, Kayser M. Human skin color is influenced by an intergenic DNA polymorphism regulating transcription of the nearby BNC2 pigmentation gene. Hum Mol Genet. 2014 Nov 1;23(21):5750–62.
158. Chahal HS, Lin Y, Ransohoff KJ, Hinds DA, Wu W, Dai H-J, et al. Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. Nat Commun. 2016 Jul 18;7:12048.
159. Vernot B, Akey JM. Resurrecting Surviving Neandeltal

Linages from Modern Human Genomes. Science (80-). 2014;343(February):1017–21.
160. Poole LA, Cortez D. Functions of SMARCAL1, ZRANB3, and HLTF in maintaining genome stability [Internet]. Vol. 52, Critical Reviews in Biochemistry and Molecular Biology. Taylor & Francis; 2017. p. 696–714.
161. Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, et al. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nat Genet. 2013 Dec 27;45(12):1431–8.
162. Shi Y, Walter MR, Pestka S, Sarkar D, Fisher PB, Krause CD. I nterleukin -10 and R elated C ytokines and R eceptors . Annu Rev Immunol. 2004 Apr;22(1):929–79.
163. Li M, Liu X, Zhou Y, Su SB. Interferon- s: the modulators of antivirus, antitumor, and immune responses. J Leukoc Biol. 2009 Jul 1;86(1):23–32.
164. Muir AJ, Arora S, Everson G, Flisiak R, George J, Ghalib R, et al. A randomized phase 2b study of peginterferon lambda-1a for the treatment of chronic HCV infection. J Hepatol. 2014 Dec;61(6):1238–46.
165. Wack A, Terczyńska-Dyla E, Hartmann R. Guarding the frontiers: The biology of type III interferons [Internet]. Vol. 16, Nature Immunology. Nature Publishing Group; 2015. p. 802–9.
166. Su Q, Hu F, Liu Y, Ge X, Mei C, Yu S, et al. Cryo-EM structure of the polycystic kidney disease-like channel PKD2L1. Nat Commun. 2018;9(1):1–12.
167. Huang AL, Chen X, Hoon MA, Chandrashekar J, Guo W, Tränkner D, et al. The cells and logic for mammalian sour taste detection. Nature. 2006;442(7105):934–8.
168. Ishimaru Y, Inada H, Kubota M, Zhuang H, Tominaga M, Matsunami H. Transient receptor potential family members PKD1L3 and PKD2L1 form a candidate sour taste receptor. Proc Natl Acad Sci. 2006 Aug 15;103(33):12569–74.
169. Huque T, Cowart BJ, Dankulich-Nagrudny L, Pribitkin EA, Bayley DL, Spielman AI, et al. Sour ageusia in two individuals implicates ion channels of the ASIC and PKD families in human sour taste perception at the anterior tongue. PLoS One. 2009;4(10).
170. DeSimone JA, Heck GL, DeSimone SK. Active ion transport in dog tongue: a possible role in taste. Science. 1981 Nov

27;214(4524):1039–41.
171. Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jørgensen ME, et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science. 2015 Sep 18;349(6254):1343–7.
172. Ameur A, Enroth S, Johansson Å, Zaboli G, Igl W, Johansson ACV, et al. Genetic Adaptation of Fatty-Acid Metabolism: A Human-Specific Haplotype Increasing the Biosynthesis of Long-Chain Omega-3 and Omega-6 Fatty Acids. Am J Hum Genet. 2012 May 4;90(5):809–20.
173. Arciero E, Biagini SA, Chen Y, Xue Y, Luiselli D, Tyler-Smith C, et al. Genes Regulated by Vitamin D in Bone Cells Are Positively Selected in East Asians. Palsson A, editor. PLoS One. 2015 Dec 31;10(12):e0146072.
174. Han J, Colditz GA, Hunter DJ. Polymorphisms in the MTHFR and VDR genes and skin cancer risk. Carcinogenesis. 2007 Jul 8;28(2):390–7.
175. Jablonski NG, Chaplin G. Human skin pigmentation as an adaptation to UV radiation. Proc Natl Acad Sci. 2010;107(Supplement_2):8962–8.
176. Jones P, Lucock M, Veysey M, Beckett E. The Vitamin D¯Folate Hypothesis as an Evolutionary Model for Skin Pigmentation: An Update and Integration of Current Ideas. Nutrients. 2018 Apr 30;10(5).
177. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug 23;38(8):904–9.
178. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep 1;19(9):1655–64.
179. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.
180. Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN. Sampling strategies for frequency spectrum-based population genomic inference. BMC Evol Biol. 2014 Dec 4;14(1).
181. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 2009

May 1;19(5):826–37.
182. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol Ecol Resour. 2017 Jan 1;17(1):78–90.
183. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.
184. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep 1;38(16):e164–e164.
185. Haller BC, Messer PW. SLiM 2: Flexible, interactive forward genetic simulations. Mol Biol Evol. 2017;34(1):230–40.
186. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009 Apr 15;25(8):1091–3.
187. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498–504.
188. Haag S, Sloan KE, Ranjan N, Warda AS, Kretschmer J, Blessing C, et al. NSUN3 and ABH1 modify the wobble position of mt-tRNA ^{Met} to expand codon recognition in mitochondrial translation. EMBO J. 2016 Oct 4;35(19):2104–19.
189. Van Haute L, Dietmann S, Kremer L, Hussain S, Pearce SF, Powell CA, et al. Deficient methylation and formylation of mt-tRNAMet wobble cytosine in a patient carrying mutations in NSUN3. Nat Commun. 2016 Jun 30;7:12039.
190. McEntee G, Minguzzi S, O'Brien K, Ben Larbi N, Loscher C, O'Fagain C, et al. The former annotated human pseudogene dihydrofolate reductase-like 1 (DHFRL1) is expressed and functional. Proc Natl Acad Sci. 2011 Sep 13;108(37):15157–62.
191. Anderson DD, Quintero CM, Stover PJ. Identification of a de novo thymidylate biosynthesis pathway in mammalian mitochondria. Proc Natl Acad Sci. 2011 Sep

13;108(37):15163–8.
192. Haugen AC, Di Prospero NA, Parker JS, Fannin RD, Chou J, Meyer JN, et al. Altered Gene Expression and DNA Damage in Peripheral Blood Cells from Friedreich's Ataxia Patients: Cellular Model of Pathology. Pearson CE, editor. PLoS Genet. 2010 Jan 15;6(1):e1000812.
193. Llop S, Tran V, Ballester F, Barbone F, Sofianou-Katsoulis A, Sunyer J, et al. CYP3A genes and the association between prenatal methylmercury exposure and neurodevelopment. Environ Int. 2017 Aug;105:34–42.
194. Li Z, Chen P, Zhou T, Chen X, Chen L. Association between CYP3A5 genotypes with hypertension in Chinese Han population: A case-control study. Clin Exp Hypertens. 2017 Apr 3;39(3):235–40.
195. Fisher DL, Plange-Rhule J, Moreton M, Eastwood JB, Kerry SM, Micah F, et al. CYP3A5 as a candidate gene for hypertension: no support from an unselected indigenous West African population. J Hum Hypertens. 2016 Dec 23;30(12):778–82.
196. Dobon B, Rossell C, Walsh S, Bertranpetit J. Is there adaptation in the human genome for taste perception and phase i biotransformation? BMC Evol Biol. 2019 Dec 31;19(1):39.
197. Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, et al. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res. 2014 Jan;42(Database issue):D903-9.
198. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017 Oct 11;550(7675):204–13.
199. Fujioka M, Takahashi N, Odai H, Araki S, Ichikawa K, Feng J, et al. A New Isoform of Human Myosin Phosphatase Targeting/Regulatory Subunit (MYPT2): cDNA Cloning, Tissue Expression, and Chromosomal Mapping. Genomics. 1998 Apr 1;49(1):59–68.
200. Okamoto R, Kato T, Mizoguchi A, Takahashi N, Nakakuki T, Mizutani H, et al. Characterization and function of MYPT2, a target subunit of myosin phosphatase in heart. Cell Signal. 2006 Sep 1;18(9):1408–16.
201. Calcraft PJ, Ruas M, Pan Z, Cheng X, Arredouani A, Hao X,

et al. NAADP mobilizes calcium from acidic organelles through two-pore channels. Nature. 2009 May 22;459(7246):596–600.
202. Sakurai Y, Kolokoltsov AA, Chen C-C, Tidwell MW, Bauta WE, Klugbauer N, et al. Ebola virus. Two-pore channels control Ebola virus host cell entry and are drug targets for disease treatment. Science. 2015 Feb 27;347(6225):995–8.
203. Zou A, Lin Z, Humble M, Creech CD, Wagoner PK, Krafte D, et al. Distribution and functional properties of human KCNH8 (Elk1) potassium channels. Am J Physiol Physiol. 2003 Dec;285(6):C1356–66.
204. Lee HH, Nemecek D, Schindler C, Smith WJ, Ghirlando R, Steven AC, et al. Assembly and architecture of biogenesis of lysosome-related organelles complex-1 (BLOC-1). J Biol Chem. 2012 Feb 17;287(8):5882–90.
205. Li W, Zhang Q, Oiso N, Novak EK, Gautam R, O'Brien EP, et al. Hermansky-Pudlak syndrome type 7 (HPS-7) results from mutant dysbindin, a member of the biogenesis of lysosome-related organelles complex 1 (BLOC-1). Nat Genet. 2003 Sep 17;35(1):84–9.
206. Polley S, Passos DO, Huang D-B, Mulero MC, Mazumder A, Biswas T, et al. Structural Basis for the Activation of IKK1/α. Cell Rep. 2016 Nov 15;17(8):1907–14.
207. Allen PM. Themis imposes new law and order on positive selection. Nat Immunol. 2009 Aug 1;10(8):805–6.
208. Fu G, Casas J, Rigaud S, Rybakin V, Lambolez F, Brzostek J, et al. Themis sets the signal threshold for positive and negative selection in T-cell development. Nature. 2013 Dec 13;504(7480):441–5.
209. Traber PG, Wu GD, Wang W. Novel DNA-binding proteins regulate intestine-specific transcription of the sucrase-isomaltase gene. Mol Cell Biol. 1992 Aug;12(8):3614–27.
210. Van Beers EH, Büller HA, Grand RJ, Einerhand AWC, Dekker J. Intestinal brush border glycohydrolases: Structure, function, and development. Crit Rev Biochem Mol Biol. 1995;30(3):197–262.
211. Marcadier JL, Boland M, Scott CR, Issa K, Wu Z, McIntyre AD, et al. Congenital sucrase-isomaltase deficiency: Identification of a common Inuit founder mutation. CMAJ. 2015 Feb 3;187(2):102–7.

Additional Files

Additional file 1: **Supplementary Figures**. **Supplementary Figure 1.** Location of the five sampled populations. **Supplementary Figure 2.** Principal component analysis of the five East African populations. **Supplementary Figure 3.** ADMIXTURE analysis of the Ethiopian samples and a set of worldwide populations. **Supplementary Figure 4.** Schematic representation of the demographic model used to simulate neutral sequences. **Supplementary Figure 5.** Relative site frequency spectrum of Afroasiatic and Gumuz populations**. Supplementary Figure 6.** Density plots of SFselect and iHS scores of neutral and real data. **Supplementary Figure 7.** PCA of the masked East African samples with a set of Europeans and Africans. **Supplementary Figure 8.** Genome-wide Manhattan plots of SFselect scores of the five populations of study. **Supplementary Figure 9.** Genome-wide Manhattan plots of the –log10(p-value) of iHS of the five populations of study.

Additional file 2: **Supplementary Tables**. **Supplementary Table 1.** The 99.99^{th} and 99.90^{th} percentile thresholds of SFselect and iHS calculated after the neutral simulations. **Supplementary Table 2.** SFselect positive selection signals shared among the five populations of study. **Supplementary Table 3.** iHS positive selection signals shared among the five populations of study**. Supplementary Table 4.** SFselect positive selection signals found in the five populations of study. **Supplementary Table 5**. iHS positive selection signals found in the five populations of study.

Additional file 3: **Supplementary Text**. Information about additional examples of shared and population-specific signals of positive selection.

Figures

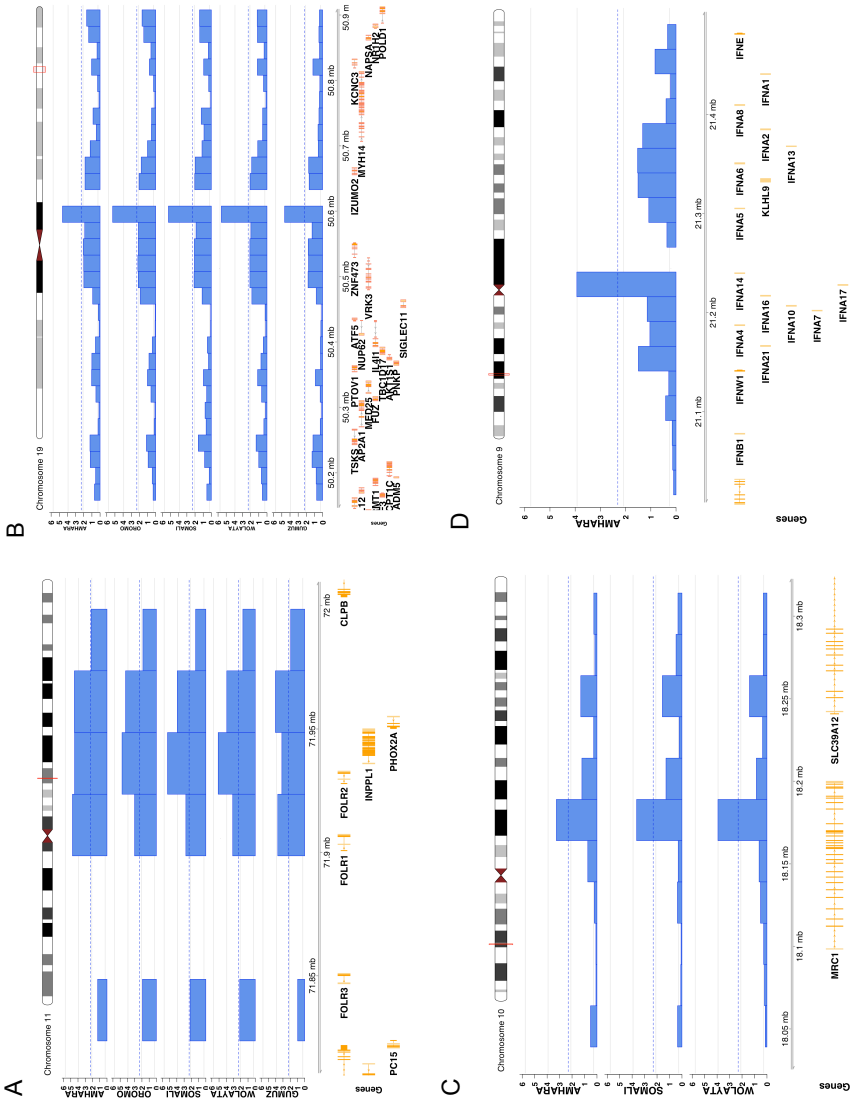


Figure 1: Genomic context of some of the significant regions from the SFselect analysis. Each bar represents to a 30 kb window, the y-axis corresponds to the SFselect score, blue dotted lines indicate the 99.99th percentile threshold obtained from extensive neutral simulations.

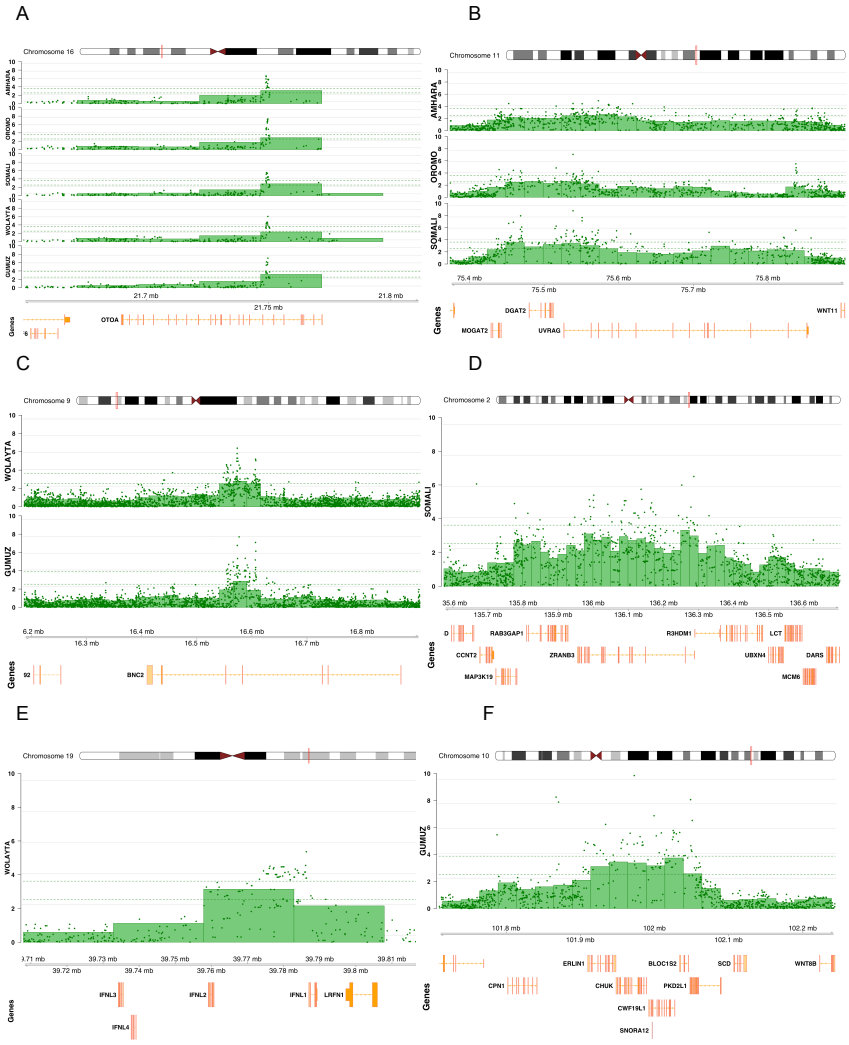


Figure 2: Genomic context of some of the significant regions from the iHS analysis. Each bar represents the absolute mean iHS of a 30 kb window, each green dot corresponds to the –log10(p-value) of a single variant. The y-axis corresponds both to the normalized mean absolute iHS score per window and the –log10(p-value) per variant. Green dotted lines indicate the 99.99th percentile thresholds of the absolute mean iHS per 30 kb windows (lower line) and the –log10(p-value) per variant (upper line) obtained from extensive neutral simulations.

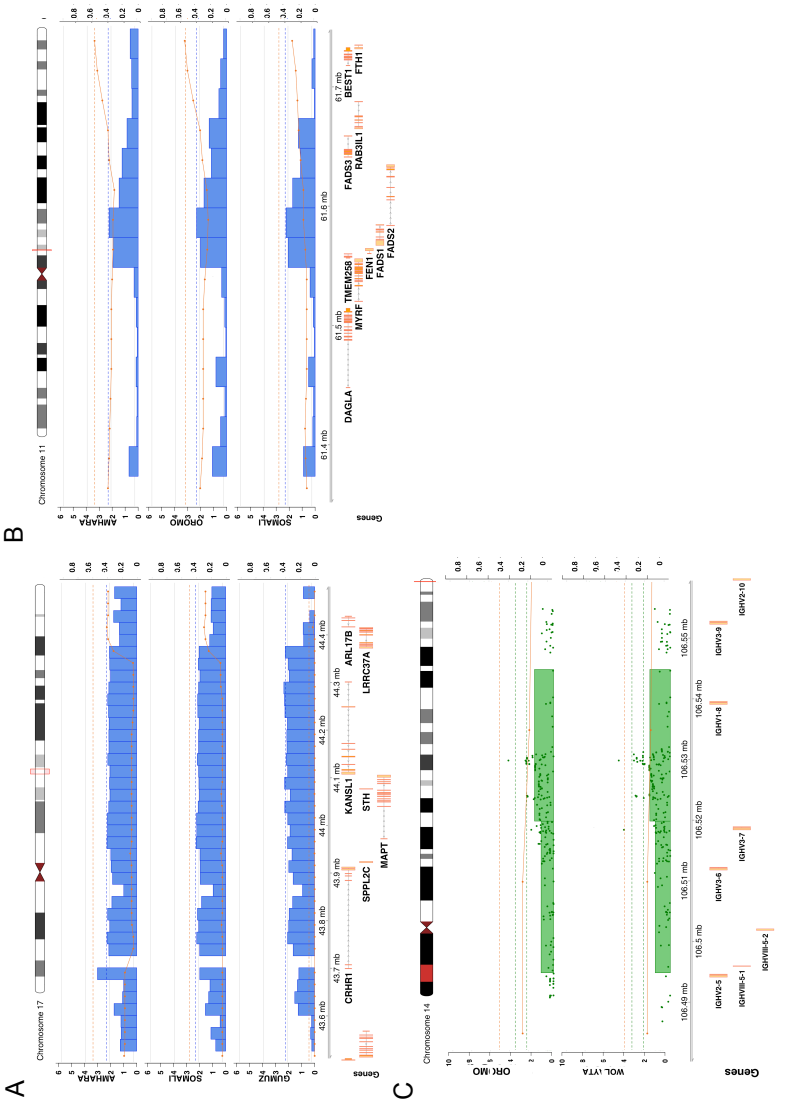


Figure 3: Genomic context of some of the regions with a strong deviation of West Asian ancestry. Orange solid lines represent the mean West Asian ancestry of the region and orange dotted lines the genome wide mean of West Asian ancestry specific of the population. Plots A and B also include in blue SFselect scores per 30 kb windows, left y-axis corresponds to SFselect scores and right y-axis the proportion of West Asian ancestry. Plot C includes in green absolute mean iHS per 30 kb region and green dots –log10(p-values) per single variant.

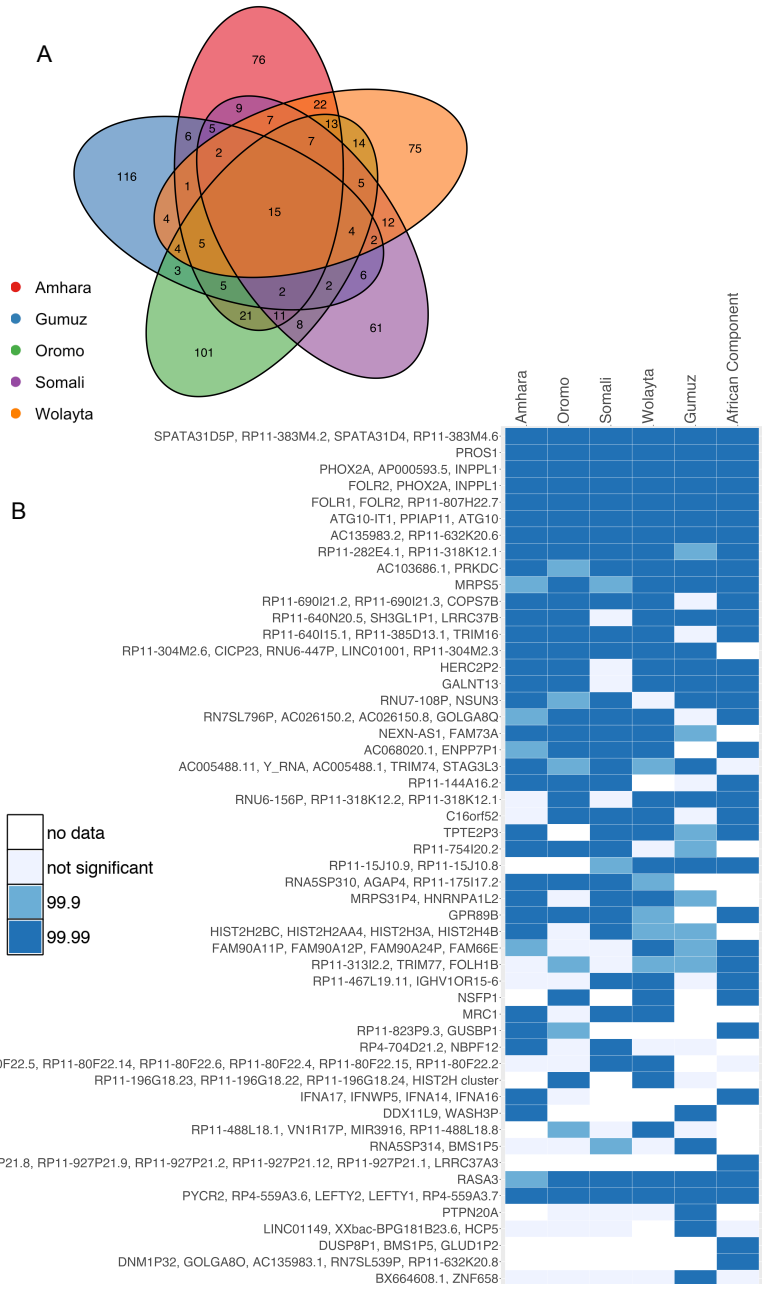


Figure 4: Common signals of positive selection between populations detected with SFselect. **A**) Venn diagram of the number of windows above the 99.99 percentile threshold shared between populations. **B**) Genes in the top 20 windows above the 99.99th percentile threshold by population. Each row represents a window and colours indicate the population corresponding significance.

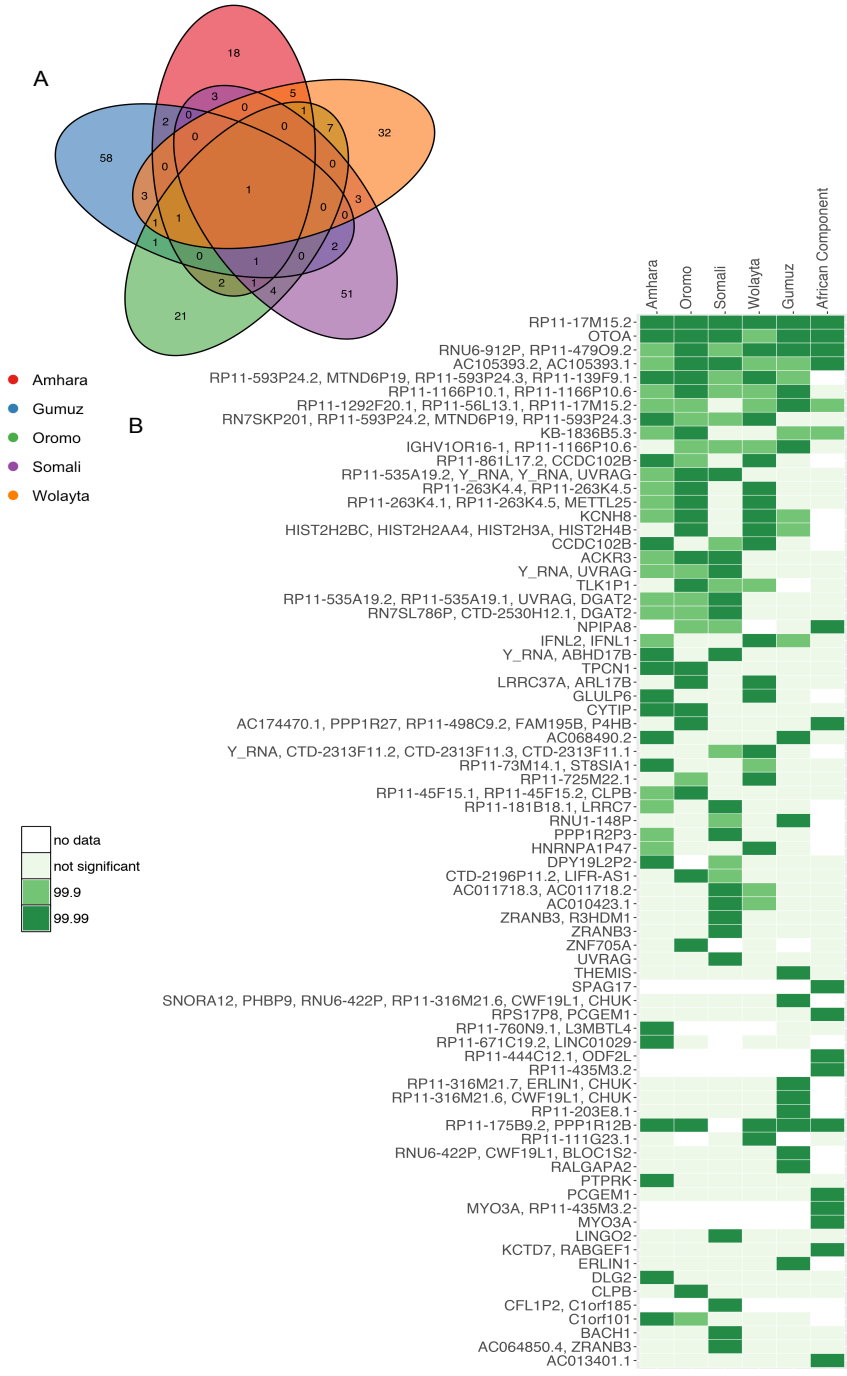


Figure 5: Common signals of positive selection between populations detected with iHS. **A**) Venn diagram of the number of windows above the 99.99th percentile threshold shared between populations. **B**) Genes in the top 20 windows above the 99.99 percentile threshold by population. Each row represents a window and colours indicate the population corresponding significance.

Tables

Population	Linguistic family	Linguistic subfamily	Number of samples
Amhara	Afroasiatic	Semitic	24
Oromo	Afroasiatic	Cushitic	24
Somali	Afroasiatic	Cushitic	24
Wolayta	Afroasiatic	Omotic	24
Gumuz	Nilo-Saharan	Nilo-Saharan	24

Table 1: Ethiopian populations, linguistic families and sample sizes included in the study.

A	Amhara (n=207)	Oromo (n=220)	Somali (n=158)	Wolayta (n=192)	Gumuz (n=182)
Amhara		79	58	72	41
Oromo			54	67	40
Somali				54	38
Wolayta					37

B	Amhara (n=35)	Oromo (n=41)	Somali (n=66)	Wolayta (n=54)	Gumuz (n=70)
Amhara		7	6	8	5
Oromo			7	11	5
Somali				4	4
Wolayta					6

Table 2: Number of shared 30 kb windows under selection between East African populations. Significant windows for each population (n) were selected after applying the 99.99 thresholds calculated after the neutral simulations. **A**) SFselect **B**) iHS.

	Whole Genome	SFselect	iHS
Amhara	0.54	0.60	0.67
Oromo	0.51	0.56	0.62
Somali	0.45	0.49	0.61
Wolayta	0.43	0.49	0.56
Gumuz	0.07	0.10	0.15

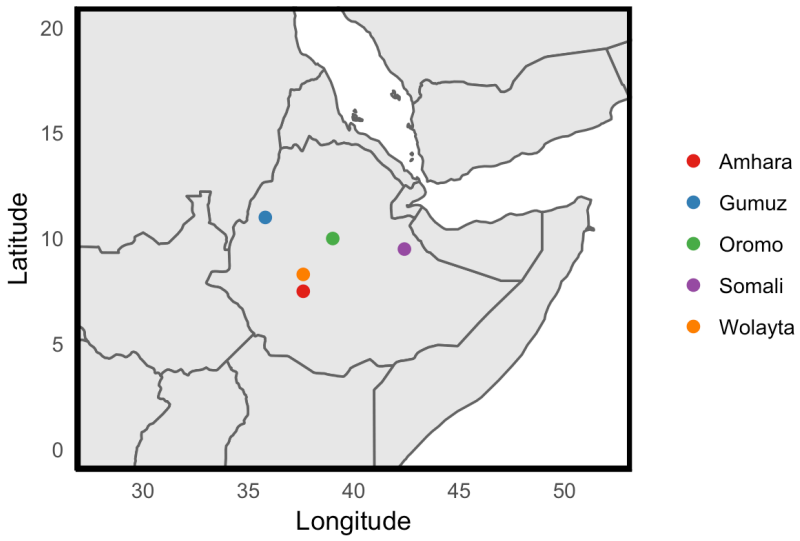
Table 3: Average West Asian ancestry proportions for each population at the genome level and among the significant windows under selection from SFselect and iHS analysis. Significant windows for each population (n) were selected after applying the 99.99 thresholds calculated after the neutral simulations.

Term	**Population**	**P-value (BH)**	**Test**
Response to virus	Amhara	0.006	SFselect
RNA surveillance	Somali	0.005	SFselect
Regulation of viral process	Wolayta	0.026	SFselect
Type I interferon binding	Amhara	0.018	SFselect
Type I interferon production	Gumuz	0.020	SFselect
Positive regulation of interferon-gamma production	Gumuz	0.027	iHS
B-cell activation and regulation of immunoglobulin production	Amhara, Somali	0.017; 0.01	SFselect
Regulation of immunoglobulin production	Amhara	0.018	SFselect
Hepatitis B	Amhara	0.03	SFselect
Tuberculosis	Amhara	0.047	SFselect
Measles	Amhara	0.04	SFselect
Leishmaniasis	Gumuz	0.049	iHS
Lupus erythematous	Somali, Wolayta, Gumuz	0.0009; 0.02; 0.026	iHS
Folic acid containing compound metabolic process	Amhara, Somali, Wolayta	0.013; 0.02; 0.02	SFselect
Folic acid metabolic process	Amhara, Wolayta	0.019; 0.01	SFselect
Metabolism of folate	Amhara	0.017	SFselect
Pterines and folate biosynthesis	Amhara	0.02	SFselect
Cellular response to UV-B	Amhara, Somali	0.002; 0.0019	SFselect
Cellular response to UV	Amhara, Somali	0.018; 0.0027	SFselect
Cellular response to radiation	Amhara, Somali	0.01; 0.001	SFselect
Cellular response to vitamin D	Amhara	0.02	SFselect
Bone mineralization	Somali	0.03	SFselect
Osteoclast differentiation	Wolayta	0.03	SFselect
Negative regulation of cardiac muscle tissue development	Gumuz	0.02	SFselect

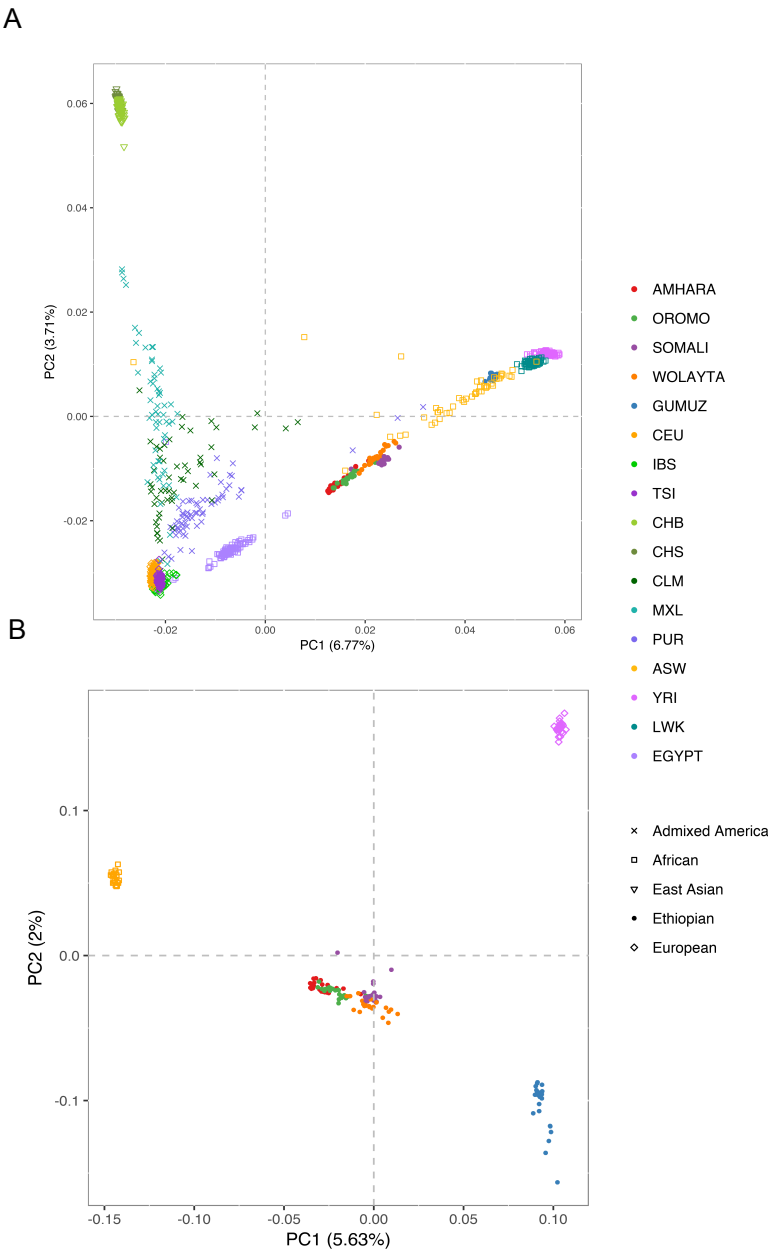
Negative regulation of striated muscle tissue development	Gumuz	0.037	SFselect
Muscle fibre development	Somali	0.036	iHS

Table 4: Signatures of polygenic adaptation through functional enrichment analysis. We have listed the most relevant terms of the analysis. The two lists of genes used for the analysis were taken from the significant windows under putative positive selection for SFselect and iHS. The genes used for this analysis are listing the genes with significant SFselect or iHS scores. A biological term was considered significant if the p-value after a Benjamini-Hochbert (BH) correction was below an alpha value of 0.05.

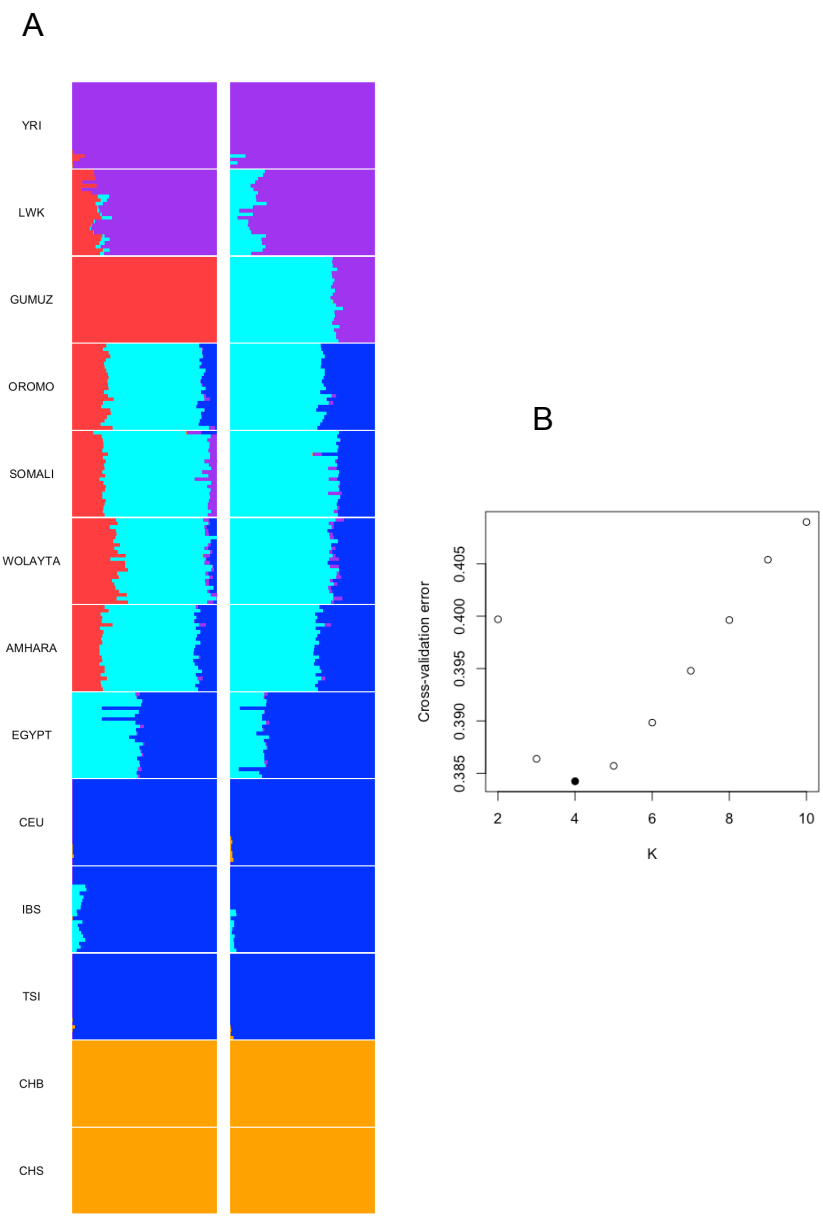
ADDITIONAL FILE 1



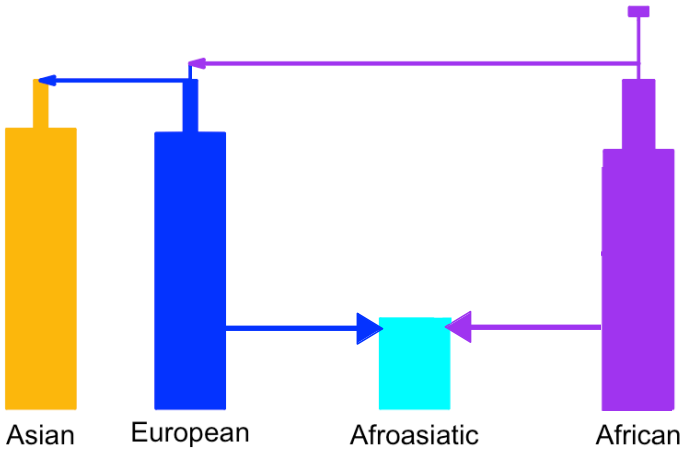
Supplementary figure 1: Location of the five sampled populations.



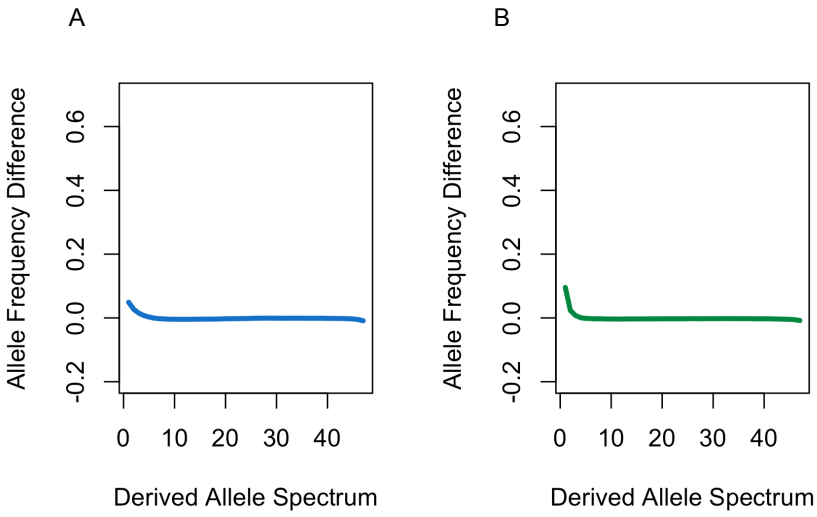
Supplementary figure 2: Principal component analysis of the five East African populations. **A**) In a worldwide context, the analysis included a set of populations from the 1000 Genomes Project. **B**) In a local context, only including the East African samples, a set of Europeans and Western Africans.



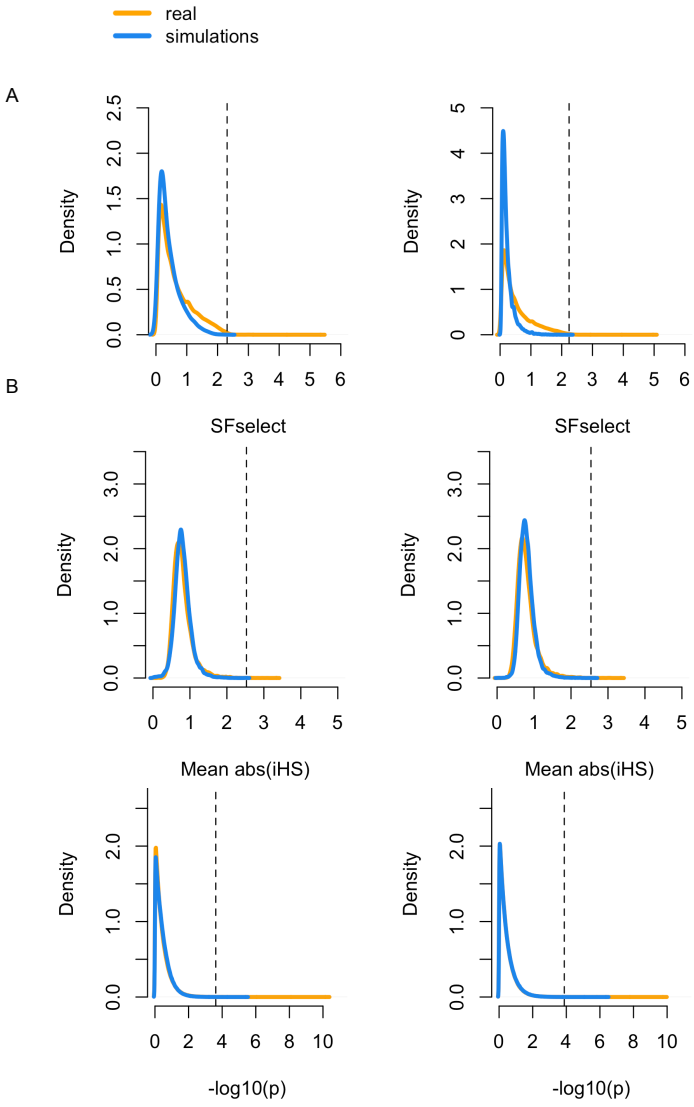
Supplementary figure 3: A) ADMIXTURE analysis of the Ethiopian samples and a set of worldwide populations. B) The lowest cross-validation error was obtained with K=4 and K=5 components. The resulting ancestry components describe Western Africa (purple), North-East Africa (light blue), East Africa (red), Europe (dark blue), East Asia (orange).



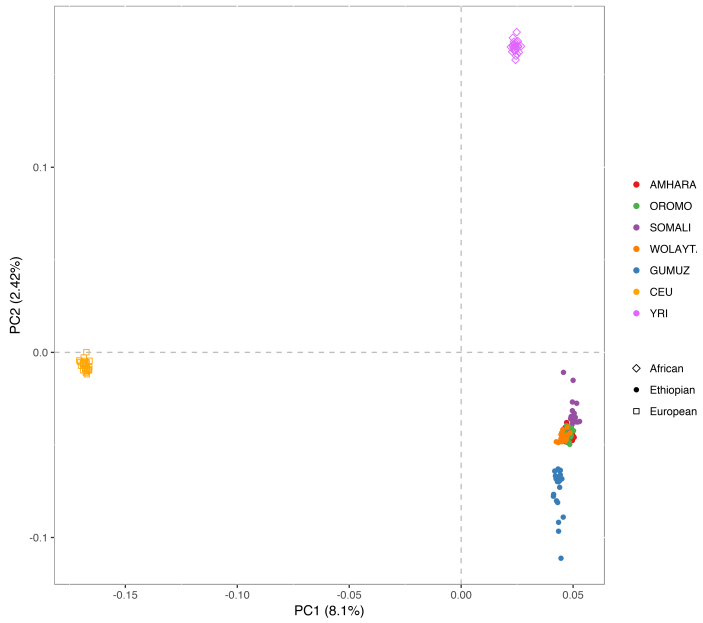
Supplementary Figure 4: Schematic representation of the demographic model used to simulate neutral sequences. Parameters where obtained from Jouganous et al 2017.



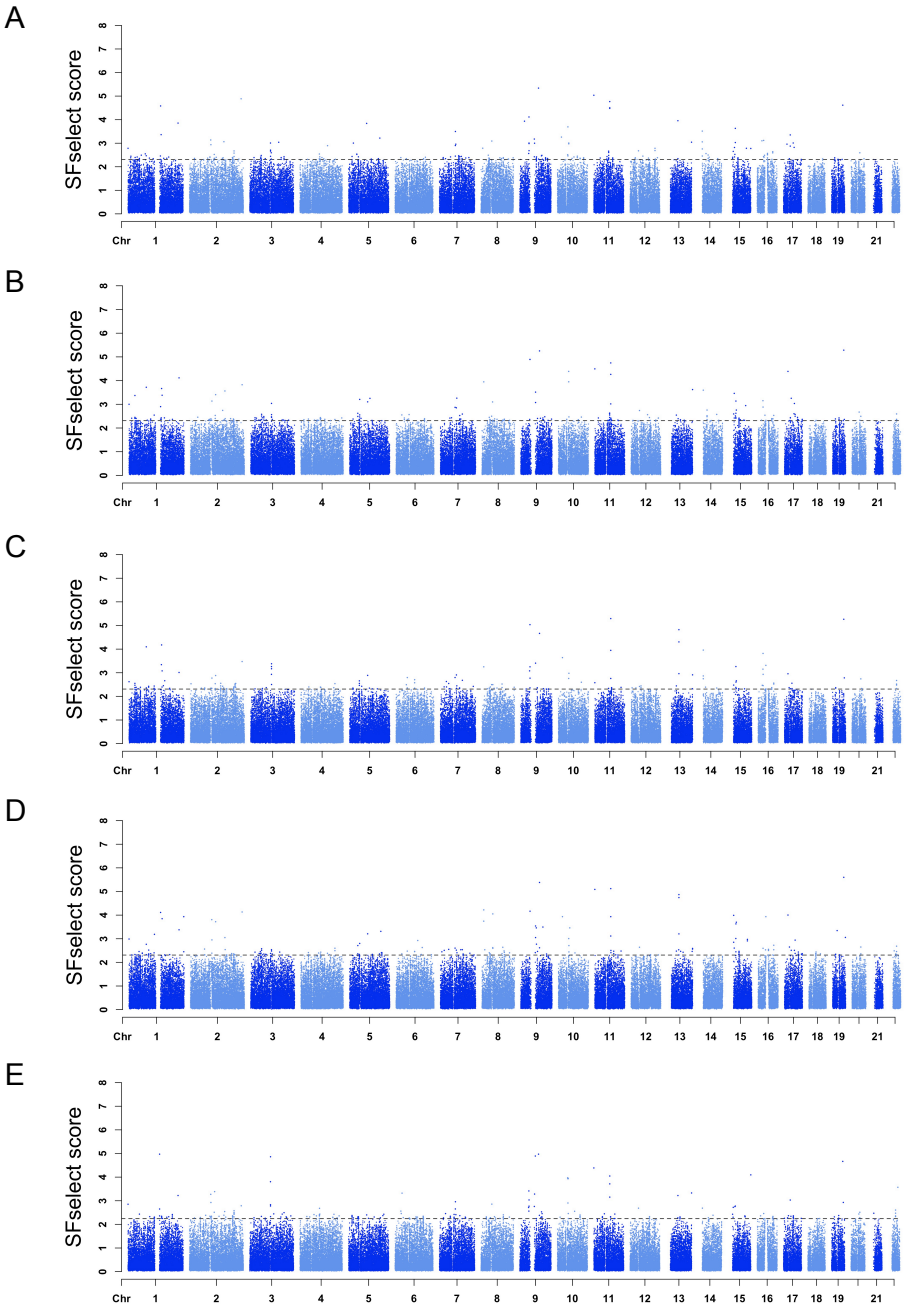
Supplementary Figure 5: Relative site frequency spectrum A) Afroasiatic B) Gumuz. The difference of derived allele frequency between neutral simulations and real data. The neutral simulations fit the real data for both Afroasiatic and Gumuz data since there are no main differences between the derived site frequency spectrum between real and simulated data. The slight increase of singletons in simulations can be explained by the lack of coverage in real data and inaccuracies of the demographic model.



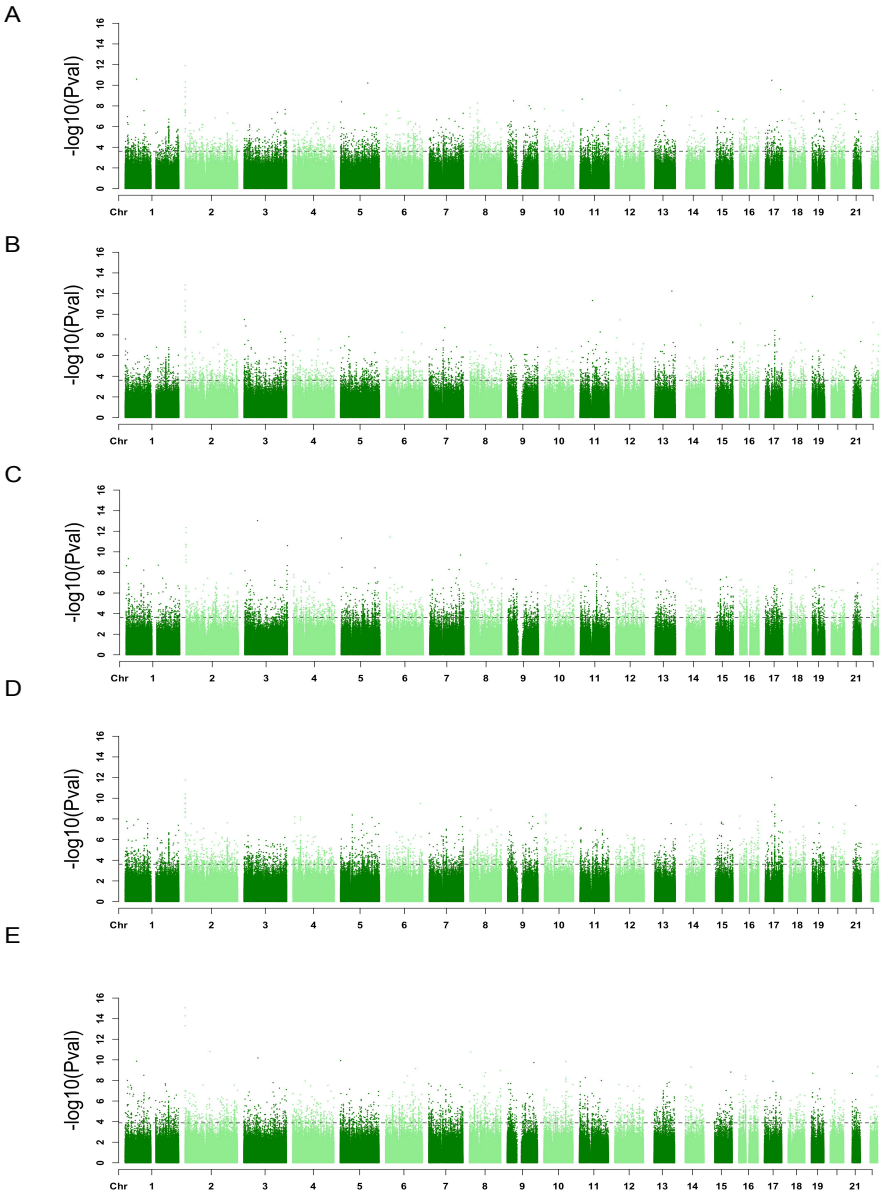
Supplementary Figure 6: Density plots of A) SFselect scores B) mean absolute iHS score in 30 kb windows C) –log10(p) per SNP, of real and simulated data for Afroasiatic (left) and Gumuz (right). The dashed line represents the 99.99th percentile threshold calculated after the simulations. Neutral and real data.



Supplementary Figure 7: PCA of the masked East African samples with a set of Europeans and Africans. The Afroasiatic samples now cluster next to the unadmixed Gumuz in comparison with Figure 2 where they clustered between Europeans and the Gumuz.



Supplementary Figure 8: Genome-wide Manhattan plots of SFselect scores A) Amhara B) Oromo C) Somali D) Wolayta E) Gumuz. Each point represents the SFselect score of a 30 kb window and the dashed black lines the 99.99th percentile threshold obtained after the neutral simulations.



Supplementary Figure 9: Genome-wide Manhattan plots of the –log10(p-value) of iHS for A) Amhara B) Oromo C) Somali D) Wolayta E) Gumuz. Each point represents the –log10(p-value) of the iHS score of a SNP normalised by allele frequency bins of 0.05. The horizontal dashed black lines represent the 99.99th percentile threshold obtained after the neutral simulations.

ADDITIONAL FILE 2

	SFselect		Mean iHS per window		-log10(p) iHS SNP based	
Percentile	99.99	99.90	99.99	99.90	99.99	99.90
Afroasiatic	2.31 (n= 20424)	1.93 (n= 20424)	2.53 (n=15809)	1.95 (n=15809)	3.62 (n=738269)	2.77 (n=738269)
Gumuz	2.24 (n= 24000)	1.56 (n= 24000)	2.54 (n=19240)	2.05 (n=19240)	3.88 (n=1055794)	2.98 (n= 1055794)

Supplementary Table 1: The 99.99th and 99.90th percentile thresholds of SFselect and iHS calculated after the neutral simulations. For SFselect and mean iHS we report in parenthesis the number of 30 kb simulated windows and for the SNP based iHS the number of SNPs used to calculate the thresholds.

Chr	Start	End	Amhara	Oromo	Somali	Wolayta	Gumuz	Genes	Type of gene
chr9	84515910	84545910	5.34	5.25	4.67	5.38	4.97	*RP11-383M4.2, RP11-383M4.6, SPATA31D4, SPATA31D5P*	lincRNA, lincRNA, PS,PS
chr11	123574	153574	5.04	4.50	2.58	5.09	4.39	*LINC01001, RP11-304M2.6, CICP23, RNU6-447P*	lincRNA, lincRNA, PS, snRNA
chr11	71898574	71928574	4.77	3.02	2.76	3.12	3.71	*FOLR1, FOLR2*	PC
chr19	50582864	50612864	4.61	4.67	5.28	5.26	5.60	-	-
chr11	71948574	71978574	4.50	4.26	3.95	3.93	4.05	*INPPL1, PHOX2A, AP000593.5*	PC, PC, PS
chr11	71923574	71953574	4.48	4.75	5.29	5.12	3.15	*FOLR2, INPPL1, PHOX2A*	PC
chr1	226096808	226126808	3.85	4.12	3.01	3.38	3.22	*LEFTY1, LEFTY2, PYCR2*	PC
chr5	81280629	81310629	3.84	3.10	2.89	3.21	2.40	*ATG10, PPIAP11*	PC, PS
chr15	32754330	32784330	3.63	2.76	2.66	3.70	2.77	*AC135983.2, RP11-632K20.6*	PC, PS
chr3	93598294	93628294	3.01	2.47	3.26	2.54	3.81	*PROS1*	PC
chr2	96448695	96478695	2.94	3.14	2.77	2.95	2.93	*LINC00342, AC008268.2*	lincRNA, PS
chr1	571808	601808	2.78	3.00	2.62	2.99	2.85	*RP5-857K21.4*	lincRNA
chr20	39678005	39708005	2.59	2.49	2.74	2.65	2.51	*TOP1*	PC
chr15	30704330	30734330	2.52	2.62	2.38	2.45	2.75	*GOLGA8R, RP11-382B18.5*	PC, PS
chr4	89360998	89390998	2.42	2.40	2.38	2.43	2.44	*HERC5, HERC6*	PC
chr1	184021808	184051808	2.41	2.32	2.34	2.33	2.32	*TSEN15*	PC

Supplementary Table 2: SFselect positive selection signals shared among the five populations of study. All the reported SFselect scores of the 30 kb windows are above the 99.99 percentile threshold calculated after the simulations. We also report the genes within the window and the type of gene (PC: protein coding gene, PS: pseudogene, lincRNA: long interspersed non-coding RNA, small nuclear RNA: snRNA).

			Amhara		Oromo		Somali		Wolayta		Gumuz			
Chr	Start	End	mean	maxP	mean	maxP	mean	maxP	mean	maxP	mean	maxP	Genes	Type of gene
chr16	21746755	21776755	3.12	6.59	2.85	7.32	2.82	5.50	2.32	6.02	3.20	7.05	*OTOA*	PC
chr8	36136376	36166376	3.09	8.28	2.88	5.62	2.42	5.16	3.17	6.46	2.29	3.98	*MTND6P19, RP11-139F9.1, RP11-593P24.2, RP11-593P24.3*	PS
chr16	32296755	32326755	2.97	6.71	2.98	5.29	3.52	6.31	3.05	5.61	3.56	8.45	*RP11-17M15.2*	PS
chr16	33496755	33526755	2.54	4.87	2.53	5.34	2.88	5.11	2.46	4.70	2.06	5.21	*BMS1P8*	PS
chr7	65258619	65288619	2.45	5.09	2.66	7.49	2.13	5.39	2.66	6.28	2.74	5.27	*RNU6-912P, RP11-479O9.2*	snRNA, PS
chr16	32121755	32151755	2.35	2.88	2.53	5.37	2.08	3.06	2.62	3.38	2.89	3.67	*HERC2P4, RP11-1166P10.7*	PS
chr16	32321755	32351755	2.29	6.71	2.06	5.29	2.29	6.12	2.19	5.61	2.05	5.61	*RP11-17M15.2*	PS
chr4	510998	540998	2.24	5.67	2.30	3.71	2.68	6.25	2.12	3.54	2.35	4.53	*PIGG*	PC
chr2	398695	428695	2.14	11.89	2.95	12.85	2.55	12.36	2.36	11.78	2.45	15.05	*AC105393.1, AC105393.2*	lincRNA
chr16	31996755	32026755	1.95	2.23	2.61	2.59	2.25	3.08	2.37	3.01	3.37	4.47	*RP11-1166P10.1, RP11-1166P10.6*	PS

Supplementary Table 3: iHS positive selection signals shared among the five populations of study. For each population the mean iHS score for the 30 kb windows is reported and the highest –log(p-value) of a variant within the window. The windows were selected for comparison if their mean iHS score was above the 99.99 percentile threshold in at least one population and in the 99.9 threshold in the rest. In gray, the only shared window that is in all populations with mean iHS scores above the 99.99 threshold. We also report the genes within the window and the type of gene (PC: protein coding gene, PS: pseudogene, lincRNA: long interspersed non-coding RNA, small nuclear RNA: snRNA).

	Chr	Start	Stop	SFselect	Genes
	chr2	232673695	232703695	4.89	*COPS7B*
	chr11	71898574	71928574	4.77	*FOLR1, FOLR2*
	chr1	147396808	147426808	4.58	*GPR89B*
	chr11	71948574	71978574	4.50	*INPPL1, PHOX2A*
	chr11	71923574	71953574	4.48	*FOLR2, INPPL1, PHOX2A*
*	chr9	21215910	21245910	3.93	*IFNA16, IFNA17, IFNA14*
	chr1	226096808	226126808	3.85	*LEFTY1, RP4-559A3.7, PYCR2, LEFTY2*
	chr5	81280629	81310629	3.84	*ATG10*
	chr15	32754330	32784330	3.63	*AC135983.2*
	chr7	72433619	72463619	3.50	*TRIM74, AC005488.1*
	chr1	149821808	149851808	3.36	*HIST2H2AA4, HIST2H3A, HIST2H4B*
	chr17	30351088	30381088	3.35	*LRRC37B*
	chr10	18164034	18194034	3.26	*MRC1*
	chr5	140705629	140735629	3.22	*PCDHGA1, PCDHGA2, PCDHGA3, PCDHGB1, PCDHGA4*
	chr8	48786376	48816376	3.09	*PRKDC*
	chr2	155023695	155053695	3.06	*GALNT13*
	chr3	130698294	130728294	3.04	*ATP2C1*
	chr17	43676088	43706088	3.01	*CRHR1*
	chr3	93598294	93628294	3.01	*PROS1*
*	chr10	51214034	51244034	3.01	*AGAP8*
	chr10	51189034	51219034	2.97	*FAM21D*
	chr17	15526088	15556088	2.96	*RP11-385D13.1, TRIM16*
	chr7	74158619	74188619	2.96	*GTF2I, NCF1*
	chr17	30326088	30356088	2.88	*SUZ12, LRRC37B*
	chr17	47876088	47906088	2.82	*KAT7*
*	chr12	112399817	112429817	2.78	*TMEM116*
	chr3	93773294	93803294	2.71	*ARL13B, DHFRL1, NSUN3*
*	chr12	112374817	112404817	2.69	*TMEM116*
	chr2	200773695	200803695	2.67	*C2orf69, TYW5*
	chr3	93798294	93828294	2.66	*NSUN3*

Supplementary Table 4A: SFselect positive selection signals found in Amhara population. All the 30 kb window scores reported are above the 99.99 percentile threshold calculated after the simulations. Only protein coding genes are reported.

* : Population-specific positive selection signals. These signals are above the 99.99 percentile threshold calculated after the simulations that are not found among the 99 percentile of the rest of the populations of study.

	Chr	Start	Stop	SFselect	Genes
	chr11	71923574	71953574	4.75	*FOLR2, INPPL1, PHOX2A*
	chr17	15526088	15556088	4.39	*RP11-385D13.1, TRIM16*
	chr10	46339034	46369034	4.39	*AGAP4*
	chr11	71948574	71978574	4.26	*INPPL1, PHOX2A*
	chr1	226096808	226126808	4.12	*LEFTY1, RP4-559A3.7, PYCR2, LEFTY2*
	chr10	46489034	46519034	3.95	*PTPN20A*
	chr2	232673695	232703695	3.83	*COPS7B*
	chr1	78321808	78351808	3.72	*FAM73A*
	chr1	147396808	147426808	3.67	*GPR89B*
	chr13	114717755	114747755	3.62	*RASA3*
	chr2	155023695	155053695	3.56	*GALNT13*
	chr2	113198695	113228695	3.41	*RGPD8*
	chr1	149796808	149826808	3.38	*HIST2H4A, HIST2H3C, HIST2H2AA3, HIST2H2AA4, HIST2H3A, HIST2H4B*
	chr1	27096808	27126808	3.37	*ARID1A, PIGV*
	chr7	74158619	74188619	3.26	*GTF2I, NCF1*
	chr17	30351088	30381088	3.25	*LRRC37B*
	chr5	45505629	45535629	3.21	*HCN1*
	chr15	30829330	30859330	3.14	*GOLGA8Q*
	chr5	81280629	81310629	3.10	*ATG10*
	chr3	93698294	93728294	3.04	*ARL13B*
	chr11	71898574	71928574	3.02	*FOLR1, FOLR2*
	chr15	74954330	74984330	2.95	*EDC3*
	chr16	22021755	22051755	2.88	*C16orf52*
	chr14	36169850	36199850	2.76	*RALGAPA1*
	chr15	32754330	32784330	2.76	*AC135983.2*
	chr2	144973695	145003695	2.74	*GTDC1*
*	chr20	32178005	32208005	2.67	*CBFA2T2*
	chr11	66898574	66928574	2.63	*KDM2A*
	chr15	30704330	30734330	2.62	*GOLGA8R*
	chr5	36880629	36910629	2.62	*NIPBL*

Supplementary Table 4B: SFselect positive selection signals found in Oromo population. All the 30 kb window scores reported are above the 99.99 percentile threshold calculated after the simulations. Only protein coding genes are reported.

* : Population-specific positive selection signals. These signals are above the 99.99 percentile threshold calculated after the simulations that are not found among the 99 percentile of the rest of the populations of study.

	Chr	Start	Stop	SFselect	Genes
	chr11	71923574	71953574	5.29	*FOLR2, INPPL1, PHOX2A*
	chr13	53167755	53197755	4.30	*HNRNPA1L2*
	chr1	147396808	147426808	4.18	*GPR89B*
	chr1	78321808	78351808	4.10	*FAM73A*
	chr11	71948574	71978574	3.95	*INPPL1, PHOX2A*
	chr16	22021755	22051755	3.82	*C16orf52*
	chr10	18164034	18194034	3.64	*MRC1*
	chr2	232673695	232703695	3.48	*COPS7B*
	chr3	93798294	93828294	3.38	*NSUN3*
	chr1	146446808	146476808	3.34	*NBPF12*
	chr15	30829330	30859330	3.26	*GOLGA8Q*
	chr3	93598294	93628294	3.26	*PROS1*
	chr3	93773294	93803294	3.17	*ARL13B, DHFRL1, NSUN3*
	chr1	149821808	149851808	3.08	*HIST2H2AA4, HIST2H3A, HIST2H4B*
	chr1	226096808	226126808	3.01	*LEFTY1, RP4-559A3.7, PYCR2, LEFTY2*
	chr17	15526088	15556088	2.96	*RP11-385D13.1, TRIM16*
*	chr16	22396755	22426755	2.93	*CDR2*
	chr13	114717755	114747755	2.91	*RASA3*
	chr7	72433619	72463619	2.91	*TRIM74, AC005488.1*
	chr5	81280629	81310629	2.89	*ATG10*
	chr2	113198695	113228695	2.88	*RGPD8*
	chr6	50655815	50685815	2.80	*TFAP2D*
	chr19	53032864	53062864	2.78	*ZNF808, ZNF701*
	chr10	46339034	46369034	2.76	*AGAP4*
	chr11	71898574	71928574	2.76	*FOLR1, FOLR2*
	chr20	39678005	39708005	2.74	*TOP1*
	chr6	83805815	83835815	2.71	*DOPEY1*
*	chr7	99258619	99288619	2.68	*CYP3A5*
	chr22	32476780	32506780	2.67	*SLC5A1*
	chr1	161546808	161576808	2.66	*FCGR3A, FCGR2B*

Supplementary Table 4C: SFselect positive selection signals found in Somali population. All the 30 kb window scores reported are above the 99.99 percentile threshold calculated after the simulations. Only protein coding genes are reported.

* : Population-specific positive selection signals. These signals are above the 99.99 percentile threshold calculated after the simulations that are not found among the 99 percentile of the rest of the populations of study.

	Chr	Start	Stop	SFselect	Genes
	chr11	71923574	71953574	5.12	*FOLR2, INPPL1, PHOX2A*
	chr13	53167755	53197755	4.74	*HNRNPA1L2*
	chr2	232673695	232703695	4.13	*COPS7B*
	chr8	48786376	48816376	4.05	*PRKDC*
	chr17	15526088	15556088	4.00	*RP11-385D13.1, TRIM16*
	chr10	18164034	18194034	3.93	*MRC1*
	chr11	71948574	71978574	3.93	*INPPL1, PHOX2A*
	chr1	149796808	149826808	3.85	*HIST2H4A, HIST2H3C, HIST2H2AA3, HIST2H2AA4, HIST2H3A, HIST2H4B*
	chr2	95748695	95778695	3.81	*MRPS5*
	chr2	113198695	113228695	3.72	*RGPD8*
	chr15	32754330	32784330	3.70	*AC135983.2*
	chr15	30829330	30859330	3.63	*GOLGA8Q*
	chr9	99740910	99770910	3.50	*HIATL2*
	chr10	51189034	51219034	3.46	*FAM21D*
	chr1	226096808	226126808	3.38	*LEFTY1, RP4-559A3.7, PYCR2, LEFTY2*
	chr5	140705629	140735629	3.31	*PCDHGA1, PCDHGA2, PCDHGA3, PCDHGB1, PCDHGA4*
	chr5	81280629	81310629	3.21	*ATG10*
*	chr1	115121808	115151808	3.19	*BCAS2, DENND2C*
	chr11	71898574	71928574	3.12	*FOLR1, FOLR2*
*	chr19	57907864	57937864	3.06	*AC003002.4, ZNF548, AC003002.6, AC004076.7, ZNF17*
	chr2	155023695	155053695	3.05	*GALNT13*
	chr10	46489034	46519034	3.01	*PTPN20A*
	chr17	47876088	47906088	2.94	*KAT7*
	chr5	45355629	45385629	2.80	*HCN1*
	chr1	78321808	78351808	2.77	*FAM73A*
	chr16	70346755	70376755	2.73	*DDX19B, RP11-529K1.3*
	chr5	36955629	36985629	2.71	*NIPBL*
	chr22	32476780	32506780	2.69	*SLC5A1*
	chr20	39678005	39708005	2.65	*TOP1*
	chr6	118030815	118060815	2.63	*NUS1*

Supplementary Table 4D: SFselect positive selection signals found in Wolayta population. All the 30 kb window scores reported are above the 99.99 percentile threshold calculated after the simulations. Only protein coding genes are reported.

* : Population-specific positive selection signals. These signals are above the 99.99 percentile threshold calculated after the simulations that are not found among the 99 percentile of the rest of the populations of study.

	Chr	Start	Stop	SFselect	Genes
	chr3	93798294	93828294	4.86	*NSUN3*
	chr11	71948574	71978574	4.05	*INPPL1, PHOX2A*
	chr10	46489034	46519034	3.97	*PTPN20A*
	chr10	46539034	46569034	3.93	*PTPN20A*
	chr3	93598294	93628294	3.81	*PROS1*
	chr11	71898574	71928574	3.71	*FOLR1, FOLR2*
*	chr9	40790910	40820910	3.41	*ZNF658*
	chr2	113198695	113228695	3.38	*RGPD8*
	chr13	114717755	114747755	3.33	*RASA3*
	chr2	95748695	95778695	3.27	*MRPS5*
	chr1	226096808	226126808	3.22	*LEFTY1, RP4-559A3.7, PYCR2, LEFTY2*
	chr11	71923574	71953574	3.15	*FOLR2, INPPL1, PHOX2A*
	chr17	30351088	30381088	3.03	*LRRC37B*
	chr7	72433619	72463619	2.96	*TRIM74, AC005488.1*
	chr19	53032864	53062864	2.93	*ZNF808, ZNF701*
	chr8	48786376	48816376	2.86	*PRKDC*
*	chr3	93723294	93753294	2.83	*ARL13B, STX19*
*	chr2	233223695	233253695	2.79	*ALPP*
	chr3	93748294	93778294	2.78	*ARL13B, DHFRL1*
	chr15	32754330	32784330	2.77	*AC135983.2*
	chr15	30704330	30734330	2.75	*GOLGA8R*
	chr4	88110998	88140998	2.68	*KLHL8*
	chr22	32451780	32481780	2.61	*SLC5A1*
	chr2	200798695	200828695	2.58	*C2orf69, TYW5, C2orf47*
*	chr9	40865910	40895910	2.56	*ZNF658*
	chr6	25955815	25985815	2.56	*TRIM38*
	chr10	77464034	77494034	2.55	*C10orf11*
*	chr2	158448695	158478695	2.54	*ACVR1C*
	chr9	97990910	98020910	2.53	*FANCC*
	chr2	111373695	111403695	2.51	*BUB1*

Supplementary Table 4E: SFselect positive selection signals found in Gumuz population. All the 30 kb window scores reported are above the 99.99 percentile threshold calculated after the simulations. Only protein coding genes are reported.

* : Population-specific positive selection signals. These signals are above the 99.99 percentile threshold calculated after the simulations that are not found among the 99 percentile of the rest of the populations of study.

	Chr	Start	Stop	meaniHS	maxP	Genes
	chr18	6247301	6277301	4.14	4.67	*L3MBTL4*
	chr1	202396808	202426808	3.57	6.71	*PPP1R12B*
	chr18	66647301	66677301	3.34	6.67	*CCDC102B*
	chr16	21746755	21776755	3.12	6.58	*OTOA*
	chr9	74490910	74520910	3.05	4.08	*ABHD17B*
	chr1	244721808	244751808	2.99	5.66	*C1orf101*
	chr18	66722301	66752301	2.98	8.46	*CCDC102B*
	chr1	244696808	244726808	2.96	5.73	*C1orf101*
	chr6	128505815	128535815	2.92	4.82	*PTPRK*
	chr11	85173574	85203574	2.92	5.16	*DLG2*
	chr12	22524817	22554817	2.89	9.50	*ST8SIA1*
	chr12	113674817	113704817	2.86	5.59	*TPCN1*
	chr1	244671808	244701808	2.77	5.18	*C1orf101*
	chr6	157480815	157510815	2.73	4.75	*ARID1B*
*	chr6	136905815	136935815	2.71	6.33	*MAP3K5*
	chr2	158323695	158353695	2.71	4.83	*CYTIP*
	chr12	113699817	113729817	2.68	4.06	*TPCN1*
	chr11	75573574	75603574	2.66	4.31	*UVRAG*
*	chr4	47210998	47240998	2.63	4.79	*GABRB1*
	chr22	46701780	46731780	2.58	4.62	*GTSE1, TRMU*
	chr9	40815910	40845910	2.54	5.23	*ZNF658*
	chr11	85198574	85228574	2.53	4.76	*DLG2*

Supplementary Table 5A: iHS positive selection signals found in Amhara population. The mean iHS score for the 30 kb windows is reported and the highest –log(p-value) of a variant within the window. All the reported iHS scores are above the reported 99.99 percentile thresholds calculated after the simulations. Only protein coding genes are reported.

* : Population-specific positive selection signals. These signals are above the 99.99 percentile threshold calculated after the simulations that are not found among the 99 percentile of the rest of the populations of study.

	Chr	Start	Stop	meaniHS	maxP	Genes
	chr3	19248294	19278294	3.33	7.46	*KCNH8*
	chr1	202396808	202426808	3.14	6.77	*PPP1R12B*
	chr17	79776088	79806088	3.08	4.25	*FAM195B, AC174470.1, PPP1R27, P4HB*
	chr2	237448695	237478695	3.07	5.37	*ACKR3*
	chr12	8299817	8329817	2.96	2.77	*ZNF705A*
	chr11	72023574	72053574	2.94	6.60	*CLPB*
	chr3	19298294	19328294	2.91	4.79	*KCNH8*
	chr2	158323695	158353695	2.87	6.16	*CYTIP*
	chr11	72048574	72078574	2.86	4.86	*CLPB*
	chr16	21746755	21776755	2.85	7.32	*OTOA*
	chr12	82849817	82879817	2.82	7.11	*METTL25*
	chr17	44351088	44381088	2.72	8.42	*ARL17B, LRRC37A*
	chr3	19273294	19303294	2.69	4.79	*KCNH8*
	chr11	75473574	75503574	2.60	3.09	*DGAT2*
	chr11	75523574	75553574	2.60	7.06	*UVRAG*
	chr3	19173294	19203294	2.59	5.31	*KCNH8*
	chr12	45674817	45704817	2.58	3.82	*ANO6*
	chr12	113674817	113704817	2.56	3.62	*TPCN1*
	chr1	149821808	149851808	2.55	3.95	*HIST2H2AA4, HIST2H3A, HIST2H4B*
	chr12	64224817	64254817	2.55	5.39	*SRGAP1*
	chr10	70414034	70444034	2.54	4.44	*TET1*
*	chr3	164673294	164703294	2.54	4.79	*SI*
	chr15	44879330	44909330	2.54	4.07	*SPG11*
	chr12	82799817	82829817	2.54	5.21	*METTL25*

Supplementary Table 5B: iHS positive selection signals found in Oromo population. The mean iHS score for the 30 kb windows is reported and the highest –log(p-value) of a variant within the window. All the reported iHS scores are above the reported 99.99 percentile thresholds calculated after the simulations. Only protein coding genes are reported.

* : Population-specific positive selection signals. These signals are above the 99.99 percentile threshold calculated after the simulations that are not found among the 99 percentile of the rest of the populations of study.

	Chr	Start	Stop	meaniHS	maxP	Genes
	chr11	75448574	75478574	3.52	7.94	*DGAT2*
	chr11	75523574	75553574	3.40	8.77	*UVRAG*
	chr2	136248695	136278695	3.32	5.99	*ZRANB3*
	chr11	75498574	75528574	3.22	6.24	*DGAT2, UVRAG*
	chr21	30808956	30838956	3.17	5.72	*BACH1*
	chr1	51596808	51626808	3.13	3.72	*C1orf185*
	chr1	70471808	70501808	3.12	6.91	*LRRC7*
	chr2	135973695	136003695	3.10	5.40	*ZRANB3*
	chr11	75548574	75578574	3.04	7.65	*UVRAG*
*	chr2	136273695	136303695	2.98	6.50	*ZRANB3, R3HDM1*
	chr11	75723574	75753574	2.96	3.80	*UVRAG*
	chr2	136073695	136103695	2.94	5.07	*ZRANB3*
*	chr9	28615910	28645910	2.94	5.38	*LINGO2*
*	chr9	28640910	28670910	2.94	5.08	*LINGO2*
	chr2	237448695	237478695	2.92	4.86	*ACKR3*
	chr2	136023695	136053695	2.91	4.86	*ZRANB3*
*	chr19	3232864	3262864	2.88	6.38	*CELF5*
*	chr15	42654330	42684330	2.86	5.87	*CAPN3*
	chr12	44674817	44704817	2.85	6.04	*TMEM117*
	chr9	40815910	40845910	2.85	5.65	*ZNF658*
	chr11	75473574	75503574	2.83	3.54	*DGAT2*
	chr1	51696808	51726808	2.83	4.72	*RNF11*
	chr16	21746755	21776755	2.82	5.50	*OTOA*
	chr12	56724817	56754817	2.82	3.85	*PAN2, IL23A, STAT2, APOF*
	chr2	136123695	136153695	2.82	5.74	*ZRANB3*
	chr4	71635998	71665998	2.80	5.71	*RUFY3*
	chr17	58826088	58856088	2.75	4.69	*BCAS3*
	chr4	71610998	71640998	2.75	6.67	*RUFY3*
	chr2	135998695	136028695	2.73	5.40	*ZRANB3*
*	chr10	57364034	57394034	2.72	2.85	*PCDH15*
	chr2	136098695	136128695	2.72	4.33	*ZRANB3*
	chr7	137258619	137288619	2.71	6.41	*DGKI*
	chr4	510998	540998	2.68	6.25	*PIGG*
	chr9	74490910	74520910	2.68	3.93	*ABHD17B*
*	chr2	135823695	135853695	2.67	3.93	*RAB3GAP1*
	chr9	74515910	74545910	2.66	4.85	*ABHD17B, C9orf85*
	chr12	45799817	45829817	2.66	5.86	*ANO6*
	chr1	51571808	51601808	2.65	4.53	*C1orf185*
*	chr13	52942755	52972755	2.63	2.27	*THSD1*
	chr7	123658619	123688619	2.63	7.45	*TMEM229A*
	chr8	145211376	145241376	2.62	5.91	*MROH1*
*	chr1	225646808	225676808	2.61	4.72	*ENAH*
	chr2	136148695	136178695	2.59	4.94	*ZRANB3*
*	chr2	135773695	135803695	2.57	4.89	*MAP3K19*
	chr12	11174817	11204817	2.57	5.91	*PRR4, TAS2R14, TAS2R19, TAS2R31, AC018630.1*
	chr4	159235998	159265998	2.56	5.33	*RXFP1*
	chr8	145161376	145191376	2.55	7.43	*SHARPIN, MAF1, KIAA1875*
	chr11	75573574	75603574	2.54	5.48	*UVRAG*
	chr7	137233619	137263619	2.53	6.41	*DGKI*
	chr7	74558619	74588619	2.53	3.73	*GTF2IRD2B*

Supplementary Table 5C: iHS positive selection signals found in Somali population. The mean iHS score for the 30 kb windows is reported and the highest –log(p-value) of a variant within the window. All the

reported iHS scores are above the reported 99.99 percentile thresholds calculated after the simulations. Only protein coding genes are reported.

* : Population-specific positive selection signals. These signals are above the 99.99 percentile threshold calculated after the simulations that are not found among the 99 percentile of the rest of the populations of study.

	Chr	Start	Stop	meaniHS	maxP	Genes
	chr1	202396808	202426808	3.59	6.66	*PPP1R12B*
	chr18	66647301	66677301	3.41	6.40	*CCDC102B*
	chr17	44351088	44381088	3.20	9.36	*ARL17B, LRRC37A*
	chr19	39757864	39787864	3.14	5.36	*IFNL2, IFNL1*
	chr18	66722301	66752301	3.09	7.24	*CCDC102B*
	chr1	149821808	149851808	3.03	5.01	*HIST2H2AA4, HIST2H3A, HIST2H4B*
	chr12	82849817	82879817	2.98	6.06	*METTL25*
	chr17	30276088	30306088	2.82	4.76	*SUZ12*
*	chr11	71823574	71853574	2.82	2.76	*ANAPC15, FOLR3*
	chr17	63001088	63031088	2.82	6.13	*GNA13*
	chr8	145211376	145241376	2.81	6.46	*MROH1*
	chr17	44326088	44356088	2.81	9.36	*ARL17B*
	chr9	16565910	16595910	2.79	6.40	*BNC2*
	chr5	144980629	145010629	2.74	5.92	*PRELID2*
	chr8	124261376	124291376	2.73	5.69	*ZHX1-C8ORF76, ZHX1*
	chr18	66672301	66702301	2.72	5.42	*CCDC102B*
	chr17	62501088	62531088	2.72	2.92	*DDX5, CEP95*
	chr12	22824817	22854817	2.69	4.37	*ETNK1*
	chr8	124211376	124241376	2.68	4.99	*FAM83A, C8orf76, ZHX1-C8ORF76*
	chr2	173473695	173503695	2.65	4.94	*PDK1*
	chr1	51571808	51601808	2.63	4.00	*C1orf185*
	chr20	20678005	20708005	2.62	4.46	*RALGAPA2*
	chr20	20653005	20683005	2.62	3.97	*RALGAPA2*
	chr17	74751088	74781088	2.60	4.32	*MFSD11*
	chr8	124236376	124266376	2.60	4.61	*C8orf76, ZHX1-C8ORF76, ZHX1*
*	chr16	85621755	85651755	2.58	7.75	*GSE1*
*	chr10	45789034	45819034	2.58	3.33	*OR13A1*
	chr1	51696808	51726808	2.57	4.10	*RNF11*
	chr3	19298294	19328294	2.56	5.36	*KCNH8*
	chr12	86999817	87029817	2.56	6.20	*MGAT4C*
	chr3	19248294	19278294	2.54	4.84	*KCNH8*

Supplementary Table 5D: iHS positive selection signals found in Wolayta population. The mean iHS score for the 30 kb windows is reported and the highest –log(p-value) of a variant within the window. All the reported iHS scores are above the reported 99.99 percentile thresholds calculated after the simulations. Only protein coding genes are reported.

* : Population-specific positive selection signals. These signals are above the 99.99 percentile threshold calculated after the simulations that are not found among the 99 percentile of the rest of the populations of study.

	Chr	Start	Stop	meaniHS	maxP	Genes
*	chr10	102014034	102044034	3.74	5.78	*CWF19L1, BLOC1S2*
*	chr10	101939034	101969034	3.45	5.57	*ERLIN1, CHUK*
*	chr10	101964034	101994034	3.37	9.85	*CHUK, CWF19L1*
	chr1	202396808	202426808	3.33	5.98	*PPP1R12B*
	chr20	20528005	20558005	3.31	7.06	*RALGAPA2*
	chr16	21746755	21776755	3.20	7.05	*OTOA*
*	chr10	101989034	102019034	3.16	6.87	*CHUK, CWF19L1*
*	chr6	128080815	128110815	3.14	6.05	*THEMIS*
*	chr10	101914034	101944034	3.11	6.23	*ERLIN1*
*	chr7	12258619	12288619	3.05	7.44	*TMEM106B*
	chr8	145161376	145191376	3.05	8.95	*SHARPIN, MAF1, KIAA1875*
	chr12	113849817	113879817	3.05	5.31	*SDS, SDSL*
*	chr6	128005815	128035815	3.00	5.00	*THEMIS*
	chr12	113824817	113854817	2.92	7.44	*PLBD2, SDS*
*	chr4	130010998	130040998	2.89	5.75	*SCLT1, C4orf33*
	chr9	16565910	16595910	2.86	7.70	*BNC2*
*	chr6	154330815	154360815	2.79	5.43	*OPRM1*
	chr2	229773695	229803695	2.75	6.00	*PID1*
*	chr16	23096755	23126755	2.75	5.04	*USP31*
*	chr16	23071755	23101755	2.73	5.11	*USP31*
	chr7	120983619	121013619	2.72	3.66	*FAM3C*
	chr6	128105815	128135815	2.70	6.05	*THEMIS*
	chr16	32046755	32076755	2.69	3.77	*AC142381.1*
	chr2	98848695	98878695	2.69	4.35	*VWA3B*
*	chr6	128055815	128085815	2.68	4.27	*THEMIS*
	chr7	74558619	74588619	2.68	3.41	*GTF2IRD2B*
	chr8	145236376	145266376	2.67	4.31	*MROH1*
*	chr7	65883619	65913619	2.67	4.09	*TPST1*
*	chr6	154305815	154335815	2.67	5.17	*OPRM1*
*	chr16	23121755	23151755	2.64	5.03	*USP31*
	chr2	98823695	98853695	2.63	4.54	*VWA3B*
*	chr7	133983619	134013619	2.62	5.29	*SLC35B4*
	chr11	16048574	16078574	2.61	5.27	*SOX6*
*	chr10	102039034	102069034	2.58	8.06	*BLOC1S2, PKD2L1*
*	chr6	33030815	33060815	2.56	6.90	HLA-DPA1,

						HLA-DPB1
	chr7	121008619	121038619	2.56	3.66	FAM3C
*	chr4	41985998	42015998	2.55	4.93	DCAF4L1, SLC30A9

Supplementary Table 5E: iHS positive selection signals found in Gumuz population. The mean iHS score for the 30 kb windows is reported and the highest –log(p-value) of a variant within the window. All the reported iHS scores are above the reported 99.99 percentile thresholds calculated after the simulations. Only protein coding genes are reported.

* : Population-specific positive selection signals. These signals are above the 99.99 percentile threshold calculated after the simulations that are not found among the 99 percentile of the rest of the populations of study.

ADDITIONAL FILE 3

Structure of the Ethiopian populations

Principal Component Analysis (PCA) of the five Ethiopian populations, together with 1000 Genomes Project samples (Supplementaty Figure 2a), shows PC1 (6.7% of variation) separating the African from OoA populations, and PC2 (3.6% of variation) distinguishing among the non-Africans. In PC1, the Northern Africans (Egypt) lie on the left, close to Europeans, and the sub-Saharan Yoruba on the far right. Ethiopians are widely spread between these extremes, illustrating their high genetic variation and complexity mostly due to variable amounts of Eurasian admixture (58). Interestingly, most of the Ethiopian populations (all the Afroasiatic speakers, but not the Gumuz) lie the closest to North Africans (and other non-Africans) of all the African groups, compatible with their extensive recent OoA admixture. Although the Afroasiatic samples cluster together, the Gumuz population lies closer to the Luhya and Yoruba than to any other samples, showing again the correlation between genetic and linguistic stratification in Ethiopia.

In order to further understand the Ethiopian samples, we performed a PCA with only these five populations, the Yoruba and a European population (CEU). Supplementary Figure 2b shows that the first component (5.6% of the variation) differentiates European from African populations, and thus between Afroasiatic and Nilotic populations, revealing that linguistic affiliation correlates with genetic similarity in Ethiopia. PC2 (2% of the variation) separates Gumuz from Yoruba, with the Afroasiatic populations in between. Among the Ethiopian Afroasiatic samples, Oromo and Amhara appear to be the most closely related with a strong overlap in the PCA.

In an ADMIXTURE analysis of a reduced set of thirteen populations (the Ethiopian samples, Egyptians and some from the 1000 Genomes Project), the clustering obtained with the best K value (K=4) showed two similar main components in the Afroasiatic samples, an Ethiopian component (light blue) at 60-70% and a second component (dark blue) shared among Europeans and North Africans. Interestingly, the Nilotic samples (Gumuz) share

the same 60-70% Ethiopian component but the remaining component is shared with the Niger-Kordofanian-speaking groups Yoruba and Luhya, indicating two different genetic components in sub-Saharan Africans. Almost as significant as K=4 (Supplementary Figure 3b), K=5 shows additional features. The Gumuz samples now have a single component (red) whereas the Afroasiatic populations show in addition a component similar to Europeans (dark blue), and a North African component (light blue).

SFselect selection analysis – additional examples

We found several windows under positive selection shared between some of the populations. One example is a window containing *NSUN3* and *DHFRL1* that is shared between the Amhara, Oromo, Somali and Gumuz populations with significant scores except for Oromo that falls close to the 99.99 percentile. It is the highest scoring protein coding signal in the Gumuz (Supplementary Table 4e). *NSUN3* is expressed in the mitochondria as well acting as an RNA methytransferase. It interacts with the anticodon stem loop of mt-tRNA^{Met} and modifies by methylating the cytosine 34. Depletion of *NSUN3* resulted in the alteration of translation of mitochondrial proteins as well as cell growth indicating the importance of epitranscriptomic modifications for mitochondria protein synthesis (188). A study reported a patient carrying mutations on *NSUN3* resulting in a non-functional protein. The patient presented mitochondrial disease symptoms with an oxidative phosphorylation deficiency in skeletal muscle (189). Dihydrofolate Reductase 1 (*DHFRL1*) was previously thought to be a pseudogene but a recent study demonstrated its functionality and expression in humans (190). This gene has a high homology with *DHFR*, the main gene in charge of maintaining active folate concentrations by the reduction of dihydrofolate to tetrahydrofolate. A study found that *DHFRL1* is localized in the mitochondria that contributes to the *de novo* mitochondrial thymidylate biosynthesis pathway and is essential for mitochondrial DNA (mtDNA) integrity (191). In addition, expression of *DHFRL1* mRNA was found elevated in Friedrich's ataxia patients that suffer mtDNA damage, indicating that *DHFRL1* could be at the base of limiting mtDNA damage (192).

Among the Somali population we found *CYP3A5* gene among the top unique genes under positive selection (Supplementary Table 4c) *CYP3A5* encodes a member of the cytochrome P450 super family of enzymes. The cytochrome P450 proteins are monooxygenases that catalyse many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. The encoded protein metabolizes drugs as well as the steroid hormones testosterone and progesterone. Expression of this gene is widely variable among populations, and a single nucleotide polymorphism that affects transcript splicing has been associated with susceptibility to hypertension. A study showed that polymorphisms in *CYP3A* genes could modify the response to dietary methyl-mercury exposure during early life development (193). Moreover, a case control study in the Han Chinese population showed an association between *CYP3A5* polymorphism and the risk of hypertension (194) but another study with a Ghanaian population found very a small or no association (195). A study did not find signals of positive selection in *CYP3A5* in the 1000 Genomes populations, thus the signal that we find here might be restricted to this population, making a case of a population-specific adaptation (196).

iHS selection analysis captures recent events of selection – additional examples

All five East African populations have some of the most significant variants in a window containing two long intergenic non-coding RNAs (lincRNAs) *AC105393.1* and *AC105393.2* (Amhara $p<10^{-11}$, Oromo $p<10^{-12}$, Somali $p<10^{-12}$, Wolayta $p<10^{-11}$, Gumuz $p<10^{-14}$). We can see in the 1000 Genomes selection browser (197) that even if the amount of data is low, we also find a signal from CEU, CHB and YRI, indicating a possible event of selection among the human species in general. Both lincRNAs are highly expressed in testis according to the last GTEx release (198).

One of the strongest signals also found in several populations (Amhara, Oromo, Wolayta and Gumuz) falls in the *PPP1R12B* gene. The signal is found in the middle of the gene where we find a high density of exons with variants holding extreme p-values $<10^{-6}$ in all four populations. *PPP1R12B* or *MYPT2* is part of the myosin phosphatase protein complex that is formed by three subunits: a catalytic subunit (PP1c-delta, protein phosphatase 1, catalytic subunit delta), a large regulatory subunit (MYPT, myosin phosphatase target) and small regulatory subunit (sm-M20). There

are two isoforms of MYPT, MYPT1 (widely expressed) and MYPT2 (specific to heart, skeletal muscle and brain) (199,200).
A very interesting signal that we found in the Amhara and Oromo is in a window containing the *TPCN1* gene. *TPCN1* encodes a voltage-gated ion two-pore channel that is activated by NAADP (201). A recent study found that two-pore channels (TPCs) are crucial for Ebola virus infection by mediating Ebola virus through the endosomal network into the cytoplasm. By inactivating TPC function they prevented Ebola virus to get into the cytoplasm and infect the cell. They further suggested that all filoviruses (e.g. Ebola and Marburg viruses) require TPCs to infect cells (202).
We have also found that Oromo and Wolayta share a window that reaches the 99.99 percentile threshold of significance that contains the gene *KCNH8* (Supplementary Table 5 Figure 5b). It is the highest scoring window encoding a protein in the Oromo population (iHS mean score 3.33 and variants with $p<10^{-7}$) The Amhara and Gumuz do not reach the 99.99 percentile threshold but the 99.9 (reported in Supplementary Table 1). *KCNH8* is mostly expressed in the central nervous system and performs diverse functions including regulation of neurotransmitter release, heart rate, insulin secretion, neuronal excitability, epithelial electrolyte transport, smooth muscle contraction and cell volume (203).
We also report the chromosome 1 histone cluster (*HIST2H*) with significant -log10(p-value) scores among Oromo and Wolayta (mean iHS scores of 2.55 and 3.01 and variants with $p<10^{-3}$ and $p<10^{-4}$ respectively) that is also among the top SFselect candidates in Amhara, Oromo, Somali and Wolayta. It is difficult to understand why a highly conserved gene region could have been under positive selection.

We will now focus on the population-specific protein-coding genes under selection (Supplementary Table 5). For the Gumuz population, we have found three interesting regions containing multiple genes that are good candidates of population specific adaptations. A first region includes *BLOC1S2*, *CHUK* and *PKD2L1*, a second region contains *THEMIS* and a third contains genes encoding T-cell Receptor Alpha Variable (TRAV) locus.
The first of the three regions, spans several windows with mean iHS significant scores (3.74, 3.45, 3.37) and variants with $p<10^{-9}$ and $p<10^{-8}$. This region contains the biogenesis of lysosome-related organelles complex 1 subunit 2 (*BLOS2* or *BLOC1S2*) which is a

protein involved in the biogenesis of melanosomes and other lysosome-related organelles. *BLOC1S2* is one of the eight subunits of the protein complex BLOC-1. Five of the eight subunits of BLOC-1 (dysbindin, Capuccino, pallidin, Muted and Snapin) have already been shown to produce strong pigmentation phenotypes in mice (204). In humans, it is known that mutations in BLOC-1 (particularly in dysbindin, BLOS3 and pallidin) produce Hermansky-Pudlak syndrome which is a genetically heterogeneous disorder that causes oculocutaneous albinism, prolonged bleeding and pulmonary fibrosis due to the anomalous vesicle trafficking to lysosomes, melanosomes and platelet dense granules (205). No specific phenotypes have been described yet for mutations in *BLOC1S2* but given the strong pigmentation in the Gumuz population and the high UV-B radiation in Ethiopia it could be an interesting candidate to study in more detail. In the GTeX database, there is a considerable number of significant eQTL in skin.

We also find in the same region the *CHUK* (also named *IKK1 or IKKA*) gene. This gene is a member of the serine/threonine protein kinase family that regulates NF-kB signalling essential for lymphoid organogenesis and adaptive immunity (206). NF-kB family is highly important for the inflammatory responses; its deregulation can cause a broad range of pathologies (metabolic diseases, chronic inflammatory diseases, autoimmune disorders and cancer).

The second genomic region that is a candidate of population-specific positive selection in the Gumuz contains the thymus-expressed molecule involved in selection (*THEMIS*) gene. This gene plays a crucial role in the positive selection of developing T-cells. It is only expressed in CD4 ad CD8 thymocytes and at a lower expression in lymph nodes and spleen (207). It acts early in the T-cell receptor signalling cascade by attenuating mild TCR signals that will increase the affinity threshold for activation and allowing positive selection of T cells with naive phenotypes in response to low-affinity self-antigens (208).

Lastly, we found evidence of adaptive selection in the third candidate region specific in the Gumuz. This region contains the T-cell Receptor Alpha Variable (TRAV) locus, which is in charge of antigen recognition. Although the mean iHS score per window do not reach the significance threshold, there are many variants that do reach significance ($p<10^{-6}$).

Another top-scoring signal found only in Oromo contains genes such as sucrase isomaltase (*SI*) (Supplementary Table 5b). *SI* is mostly expressed in the intestinal brush border (209) and it is essential for the digestion of dietary carbohydrates such as starch, sucrose and isomaltose (210). Congenital sucrase isomaltase deficiency is a rare hereditary disease that causes chronic diarrhoea due to the reduction or absence of *SI* that causes carbohydrate malabsorption (211).

Unbalanced ancestry regions – additional examples

Interestingly, another region of 1 Mb containing an olfactory cluster of genes in the genomic coordinate chr6:27200000-28500000 shows a high proportion of African ancestry, especially in the Amhara (~85%). No clear signals of old or recent selection are found.

Bibliography

1. Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. Am J Hum Genet. 2015;96(6):986–91.
2. Haag S, Sloan KE, Ranjan N, Warda AS, Kretschmer J, Blessing C, et al. NSUN3 and ABH1 modify the wobble position of mt-tRNA ^{Met} to expand codon recognition in mitochondrial translation. EMBO J. 2016 Oct 4;35(19):2104–19.
3. Van Haute L, Dietmann S, Kremer L, Hussain S, Pearce SF, Powell CA, et al. Deficient methylation and formylation of mt-tRNAMet wobble cytosine in a patient carrying mutations in NSUN3. Nat Commun. 2016 Jun 30;7:12039.
4. McEntee G, Minguzzi S, O'Brien K, Ben Larbi N, Loscher C, O'Fagain C, et al. The former annotated human pseudogene dihydrofolate reductase-like 1 (DHFRL1) is expressed and functional. Proc Natl Acad Sci. 2011 Sep 13;108(37):15157–62.
5. Anderson DD, Quintero CM, Stover PJ. Identification of a de novo thymidylate biosynthesis pathway in mammalian mitochondria. Proc Natl Acad Sci. 2011 Sep 13;108(37):15163–8.
6. Haugen AC, Di Prospero NA, Parker JS, Fannin RD, Chou J, Meyer JN, et al. Altered Gene Expression and DNA Damage in Peripheral Blood Cells from Friedreich's Ataxia Patients: Cellular Model of Pathology. Pearson CE, editor. PLoS Genet. 2010 Jan 15;6(1):e1000812.
7. Llop S, Tran V, Ballester F, Barbone F, Sofianou-Katsoulis A, Sunyer J, et al. CYP3A genes and the association between prenatal methylmercury exposure and neurodevelopment. Environ Int. 2017 Aug;105:34–42.
8. Li Z, Chen P, Zhou T, Chen X, Chen L. Association between CYP3A5 genotypes with hypertension in Chinese Han population: A case-control study. Clin Exp Hypertens. 2017 Apr 3;39(3):235–40.
9. Fisher DL, Plange-Rhule J, Moreton M, Eastwood JB, Kerry SM, Micah F, et al. CYP3A5 as a candidate gene for hypertension: no support from an unselected indigenous West African population. J Hum Hypertens. 2016 Dec

23;30(12):778–82.
10. Dobon B, Rossell C, Walsh S, Bertranpetit J. Is there adaptation in the human genome for taste perception and phase i biotransformation? BMC Evol Biol. 2019 Dec 31;19(1):39.
11. Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, et al. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res. 2014 Jan;42(Database issue):D903-9.
12. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017 Oct 11;550(7675):204–13.
13. Fujioka M, Takahashi N, Odai H, Araki S, Ichikawa K, Feng J, et al. A New Isoform of Human Myosin Phosphatase Targeting/Regulatory Subunit (MYPT2): cDNA Cloning, Tissue Expression, and Chromosomal Mapping. Genomics. 1998 Apr 1;49(1):59–68.
14. Okamoto R, Kato T, Mizoguchi A, Takahashi N, Nakakuki T, Mizutani H, et al. Characterization and function of MYPT2, a target subunit of myosin phosphatase in heart. Cell Signal. 2006 Sep 1;18(9):1408–16.
15. Calcraft PJ, Ruas M, Pan Z, Cheng X, Arredouani A, Hao X, et al. NAADP mobilizes calcium from acidic organelles through two-pore channels. Nature. 2009 May 22;459(7246):596–600.
16. Sakurai Y, Kolokoltsov AA, Chen C-C, Tidwell MW, Bauta WE, Klugbauer N, et al. Ebola virus. Two-pore channels control Ebola virus host cell entry and are drug targets for disease treatment. Science. 2015 Feb 27;347(6225):995–8.
17. Zou A, Lin Z, Humble M, Creech CD, Wagoner PK, Krafte D, et al. Distribution and functional properties of human KCNH8 (Elk1) potassium channels. Am J Physiol Physiol. 2003 Dec;285(6):C1356–66.
18. Lee HH, Nemecek D, Schindler C, Smith WJ, Ghirlando R, Steven AC, et al. Assembly and architecture of biogenesis of lysosome-related organelles complex-1 (BLOC-1). J Biol Chem. 2012 Feb 17;287(8):5882–90.
19. Li W, Zhang Q, Oiso N, Novak EK, Gautam R, O'Brien EP, et al. Hermansky-Pudlak syndrome type 7 (HPS-7) results from mutant dysbindin, a member of the biogenesis of

lysosome-related organelles complex 1 (BLOC-1). Nat Genet. 2003 Sep 17;35(1):84–9.
20. Polley S, Passos DO, Huang D-B, Mulero MC, Mazumder A, Biswas T, et al. Structural Basis for the Activation of IKK1/α. Cell Rep. 2016 Nov 15;17(8):1907–14.
21. Allen PM. Themis imposes new law and order on positive selection. Nat Immunol. 2009 Aug 1;10(8):805–6.
22. Fu G, Casas J, Rigaud S, Rybakin V, Lambolez F, Brzostek J, et al. Themis sets the signal threshold for positive and negative selection in T-cell development. Nature. 2013 Dec 13;504(7480):441–5.
23. Traber PG, Wu GD, Wang W. Novel DNA-binding proteins regulate intestine-specific transcription of the sucrase-isomaltase gene. Mol Cell Biol. 1992 Aug;12(8):3614–27.
24. Van Beers EH, Büller HA, Grand RJ, Einerhand AWC, Dekker J. Intestinal brush border glycohydrolases: Structure, function, and development. Crit Rev Biochem Mol Biol. 1995;30(3):197–262.
25. Marcadier JL, Boland M, Scott CR, Issa K, Wu Z, McIntyre AD, et al. Congenital sucrase-isomaltase deficiency: Identification of a common Inuit founder mutation. CMAJ. 2015 Feb 3;187(2):102–7.
26. Meylan E, Tschopp J, Karin M. Intracellular pattern recognition receptors in the host response [Internet]. Vol. 442, Nature. Nature Publishing Group; 2006. p. 39–44.
27. Kato H, Takeuchi O, Sato S, Yoneyama M, Yamamoto M, Matsui K, et al. Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. Nature. 2006 May 9;441(1):101–5.
28. Fumagalli M, Cagliani R, Riva S, Pozzoli U, Biasin M, Piacentini L, et al. Population genetics of IFIH1: Ancient population structure, local selection, and implications for susceptibility to type 1 diabetes. Mol Biol Evol. 2010 Nov 1;27(11):2555–66.
29. Hosoda Y, Yoshikawa M, Miyake M, Tabara Y, Shimada N, Zhao W, et al. CCDC102B confers risk of low vision and blindness in high myopia. Nat Commun. 2018 Dec 3;9(1):1782.

2. POSITIVE SELECTION ANALYSIS OF A KHOESAN POPULATION

Sandra Walsh[1], Elizabeth Atkinson[2], Begoña Dobón[1], Sandra Acosta[1], Deepti Gurdasani[3], Tommy Carensten[3], Caitlin Uren[2], Javier Prado-Martinez[3], Martin Pollard[3], Manj Sandhu[4], Marlo Möller[5], Eileen Hoal[5], Hafid Laayouni[1], Brenna Henn[2,*], Jaume Bertranpetit[1,*]

1. Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Doctor Aiguader, 88. 08003 Barcelona, Catalonia, Spain.
2. Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY, USA; Department of Anthropology and the Genome Center, University of California, Davis, Davis, CA 95616, USA.
3. Department of Human Genetics, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK.
4. Department of Medicine, University of Cambridge, Cambridge, CB2 0QQ, UK.
5. Molecular Biology and Human Genetics, MRC Centre for Molecular and Cellular Biology, DST/NRF Centre of Excellence for Biomedical TB Research, Faculty of Health Sciences, Stellenbosch University, Tygerberg, South Africa.

* Co-corresponding authors: bmhenn@ucdavis.edu, jaume.bertranpetit|@upf.edu

Keywords
KhoeSan, southern Africa, positive selection, selective sweeps

Abstract

The analysis of adaptive selection of tha Nama, a KhoeSan semi-nomadic pastoralist from southern African belonging to the deepest extant human lineage, can give unvaluable an insight into the evolutionary history of humans. In this study, we have uncovered the Nama specific positive selection signals accounting for recent migration events. Interestingly we have found two typical European signals that have been widely studied, the light pigmentation *SLC24A5* gene and the lactase persistence *LCT* gene, of which the mutation in Nama is the Eastern African lactase persistence allele. Both events represent a clear evidence of selection after admixture.
Additionaly, the evolutionary history of human populations has been marked by dietary changes in different environments, thus, it is important to study those adaptations since many worldwide modern diseases are associated with diet. We have described several interesting adaptations involving the lipid metabolism and cardiovascular function and genes clearly related to the body mass index, along with genes related to some micronutrients (iodine). A region containing the *TG* (thyroglobulin) gene is a good candidate of an adaptation since Africa has suffered for many years of iodine deficiency. And much more attention should be payed to cardiovascular diseases specific for specific African populations.
A very high number of genes with signals of adaptation are those related to muscle or to connective tissue. Even if the direct physiological action cannot be established, their accumulation postulates a population-specific adaptation in relation to muscle activity. Last but not least, in the class where more candidate genes are concentrated are those of brain function and development. It pops up both in the gene ontology enrichment analysis as well as in the genes in highly selected regions. These findings are in agreement with previous studies that found many brain specific adaptations across the human lineage.

Introduction

The KhoeSan peoples are a wide group of indigenous populations inhabiting countries from the south of Africa. The term KhoeSan (*khoe* means 'person' and *san* means 'bushmen' in Khoekhoe) reflects an ethnological division into two groups: the Khoekhoe and the San (1). The KhoeSan constitute a unique group of people because of their subsistence strategies and their linguistic and genetic background. The KhoeSan languages, which are one of the four language families in Africa, are characterised by the particular use of click consonants. Some groups of KhoeSan practice a hunter-gatherer subsistence strategy, which has been by large the main model of subsistence of our species during its evolutionary history but the rarest nowadays. At the genetic level the KhoeSan belong to the deepest extant human lineage (2), where recent estimates date the split from the rest of present-day human populations around 300,000 years before present (BP) (3), being thus the deepest divergence among modern humans.

The genetic structure of the KhoeSan populations is far from being trivial; it does not simply reflect a correlation between language, culture or subsistence strategy but rather indicates a correlation with geography and ecological barriers (e.g. the Kalahari desert) (4). Several studies agree that present day KhoeSan populations are a mixture of three different KhoeSan ancestries, Northern (e.g. mostly represented by Ju|'Hoansi and !Xun), Central (e.g. G|ui, G||ana and Naro) and Southern or Circum-Kalahari KhoeSan (e.g. the Nama, ≠Khomani and Karretjie) and the three contain different KhoeSan family languages, modes of subsistence and cultural practices (4–6). At the same time that the complex structure of the KhoeSan populations has been unravelled, events of recent admixture have been described. The main sources and proportions of this gene flow, which are unequally distributed among the KhoeSan, were from East Africa from Bantu-speaking populations and from European colonists. East African admixture, dated in the last 2000 years (7,8), is higher among the Khoekhoe groups (2,7,9). This migration not only introduced new genetic variation in the KhoeSan (in the order of 23-30% from an East African source) (3) but also suggests that initial pastoralist practices where introduced in southern Africa by demic diffusion rather than by cultural diffusion (8–10). This migrant East African population was already an admixed group with an average 30% of Eurasian-Levant ancestry, which makes the ancestors of the nowadays Afro-asiatic populations from Ethiopia

the probable donor population (3). Later and independently of the migration of East African populations, there is evidence that the Bantu-speaking groups reached the south of Africa through two different waves of the Bantu expansion (11–14). The first wave represented by the ancestors of the Eastern Bantu-speaking groups that migrated from western Africa, reached East Africa 3000 BP and then expanded to southern Africa by 1300 BP. The second wave was initiated by the ancestors of the Western Bantu-speakers that migrated through the Atlantic coast from the west to the south of Africa. The Bantu expansion had a great impact on the populations of southern Africa at the linguistic, cultural and genetic level (4,7). Lastly, the arrival of European colonists in the area some 300 years ago also shaped the KhoeSan genetic background.

The migrations that affected KhoeSan genomes have caused cultural and lifestyle changes (e.g. adoption of pastoralism and farming) that have left signatures of positive selection in their genomes. One of the most studied cases in the KhoeSan (and probably in worldwide populations) is the lactase persistence (LP) adaptation. Many studies have focused on LP (10,15), especially in Europe and East African populations from Tanzania and Kenya where many LP alleles have been reported (16). In Africa, LP cases have been described in eastern and southern Africa and the evidence points towards a demic diffusion of pastoralism from eastern to southern Africa followed by a selective sweep affecting the lactase gene region in some particular pastoralist populations (e.g. the Nama). The arrival of pastoralism in the south from eastern Africa dates ~2000 BP, the Nama, a pastoralist population, show a 50% of LP, and their frequency of the LP allele (rs145946881) of eastern African origin is 33%, way above the expected 10% resulting after the gene flow of eastern African pastoralists (15). Another interesting signal of adaptive selection is related to skin pigmentation. A recent study showed (using Nama and ≠Khomani individuals) that the European allele associated with light skin pigmentation (rs1426654) was introduced after the migration from eastern Africa of an already admixed population (30% Eurasian, 70% East African) that in turn was carrying the European allele. Their best-fit model strongly suggests that this allele underwent strong positive selection after its introduction around ~1500 BP (17). The influx of infectious diseases by the migrant groups have most certainly also affected the genetic pool of KhoeSan populations. Several epidemics such as the flu and smallpox have

been documented to greatly affect KhoeSan populations leaving genomic footprints of adaptation in the extant populations (18). Additionally, a recent study with two 2000 year old ancient south African samples found recent signals of adaptation in taste receptors and in groups of genes or pathways related to response to radiation (9).

The complex evolutionary history of the KhoeSan populations, especially the complex recent pattern of gene flow, makes it a great challenge to study adaptive selection. There are many methods of detection of positive selection, but a few that take into account admixture. The most common method that may account for admixture is the haplotype based method XP-EHH (19), which uses a reference population (probable source of gene flow) to remove the potential spurious effects caused by the haplotypes from the donor population. Moreover, the newly introduced variation by admixture can provide the recipient population with a selective advantage. Many cases of adaptive selection after admixture have been described (20–22), even with genetic variation from archaic hominin populations (23–25).

This study is an in-depth analysis, using whole-genome sequences, of adaptive selection in a pastoralist population from southern Africa, the Nama. This can provide a better a unique insight into the evolutionary history of the human species.

Results

Population structure

The principal component analysis (PCA) separates African from non-African populations at the first PC (Figure 1A), and shows a very strong differentiation of the Nama in relation to other Africans and a high diversity among the Nama individuals, much higher than for any other population. The fourth PC (Figure 1B) separates two different African components, the Gumuz as a representative of the East African gene pool (26) and the Bantu, enclosing two geographically very distant populations, Zulu from southern Africa and living mainly in the province of KwaZulu-Nata and the Yoruba from western Africa, living in Nigeria. The diversity of Nama individuals encompass both the first, the second and the fourth components, a sign of the diverse admixture with Eurasian-Levant, with Bantus and with East Africans.

The admixture analysis (Figure 1A) shows similar results, with a clear Nama component and two other, separated, African components, the Bantu (with Yoruba and Zulu, the last having a small amount of Nama admixture) and Eastern African, with the Gumuz as the best representative.

To have a better insight in the ancestry of the Nama, we have performed a more analytical approach. This step is fundamental both for demographic inference and for reliability of genome masking.

Adaptive selection analysis

These results give us, under a well-tested demographic model, the tools to perform an in-depth analysis for adaptive selection signals in the Nama genome needed because of the recent admixture between the Nama, Bantu-speaking and European populations. We therefore have designed the study of adaptive selection in such a way not to be influenced by admixture neither from Europeans (in a wide sense, including a both the Levant component and the more recent arrival of Europeans) nor from Bantu speaking neighbours. We have first masked (marked as missing data) the European component (12.8%) from the Nama genomes by LAI analysis. Then, since the Bantu component of the Nama (10,4%) could also influence our results, we have used XP-EHH (19), a cross-population haplotype based method to detect positive selection using the Zulu population (a Bantu-speaking group from South Africa) as reference.

In order to have an overview of the nature of the signals of selection, we ranked the XP-EHH scores per variant (for each SNP) and took the 1% most extreme values and annotated them. We also grouped variants within a 100-kilobase (kb) window and calculated the mean XP-EHH score, ranked the windows and annotated the extreme 1%. Tables 1A-B show the summary of the annotation of the putative selection signals by SNPs and by 100Kb windows. It is interesting to note the strong amount of selection signals in places of the genome with no annotation (43%); this fact, similar to what is found in GWAS studies when the criteria of finding annotation is strict, may likely be due to a dearth of annotation of the genome, mainly for regions important for gene regulation. From the 436,250 positions, that represent the top 1% of positive XP-EHH values, 57% of them had a gene annotation. The much higher values when considering SNPs (Table 1B) indicates that the top high values of XP-EHH are highly concentrated in cluster, seen as peaks in the figures and they highly enriched in coding elements, and specially in protein-coding genes (Table 1B).
After filtering repeated annotations (multiple positions in tandem could have the same annotation), we obtained a total of unique 2812 coding elements that physically overlaps with the SNPs of the top 1%. From these 2812 coding elements, 70% of them are protein-coding genes (Table 1C). There is a very strong enrichment of protein-coding genes (40% of the extreme selection signals) even if there is a very high amount of positions with strong selection signals with no annotation. Other elements are also at high proportion, including a 12.5% of lincRNA and 5.3% of pseudogenes among the annotated elements. When considering windows with strong signals (Table 1D), the proportion of pseudogenes is much higher, 16.5%, likely due to their proximity to protein-coding genes.

Enrichment analysis of the XP-EHH putative selection candidates
We next examined if genes with specific functional categories were overrepresented among the strongest candidates to be under selection. Overrepresentation analyses are useful methods that give a functional overview on where selection acted and provide a better understanding of candidate lists of genes, allowing the detection of complex adaptations.
We first conducted the analysis with traseR (27), a trait-associated SNP enrichment tool based on GWAS studies. In Table 2 we can

see the most significant gene categories and in Supplementary Table 2B a list of the genes and variants that have been related to the phenotype that will be discussed later. There is a clear enrichment mostly related to cardiovascular function, like blood viscosity (p-adjusted 0.01), coronary disease (p-adjusted 0.01) or hearth function tests (p-adjusted 0.076) and also to lipid metabolism such as HDL and LDL cholesterol (p-adjusted 0.04 and 0.06 respectively). These results might be pointing on one hand towards a recent metabolic adaptation of the Nama with interesting biomedical implications and on the other hand to cardiovascular function related genes. Another interesting category is skin pigmentation (p-adjusted 0.03), discussed later.

We next used WebGestalt (28) to perform a gene based enrichment analysis with the most common databases of gene functional categories and pathways: Gene Ontology (Biological Process and Molecular Function) and KEGG. In Biological Process (Supplementary table 2), there is a strong enrichment in terms related to processes of the nervous system, for example nervous system development (p-adjusted 1.35E-06), cell morphogenesis involved in neuron differentiation (p-adj 9.64E-06), axogenesis (p-adjusted 6.21E-05), neuron development (p-adjusted 1.02E-04), synaptic signalling (p-adjusted 2.03E-02) and sensory perception of sound (p-adjusted 3.89E-02); in fact 17 out of the 19 top and significant values are related to the nervous system and the top 11 are directly related to neuron development.

The Molecular Function enrichment (Supplementary table 3) shows much less extreme cases, with a very strong amount of functions related to transport or channel activity (16/28 significant cases) like ion channel activity, transmembrane transporter activity, cation channel activity, metal ion transmembrane activity, calcium channel activity. Many other cases refer to binding (9/28) like to calcium ion, carbohydrate, actin, actinin, ATP, zinc ion, vitamin D or muscle alpha-actinin. Interestingly we also found sialyltransferase activity category (p-adj 1.63E-03) and glutamate receptor activity.

The enrichment with KEGG database (Table 3) revealed brain related categories confirming the Biological Process of Gene Ontology results (circadian entrainment, glutamatergic synapse, retrograde endocannabinoid signalling and long-term potentiation) and insulin secretion (p-adj 3.85E-02).

Some of these categories will be found in the search of the top regions, but others will not, a fact due to be categories formed by

genes that have a clear selection signal, but not with an extreme value.

Regions of interest

We performed a detailed analysis of the main selection candidates focusing on 100 kb windows with very high XP-EHH scores that contain clusters of extreme values of XP-EHH per SNP. Figure 4 shows a circular Manhattan plot of XP-EHH of Nama versus Zulu, and iHS results both for Nama and Zulu, with some annotations. Table 4 shows the top scoring windows containing protein-coding genes of special interest that will be discussed; the full list is in Supplementary tables 4 and 5. A full description of the regions with the genes and regulatory elements is in the Supplementary text. Here it is only a short description.

Region 1. It includes genes *METTL13*, *DNM3 and VAMP4* (Figure 5), having interesting brain-related functions (29–33) and strong recent specific adaptation in Nama, likely due to more than a single selective event. In the region between *DNM3 and VAMP4* there are several regulatory elements (see Supplementary Text) and many variants at high derived allele frequency in Nama (50%) and absent in other populations that are located at the most robust selective sweep of the whole genome; nonetheless, the relationship of their sequence variation with their functional implications is not well established.

Region 2. This is a broad region with high selection scores, encompassing the area around the lactase (LCT) gene and a larger region at its 5' (Figure 6A and B). 50% of Nama population is lactase persistent (LP) (15) but after masking, the European allele conferring LP (13910C>T) is zero; they have the East African substitution 14010G>C that confers LP at a frequency of 33%. In our results, there are not signals of selection coming from the *MCM6* intron 13, where most regulation signals for LP have been described (10,16,34,35), but strong signals are found inside the *LCT* and in its 5' region. The extended haplotypes of the region in the Nama are very similar to those found in East African populations, like present Amhara (Figure 6C-E and Supplementary Text).

Region 3. Containing mainly gene *SLC24A5* (Figure 7A), strongly associated through rs1426654 with light skin pigmentation in Europeans, but of minor importance for skin pigmentation in KhoeSan (36) even if they have a high frequency of the European allele. It is likely that, besides the selection that drove this allele to a

high frequency in Nama (17), an adaptive sweep also happened in the Nama haplotypes (as the Europeans had been masked) in the 5' region of the gene.

Region 4. A member of the fibrillin gene family, *FBN3,* (Figure 7B) is important in maintaining the structure and integrity of the extracellular matrix in connective tissues (37,38) and in GWAS has been related to attention deficit hyperactivity disorder and facial morphology (39,40). The highest scoring variants of this signal intersect with two cis-regulatory elements (see Supplementary Text).

Region 5. In this region we find the thyroglobulin (TG) gene (Figure 7C) that produces the substrate for the synthesis of thyroid hormones, required for the regulation of metabolism (41). There is an interesting non-synonymous variant with a high Phred-scaled CADD score (11.73) among our selected variants, rs73354644 (A2422T), with the highest derived frequency in Nama (0.36), and absent outside Africa and a top candidate for selection.

Region 6, encompassing mainly the hardly known *NCKAP5* gene (Figure 7D) that has been associated in GWAS with body mass index (42). There is a non-synonymous substitution (rs12611515) at a frequency of 0.80 in the Nama (Supplementary Table 6) with a very high Phred score of 21, but nothing is known on its possible phenotype implications.

Region 7 Two strong peaks of XP-EHH high scores correspond clearly to *DDP4* and *SLC4A10* genes (Figure 7E). The *DPP4* (also called *CD26*) product plays a major role in glucose metabolism (43,44), is an enzyme expressed on the surface of most cell types and is associated with immune regulation, signal transduction, and apoptosis; it has been shown to negatively regulate lymphocyte trafficking, and its inhibition enhances T cell migration and tumour immunity (45). *SLC4A10* gene belongs to a small family of sodium-coupled bicarbonate transporters (NCBTs) that regulate the intracellular pH of neurons, and would be crucial to maintain a faster neuronal excitability (46,47). Related by GWAS to educational attainment and psychiatric conditions (48,49). There are two variants in *SLC4A10* in our data (rs113208259, rs113278723) with high Phred scores 10 and 19.6 where Nama have very high derived allele frequencies (both 0.44), even if no phenotypic information exists for them.

Region 8. It is a region (Figure 7F) corresponding to the 3' end of two genes, *NPFFR2* (or *GPR74*) and *ADAMTS3*. The product of

NPFFR2 is a member of a subfamily of G-protein-coupled neuropeptide receptors and this gene is related to metabolism (regulation of body mass index) via a brain physiological process (50–56). *ADAMTS3* plays an important role in the collagen synthesis, and is related to height (57,58). Since the strongest peak of scores lies between both genes, more evidence is needed to associate the signal of selection to a specific gene.

Region 9. Strong signal overlapping with *MRAP2* (Figure 7G) that regulates food intake and energy expenditure and thus related to body weight and obesity (59–61). A second peak in the region corresponds to *KIAA1009* (*CEP162*) gene that acts by specifically recognizing and binding the axonemal microtubule. The region contains several regulatory elements for neuronal cells (See supplementary Text).

Region 10. The main signal corresponds to gene *MMP2, (*metalloproteinase 2; Figure 7H), related both to cardiovascular metabolism and to extracellular matrix maintenance (62). MMP2 is one of the most studied metalloproteinases in cardiovascular research (63,64). There is an interesting variant in a splice site (rs243834), with very low frequency in Nama in comparison to other African and non-African populations and has a potential regulatory function (see Supplementary Text).

Region 11. This region contains the gene *CETP,* encoding a cholesteryl ester transfer protein (Figure 7I). It is an important protein for HDL degradation metabolism since it enables the transfer of cholesteryl ester from HDL toward triglyceride rich lipoproteins and LDL and contributing to lower HDL cholesterol (65).

Region 12. Another region that we have found related to cardiovascular function contains *ZFHX3,* a transcription factor that regulates myogenic and neuronal differentiation (Figure 7J) with variants that have been related to atrial fibrillation, ischemic stroke and cardio embolic stroke (66–69). Other variants have interesting regulatory function in cardiac muscle cells (see Supplementary Text).

Region 13. This region (Figure 7K) contains *TRPM3* that encodes a member of the family of transient receptor potential channels, which is mainly expressed in sensory neurons and acts as a sensor of nociceptive heat (70,71). It is expressed both in the central and peripheral nervous system. Mutations in the gene cause mental disabilities (72).

Discussion

The analysis of adaptive selection of the Nama, a southern African semi nomadic pastoralist population, has given an insight into the potentially recent adaptations that this population has undergone. We have uncovered the Nama specific positive selection signals by masking the European ancestry haplotypes (and thus excluding any spurious selection signals that could have originated in the admixed Europeans), and by applying a cross population test using the Zulu population as a reference in order to exclude any signals deriving from populations of Western African Bantu origin. In both cases the admixture is important and could have given spurious signals.

Interestingly and contrary as we would expect after the masking process, we have found two typical European signals that have been widely studied, the light pigmentation *SLC24A5* gene and the lactase persistence *LCT* gene, of which the mutation in Nama is the Eastern African lactase persistence allele. These two selection signals are reflecting common events of selection as in both cases the derived allele comes from an external population (Europe for *SLC24A5*, East Africa for *LCT*) through migration and selection kept acting on them, towards lactase tolerance for the *LCT* variant that was introduced from East Africa, and towards less dark pigmentation for the *SLC24A5* introduced both by the West Eurasian component of East Africans and the most recent European colonisation followed by a very strong selective sweep (17). The variants that we find under positive selection at the 5' end of the gene are likely to be key in the gene function (Figure 7A and ref (36). It is interesting to note that the genetic effect does not have to be through a process of adaptive selection, as sexual selection could have been acting and could have produced the detected event.

The evolutionary history of human populations has been marked by dietary changes in different environments that were likely associated with major cultural shifts such as the advent of agriculture and pastoralism. It is important to study those adaptations since many worldwide modern diseases are associated with diet. In our study we have described several interesting potential adaptations involving the lipid metabolism (traseR enrichment analysis, region 11), insulin pathway (KEGG pathway analysis) and genes clearly related to the body mass index (in regions 6, 8, 9). In fact a striking number of regions found under

strong selection have functions associated with the lipid metabolism and the cardiovascular system. This and the results from the enrichment analysis clearly support the fact that the Nama have had many adaptations related to the fat metabolism and cardiovascular function, the last being a top signal for enrichment and for region analysis (region 10 and 12). Surprisingly, Africa is the continent where we find the highest rates of raised blood pressure (73), which is a major driver of cardiovascular diseases (CVD) but we have not been able to find data for the Nama.

Why do we find so many genes related to cardio metabolic functions under strong selection? Are the Nama a sort of exception among Africans with less CVDs? Are those adaptations conferring higher rates of CVD? We do not have a good understanding of the function of the genetic variants that may have been selected for. More genetic and epidemiological studies in Africa are needed to understand the genetic basis of such putative adaptations.

Among other high scoring regions, we find adaptations related to metabolism through hormone synthesis and regulation. The region that contains the *TG* gene could be at the base of an interesting adaptation related to thyroid hormone and iodine deficiency in Africa. On the other hand, *DPP4*, a gene related to the glucose metabolism, through insulin, has been has been studied to treat diabetes type 2.

Infection forces are likely to have been important in the adaptation of human populations. In this study we have found two genes that may be clearly interpreted (*VAMP4* and *DPP4*) but surprisingly we don't find as many genes or strong signals as in other populations.

Finally, in the class where more genes are concentrated are those of nervous system. It pops up both in the gene ontology enrichment analysis (in Biological Process, Molecular Function and KEGG pathways) as well as in the genes in highly selected regions 1, 8, 13 and possibly 7. The strong concentration of nervous system related categories in the enrichment analysis points to a possible selective process affecting many different genes even if not in a very strong fashion, in what could be a case of polygenic selection. More functional studies are needed to fully understand the functions of these selected genes.

All in all our study has revealed many potential adaptations in a KhoeSan pastoralist population from Namibia, considered as one of the populations of mankind holding the most ancestral variation. Our findings suggest that, among others, genes important for brain

function and development, along those of muscle and connective tissues are those of greatest selective impact for the genes involved. Others related to dietary and environmental changes have been crucial in shaping present African genomes, along those related to specific infectious diseases.

Even if there is a gap in the possibilities of interpreting properly the signals found here, many novel findings ask for the need of genetic variation studies in Africa in order to be able to understand the adaptations and the impact of certain diseases in the African continent.

Methods

Summary of the data

This dataset consists of a set of low coverage (4x-8x) Illumina HiSeq curated sequences from the AGVP (African Genome Variation Project). The data has been processed, called and filtered with GATK (74) and phased with SHAPEIT2 (75). No missing data, no indels, only biallelic sites. All SNPs in the database have been used.

PCA and ADMIXTURE

To perform the PCA and ADMIXTURE analysis a set of all Nama individuals and 18 individuals from 11 populations (Zulu, YRI, Gumuz, Amhara, Oromo, Somali, Wolayta, Egypt, CEU, GBR, CHB) was used. A minor allele frequency filter of 0.05 was applied and variants were pruned using the PLINK software (76) with parameters --indep 50 5 2 to remove the effect of linkage disequilibrium.

The PCA analysis was done using *smartpca* from the Eigensoft 6.0.1 software (77). Population structure analysis was performed with the ADMIXTURE software (78).

Masking of CEU ancestry

SNP ancestry calls were used to mask our data and avoid the European (CEU) ancestry regions. From the ancestry calls it was not possible to know which was the specific chromosome of the pair that contained the European ancestry in each individual of the AGR data. Hence, regions of CEU ancestry from a chromosome were masked in both chromosomes.

iHS and XP-EHH analysis

In order to scan for recent selection events we decided to perform the analysis with the rehh 2.0 program (79) that allows calculating the iHS and XP-EHH scores (19) per SNP allowing for missing data (CEU ancestry masked data). Since no population specific genetic map is available for these populations and given that iHS and XP-EHH are strongly affected by it when performing the extended haplotype homozygosity calculations, no genetic map was used to calculate the statistics.

Given that our Nama individuals have, beyond the European component, a Bantu component, iHS results could be influenced by it. To confirm the Nama ancestry from our selection candidates we used a cross-population test (XP-EHH) test using Zulu population from South Africa as a reference. Note that we only analysed the positive values of XP-EHH since they represent events of selection in Nama whereas negative values are events that happened in the Zulu population. Thus the present analysis for the Nama is designed in such a way not to be influenced by admixture neither from Europeans nor from Bantu speaking neighbours.

Enrichment analysis of putative selection candidates

Two softwares were used to perform the gene enrichment analysis, WebGestalt (28) and traseR (27). WebGestalt is a functional enrichment analysis web tool. We used 1% of extreme positive scoring positions to perform the analysis. We annotated the positions with BEDTools Intersect (80) and used the genes found to run WebGestalt. The background used was all positive positions mapping the genetic elements. We found in the 1% 2812 coding elements. The background has 36364 coding elements in total.

By the other hand, traseR performs enrichment analyses of trait-associated SNPs obtained from GWAS analysis in arbitrary genomic intervals with flexible options, focusing in trait-associated SNPs. From the positive XP-EHH values (43,625,000), we took the 1% extreme SNPs (436,250) as putative positions under selection in the Nama or at least highly enriched. The background used is all the positions from the positive XP-EHH that have an rsID (35,559,195). To calculate p-values the binomial test option was used and an FDR correction was applied to obtain q-values. Only the trait-associated SNPs (without linkage disequilibrium interpolation) with a GWAS p-value greater than 5e-8 were used and only reported overrepresented categories.

Annotation of variants

We used ANNOVAR software (81) to annotate the variants of interest with different databases such as RefSeq, ClinVar, GWAS catalogue, dbSNP and Phred-scaled CADD scores (http://cadd.gs.washington.edu/info). Phred-scaled CADD scores

follow a logarithmic distribution, therefore a variant with a phred-scaled CADD score of 10 is in the 10th percentile highest CADD scores and a score of 20 corresponds to the 1st percentile. We also calculated allele frequencies of the Nama and other populations (Zulu, Gumuz, Amhara, Wolayta, CEU, CHB) to have, for those positions under selection, comparative allele frequencies in other populations. To try to unravel the putative regulatory functions of our candidate regions under positive selection, we have used the ENCODE Encyclopedia (82) of cis-Regulatory Elements and SCREEN, the web-based visualization engine (http://screen.encodeproject.org/).

Bibliography

1. Barnard A. Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan Peoples. Cambridge: Cambridge University Press; 1992.
2. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Blum MGB, Soodyall H, et al. Genomic Variation in Seven Khoe-San. 2012;1187(October):374–9.
3. Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science (80-). 2017 Nov 3;358(6363):652–5.
4. Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, van Helden PD, et al. Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. Genetics. 2016 Sep 1;204(1):303–14.
5. Montinaro F, Busby GBJ, Gonzalez-Santos M, Oosthuitzen O, Oosthuitzen E, Anagnostou P, et al. Complex Ancient Genetic Structure and Cultural Transitions in Southern African Populations. Genetics. 2017;205(1):303–16.
6. Vicente M, Jakobsson M, Ebbesen P, Schlebusch CM. Genetic Affinities among Southern Africa Hunter-Gatherers and the Impact of Admixing Farmer and Herder Populations. Heyer E, editor. Mol Biol Evol. 2019 Sep 1;36(9):1849–61.
7. Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A. 2014 Feb 18;111(7):2632–7.
8. Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, et al. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc Natl Acad Sci. 2008 Aug 5;105(31):10693–8.
9. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, et al. Reconstructing Prehistoric African Population Structure. Cell. 2017 Sep 21;171(1):59-71.e21.
10. Breton G, Schlebusch CM, Lombard M, Sjödin P, Soodyall H, Jakobsson M. Lactase Persistence Alleles Reveal Partial East African Ancestry of Southern African Khoe Pastoralists. Curr Biol. 2014 Apr 14;24(8):852–8.

11. Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science (80-). 2017 May 5;356(6337):543–6.
12. Filippo C de, Bostoen K, Stoneking M, Pakendorf B. Bringing together linguistic and genetic evidence to test the Bantu expansion. Proc R Soc B Biol Sci. 2012 Aug 22;279(1741):3256.
13. Li S, Schlebusch C, Jakobsson M. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. Proc R Soc B Biol Sci. 2014 Oct 22;281(1793):20141448.
14. Semo A, Gayà-Vidal M, Fortes-Lima C, Alard B, Oliveira S, Almeida J, et al. Mozambican genetic variation provides new insights into the Bantu expansion. bioRxiv. 2019 Jul 10;697474.
15. Macholdt E, Lede V, Barbieri C, Mpoloka SW, Chen H, Slatkin M, et al. Tracing pastoralist migrations to southern Africa with lactase persistence alleles. Curr Biol. 2014 Apr 14;24(8):875–9.
16. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007 Jan 10;39(1):31–40.
17. Lin M, Siford RL, Martin AR, Nakagome S, Möller M, Hoal EG, et al. Rapid evolution of a skin-lightening allele in southern African KhoeSan. Proc Natl Acad Sci U S A. 2018;115(52):13324–9.
18. Jakobsson M, Owers KA, Schlebusch CM, Skoglund P, Soodyall H. Adaptation to infectious disease exposure in indigenous Southern African populations.
19. Vitti JJ, Grossman SR, Sabeti PC. Detecting Natural Selection in Genomic Data. 2013;
20. Jin W, Xu S, Wang H, Yu Y, Shen Y, Wu B, et al. Genome-wide detection of natural selection in African Americans pre- and post-admixture. Genome Res. 2012 Mar;22(3):519–27.
21. Jeong C, Alkorta-Aranburu G, Basnyat B, Neupane M, Witonsky DB, Pritchard JK, et al. Admixture facilitates genetic adaptations to high altitude in Tibet. Nat Commun. 2014 Feb 10;5:3281.
22. Pierron D, Heiske M, Razafindrazaka H, Pereda-Loth V,

Sanchez J, Alva O, et al. Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. Nat Commun. 2018;9(1):1–9.
23. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014 Aug 2;512(7513):194–7.
24. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nat Rev Genet. 2015 Jun;16(6):359–71.
25. Dolgova O, Lao O. Evolutionary and Medical Consequences of Archaic Introgression into Modern Human Genomes. Genes (Basel). 2018 Jul 18;9(7).
26. Dobon B, Hassan HY, Laayouni H, Luisi P, Ricaño-Ponce I, Zhernakova A, et al. The genetics of East African populations: A Nilo-Saharan component in the African genetic landscape. Sci Rep. 2015 Sep 28;5(1):9996.
27. Chen L, Qin ZS. traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals: Table 1. Bioinformatics. 2016 Apr 15;32(8):1214–6.
28. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. 2017 Jul 3;45(W1):W130–7.
29. Lu J, Helton TD, Blanpied TA, Rácz B, Newpher TM, Weinberg RJ, et al. Postsynaptic Positioning of Endocytic Zones and AMPA Receptor Cycling by Physical Coupling of Dynamin-3 to Homer. Neuron. 2007 Sep 20;55(6):874–89.
30. Romeu A, Arola L. Classical dynamin DNM1 and DNM3 genes attain maximum expression in the normal human central nervous system. BMC Res Notes. 2014 Mar 28;7:188.
31. Costas J, Carrera N, Alonso P, Gurriarán X, Segalàs C, Real E, et al. Exon-focused genome-wide association study of obsessive-compulsive disorder and shared polygenic risk with schizophrenia. Transl Psychiatry. 2016 Mar 29;6(3):e768–e768.
32. Nicholson-Fish JC, Kokotos AC, Gillingwater TH, Smillie KJ, Cousin MA. VAMP4 Is an Essential Cargo Molecule for Activity-Dependent Bulk Endocytosis. Neuron. 2015 Dec 2;88(5):973–84.
33. Mather KA, Armstrong NJ, Wen W, Kwok JB, Assareh AA,

Thalamuthu A, et al. Investigating the Genetics of Hippocampal Volume in Older Adults without Dementia. Arking DE, editor. PLoS One. 2015 Jan 27;10(1):e0116920.
34. Wagh K, Bhatia A, Alexe G, Reddy A, Ravikumar V, Seiler M, et al. Lactase Persistence and Lipid Pathway Selection in the Maasai. Johnson N, editor. PLoS One. 2012 Sep 28;7(9):e44751.
35. Jones BL, Raga TO, Liebert A, Zmarz P, Bekele E, Danielsen ET, et al. Diversity of Lactase Persistence Alleles in Ethiopia: Signature of a Soft Selective Sweep. Am J Hum Genet. 2013 Sep 5;93(3):538–44.
36. Martin AR, Lin M, Granka JM, Myrick JW, Liu X, Sockell A, et al. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. Cell. 2017 Nov 30;171(6):1340-1353.e14.
37. Sabatier L, Miosge N, Hubmacher D, Lin G, Davis EC, Reinhardt DP. Fibrillin-3 expression in human development. Matrix Biol. 2011 Jan;30(1):43–52.
38. Davis MR, Summers KM. Structure and function of the mammalian fibrillin gene family: Implications for human connective tissue diseases. Mol Genet Metab. 2012 Dec;107(4):635–47.
39. Hawi Z, Yates H, Pinar A, Arnatkeviciute A, Johnson B, Tong J, et al. A case–control genome-wide association study of ADHD discovers a novel association with the tenascin R (TNR) gene. Transl Psychiatry. 2018 Dec 18;8(1):284.
40. Lee MK, Shaffer JR, Leslie EJ, Orlova E, Carlson JC, Feingold E, et al. Genome-wide association study of facial morphology reveals novel associations with FREM1 and PARK2. Li Y, editor. PLoS One. 2017 Apr 25;12(4):e0176566.
41. Di Jeso B, Arvan P. Thyroglobulin From Molecular and Cellular Biology to Clinical Endocrinology. Endocr Rev. 2016 Feb 1;37(1):2–36.
42. Kichaev G, Bhatia G, Loh P-R, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am J Hum Genet. 2019 Jan 3;104(1):65–75.
43. Rufinatscha K, Radlinger B, Dobner J, Folie S, Bon C, Profanter E, et al. Dipeptidyl peptidase-4 impairs insulin signaling and promotes lipid accumulation in hepatocytes.

Biochem Biophys Res Commun. 2017 Apr 1;485(2):366–71.
44. Pratley RE, Salsali A. Inhibition of DPP-4: a new therapeutic approach for the treatment of type 2 diabetes. Curr Med Res Opin. 2007 Apr;23(4):919–31.
45. Hollande C, Boussier J, Ziai J, Nozawa T, Bondet V, Phung W, et al. Inhibition of the dipeptidyl peptidase DPP4 (CD26) reveals IL-33-dependent eosinophil-mediated control of tumor growth. Nat Immunol. 2019 Mar 18;20(3):257–64.
46. Parker MD, Boron WF. The Divergence, Actions, Roles, and Relatives of Sodium-Coupled Bicarbonate Transporters. Physiol Rev. 2013 Apr;93(2):803–959.
47. Jacobs S, Ruusuvuori E, Sipila ST, Haapanen A, Damkier HH, Kurth I, et al. Mice with targeted Slc4a10 gene disruption have small brain ventricles and show reduced neuronal excitability. Proc Natl Acad Sci. 2008 Jan 8;105(1):311–6.
48. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018 Aug 23;50(8):1112–21.
49. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014 Jul 22;511(7510):421–7.
50. Dahlman I, Dicker A, Jiao H, Kere J, Blomqvist L, van Harmelen V, et al. A common haplotype in the G-protein-coupled receptor gene GPR74 is associated with leanness and increased lipolysis. Am J Hum Genet. 2007 Jun;80(6):1115–24.
51. Zhang L, Ip CK, Lee I-CJ, Qi Y, Reed F, Karl T, et al. Diet-induced adaptive thermogenesis requires neuropeptide FF receptor-2 signalling. Nat Commun. 2018 Dec 9;9(1):4722.
52. Nicklous DM, Simansky KJ. Neuropeptide FF exerts pro- and anti-opioid actions in the parabrachial nucleus to modulate food intake. Am J Physiol Integr Comp Physiol. 2003 Nov;285(5):R1046–54.
53. Cador M, Marco N, Stinus L, Simonnet G. Interaction between neuropeptide FF and opioids in the ventral tegmental area in the behavioral response to novelty. Neuroscience. 2002 Mar 12;110(2):309–18.

54. Panula P, Aarnisalo AA, Wasowicz K. Neuropeptide FF, a mammalian neuropeptide with multiple functions. Prog Neurobiol. 1996 Mar 1;48(4–5):461–87.
55. Huang EYK, Li JY, Wong CH, Tan PPC, Chen JC. Dansyl-PQRamide, a possible neuropeptide FF receptor antagonist, induces conditioned place preference. Peptides. 2002 Mar;23(3):489–96.
56. Bray L, Froment C, Pardo P, Candotto C, Burlet-Schiltz O, Zajac J-M, et al. Identification and Functional Characterization of the Phosphorylation Sites of the Neuropeptide FF $_2$ Receptor. J Biol Chem. 2014 Dec 5;289(49):33754–66.
57. Bekhouche M, Colige A. The procollagen N-proteinases ADAMTS2, 3 and 14 in pathophysiology. Matrix Biol. 2015 May 1;44–46:46–53.
58. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014 Nov 5;46(11):1173–86.
59. Rouault AAJ, Srinivasan DK, Yin TC, Lee AA, Sebag JA. Melanocortin Receptor Accessory Proteins (MRAPs): Functions in the melanocortin system and beyond. Biochim Biophys Acta - Mol Basis Dis. 2017 Oct 1;1863(10):2462–7.
60. Asai M, Ramachandrappa S, Joachim M, Shen Y, Zhang R, Nuthalapati N, et al. Loss of Function of the Melanocortin 2 Receptor Accessory Protein 2 Is Associated with Mammalian Obesity. Science (80-). 2013 Jul 19;341(6143):275–8.
61. Schonnop L, Kleinau G, Herrfurth N, Volckmar A-L, Cetindag C, Müller A, et al. Decreased melanocortin-4 receptor function conferred by an infrequent variant at the human melanocortin receptor accessory protein 2 gene. Obesity. 2016 Sep;24(9):1976–82.
62. Xie Y, Mustafa A, Yerzhan A, Merzhakupova D, Yerlan P, N Orakov A, et al. Nuclear matrix metalloproteinases: functions resemble the evolution from the intracellular to the extracellular compartment. Cell Death Discov. 2017 Dec 14;3(1):17036.
63. DeCoux A, Lindsey ML, Villarreal F, Garcia RA, Schulz R. Myocardial matrix metalloproteinase-2: inside out and upside down. J Mol Cell Cardiol. 2014 Dec;77:64–72.
64. Lalu MM, Pasini E, Schulze CJ, Ferrari-Vivaldi M, Ferrari-

Vivaldi G, Bachetti T, et al. Ischaemia–reperfusion injury activates matrix metalloproteinases in the human heart. Eur Heart J. 2005 Jan 1;26(1):27–35.
65. Mabuchi H, Nohara A, Inazu A. Cholesteryl ester transfer protein (CETP) deficiency and CETP inhibitors. Mol Cells. 2014 Nov;37(11):777–84.
66. Kao Y-H, Hsu J-C, Chen Y-C, Lin Y-K, Lkhagva B, Chen S-A, et al. ZFHX3 knockdown increases arrhythmogenesis and dysregulates calcium homeostasis in HL-1 atrial myocytes. Int J Cardiol. 2016 May 1;210:85–92.
67. Benjamin EJ, Rice KM, Arking DE, Pfeufer A, van Noord C, Smith A V, et al. Variants in ZFHX3 are associated with atrial fibrillation in individuals of European ancestry. Nat Genet. 2009 Aug 13;41(8):879–81.
68. Liu Y, Ni B, Lin Y, Chen X, Fang Z, Zhao L, et al. Genetic Polymorphisms in ZFHX3 Are Associated with Atrial Fibrillation in a Chinese Han Population. Ai X, editor. PLoS One. 2014 Jul 1;9(7):e101318.
69. Gudbjartsson DF, Holm H, Gretarsdottir S, Thorleifsson G, Walters GB, Thorgeirsson G, et al. A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. Nat Genet. 2009 Aug 13;41(8):876–8.
70. Vriens J, Voets T. Sensing the heat with TRPM3. Pflügers Arch - Eur J Physiol. 2018 May 5;470(5):799–807.
71. Vriens J, Owsianik G, Hofmann T, Philipp SE, Stab J, Chen X, et al. TRPM3 Is a Nociceptor Channel Involved in the Detection of Noxious Heat. Neuron. 2011 May 12;70(3):482–94.
72. Dyment DA, Terhal PA, Rustad CF, Tveten K, Griffith C, Jayakar P, et al. De novo substitutions of TRPM3 cause intellectual disability and epilepsy. Eur J Hum Genet. 2019 Oct 5;27(10):1611–8.
73. Global Health Observatory (GHO) data. WHO | Raised blood pressure [Internet]. WHO. World Health Organization; 2015.
74. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep 1;20(9):1297–303.
75. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012 Feb 4;9(2):179–81.

76. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.
77. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug 23;38(8):904–9.
78. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep 1;19(9):1655–64.
79. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol Ecol Resour. 2017 Jan 1;17(1):78–90.
80. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.
81. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep 1;38(16):e164–e164.
82. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 5;489(7414):57–74.

FIGURES

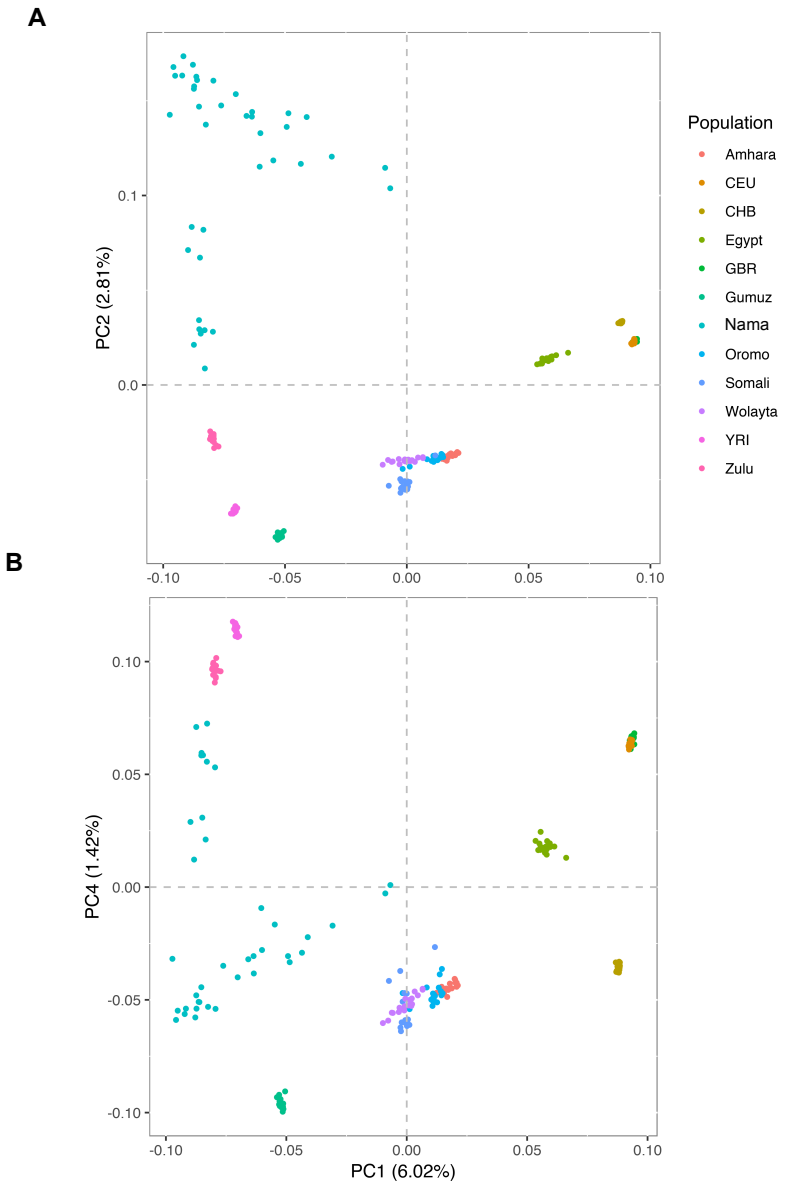


Figure 1: Principal Component Analysis of the Nama individuals and a set of 11 worldwide populations. A 5% minor allele frequency filter and linkage disequilibrium pruning of the data was done for the analysis (see Methods) A) Plot of PC1 and PC2 shows the differentiation between African and non-African populations and a strong differentiation between the Nama and other Africans. B) Plot of PC1 and PC4 shows that PC4 separates the East African (Gumuz) and West/South Bantu (YRI and Zulu) components.

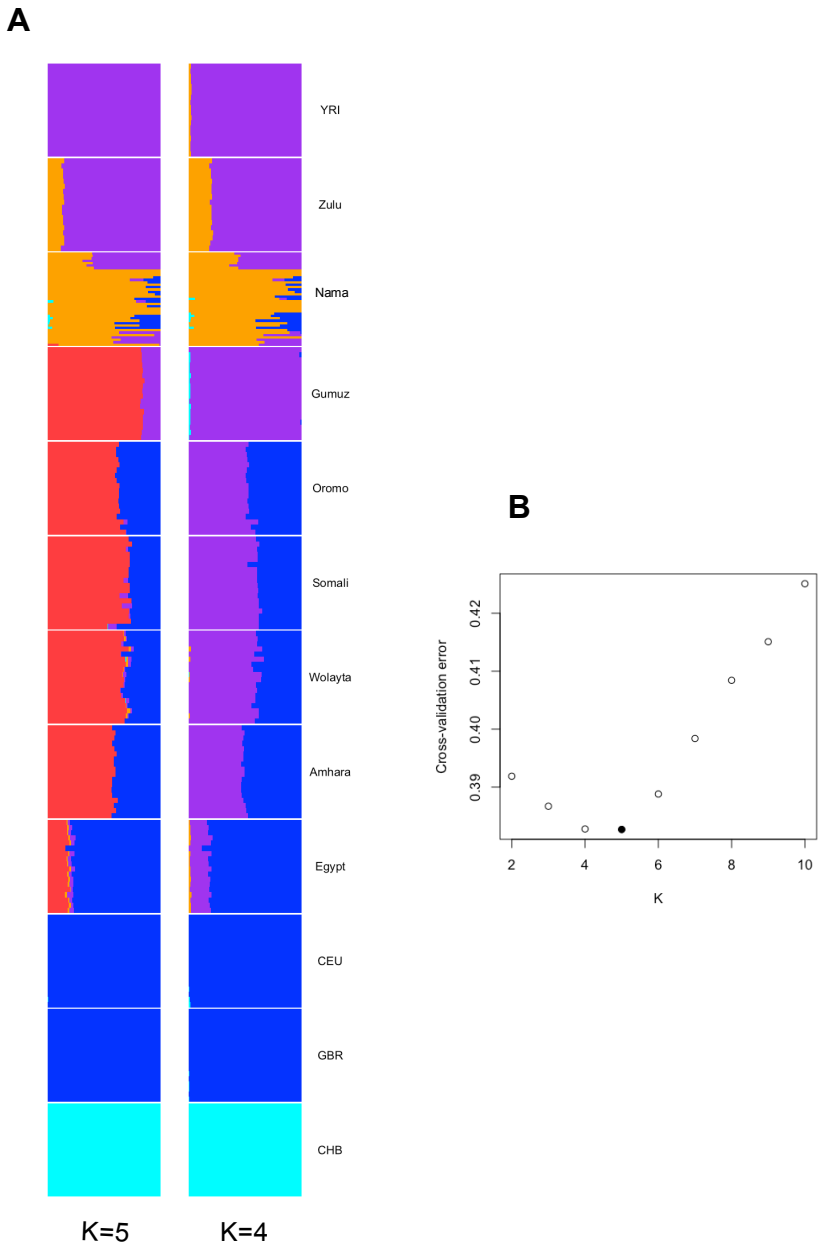


Figure 2: Ancestry proportions per individual inferred from ADMIXTURE (see Methods). A) ADMIXTURE plot at K=4 and K=5. The ancestry components obtained describe West Africans (purple), KhoeSan (yellow), East Africa (red), European (dark blue), East Asia (light blue).

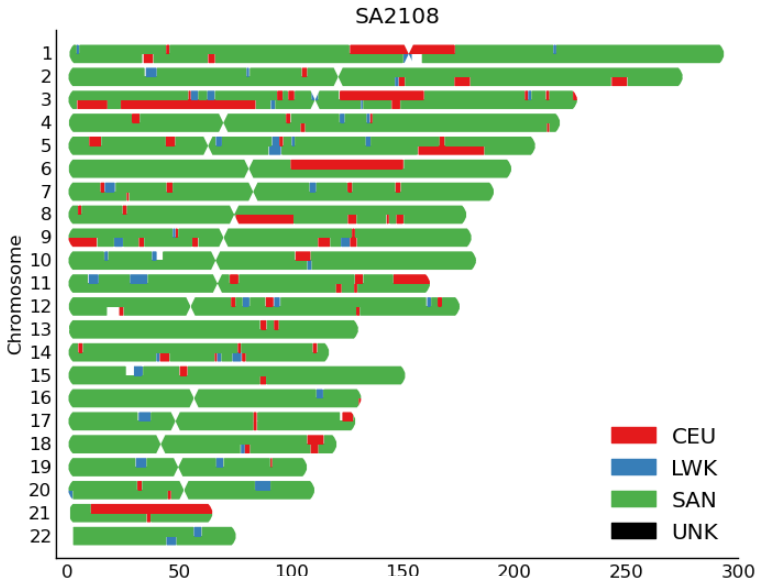


Figure 3: Painted chromosomes showing ancestry tracts contributed from ancestral populations in a Nama example individual.

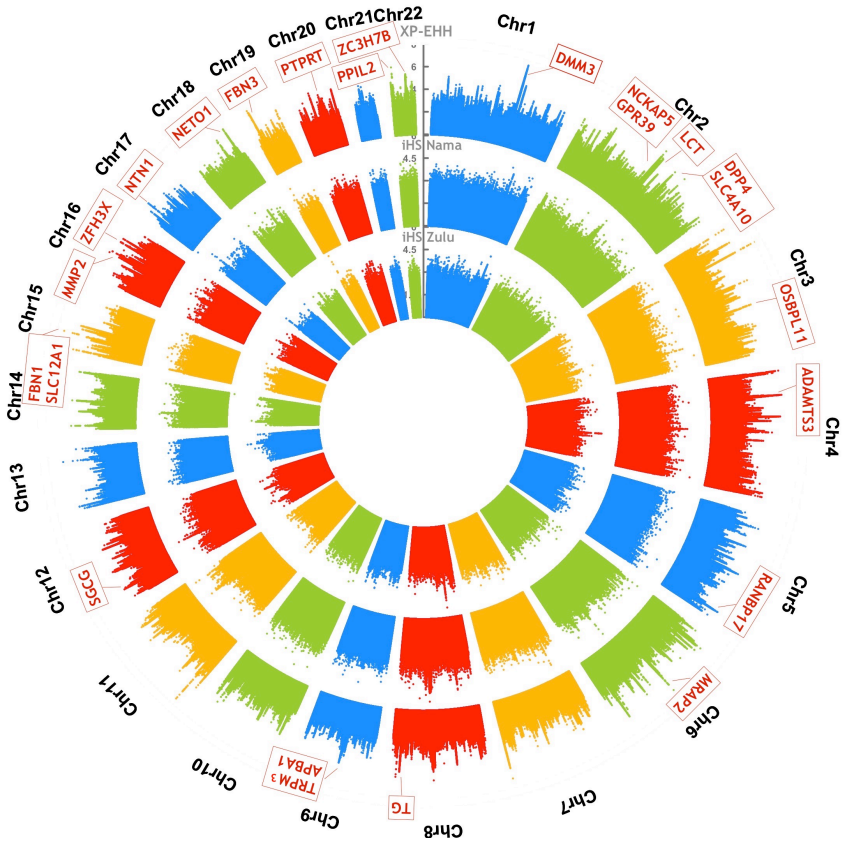


Figure 4: Circular Manhattan Plots of XP-EHH scores of the Nama and iHS scores of the Nama and Zulu. The XP-EHH analysis was done with the European-masked Nama genome and Zulu was used as the reference population so the effect of recent admixture was corrected. Red boxes contain the names of the highest scoring genes from the XP-EHH analysis.

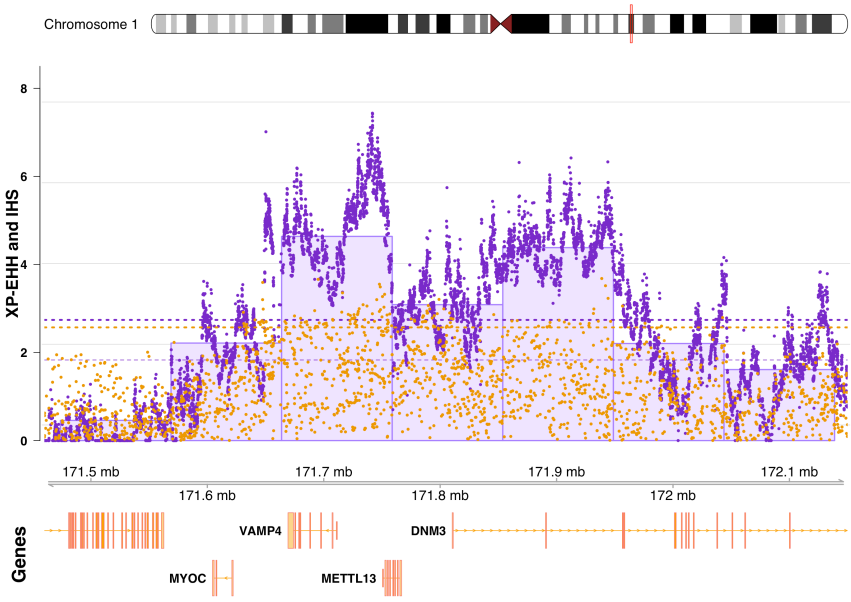


Figure 5: Genomic context of Region 1, a candidate of positive selection in the Nama. Each purple bar represents the mean XP-EHH positive score of a 100 kb window, purple dots correspond to the positive XP-EHH score of a single genetic position, yellow dots correspond to the absolute iHS scores of the Nama. Purple dotted lines indicate the 99th percentile thresholds of the positive XP-EHH scores per 100 kb windows (lower line) and the positive XP-EHH score per genetic position (upper line), yellow dotted lines indicate the 99th percentile thresholds of the absolute iHS scores.

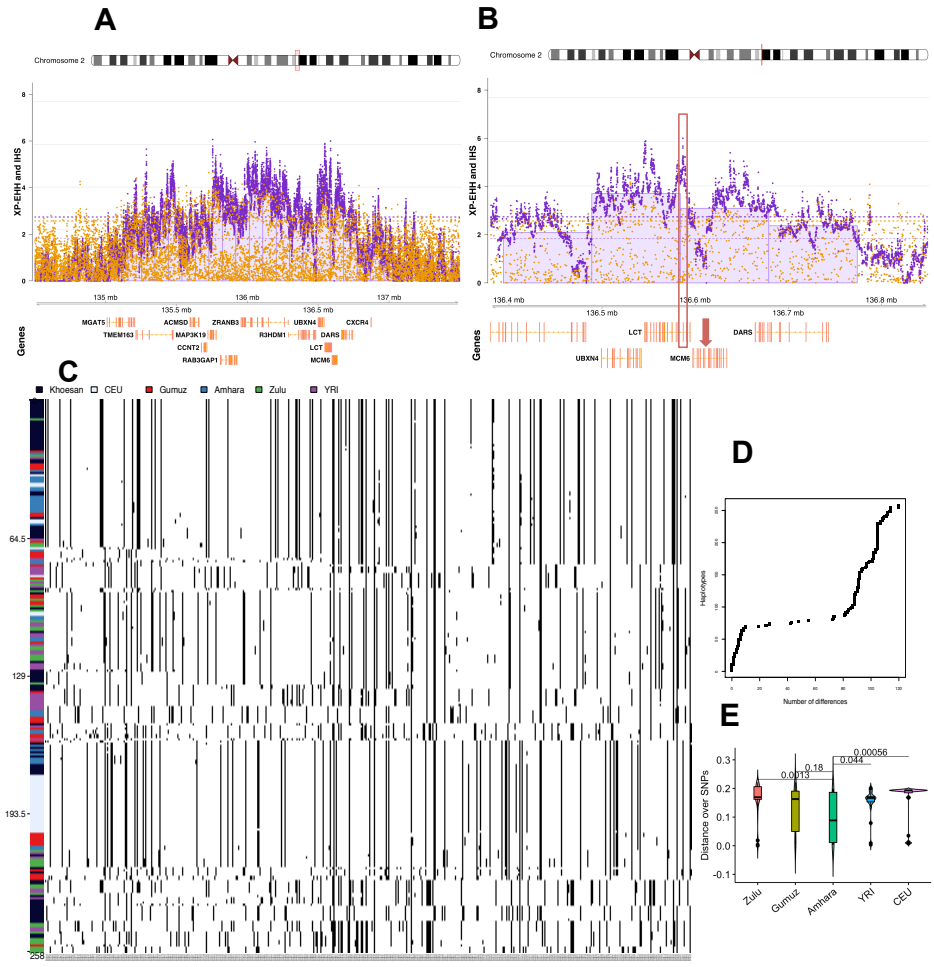
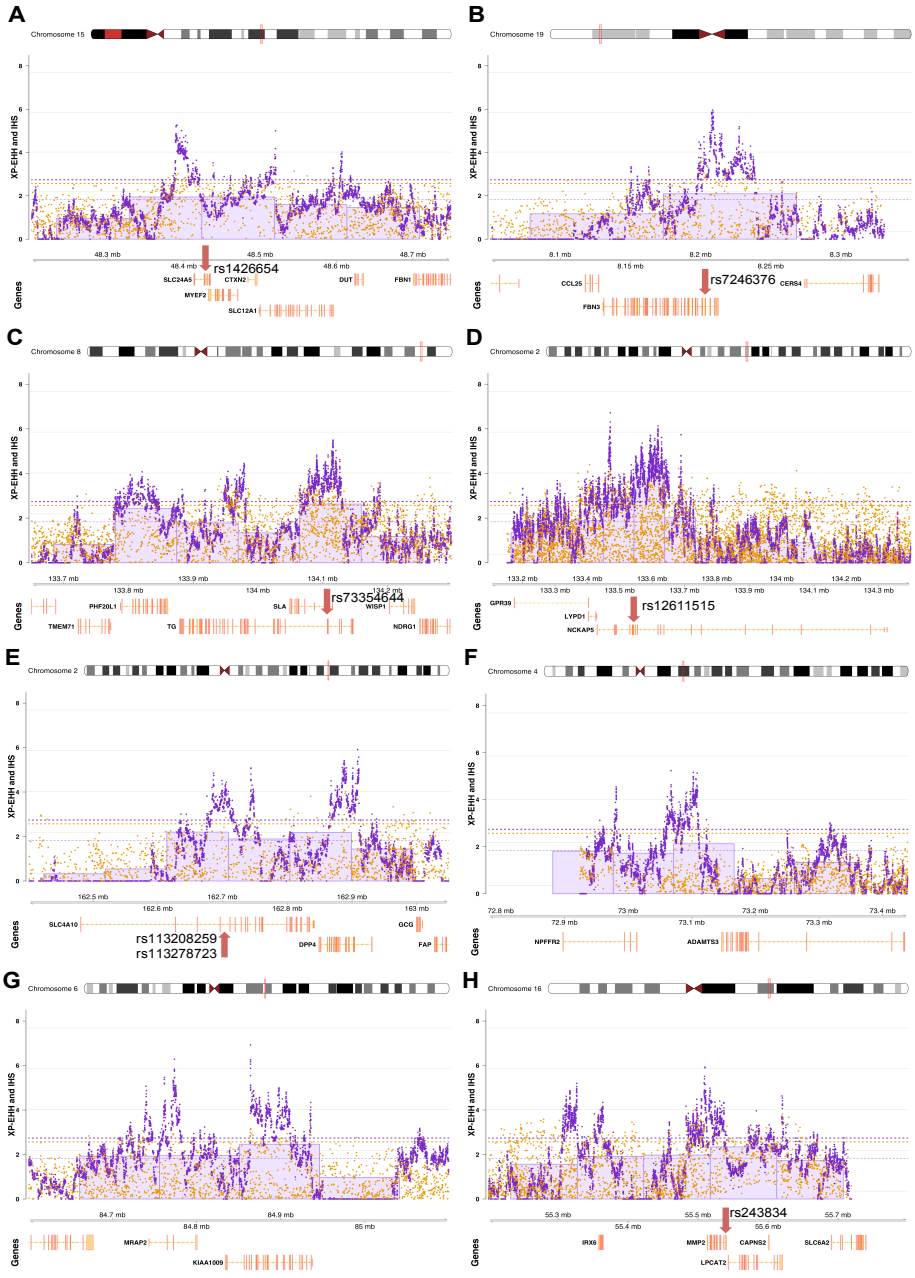


Figure 6 A-E: A) Genomic context of Region 2, a candidate of positive selection in the Nama around the *LCT* locus B) Zoom in of the *LCT* locus. The red arrow points to intron 13 of MCM6 gene. Each purple bar represents the mean XP-EHH positive score of a 100 kb window, purple dots correspond to the positive XP-EHH score of a single genetic position, yellow dots correspond to the absolute iHS scores of the Nama. Purple dotted lines indicate the 99th percentile thresholds of the positive XP-EHH
scores per 100 kb windows (lower line) and the positive XP-EHH score per genetic position (upper line), yellow dotted lines indicate the 99th percentile thresholds of the absolute iHS scores. C) Haplotype visualization of the LCT locus and its regulatory region. The number of differences among haplotypes is used to perform the clustering of haplotypes, hence haplotypes are plotted from the most similar (upper part of figure) to the least similar (bottom part of figure) D) Number of differences between the reference non-admixed Nama haplotype and the haplotypes belonging to the Nama and other populations (Zulu, Gumuz, Amhara, YRI and CEU) E) Distribution of the number of differences between the Nama reference and a population. Kruskal-Wallis Test was used to analyse the differences of distance from the Nama reference according to the population, significant differences where found (chi-squared = 20.438, df = 4, p=0.00041) and Wilcoxon test and the statistical significance of the difference between distributions. Wilcoxon test indicated that the distance of Amhara was significantly lower than the distance of Zulu (p=0.0013), YRI (0.044) and CEU (p=0.00056).


A
Chromosome 15
XP-EHH and IHS
Genes
48.3 mb
48.4 mb
48.5 mb
48.6 mb
48.7 mb
rs1426654
SLC24A5
CTXN2
DUT
FBN1
MYEF2
SLC12A1
B
Chromosome 19
8.1 mb
8.15 mb
8.2 mb
8.25 mb
8.3 mb
rs7246376
CCL25
CERS4
FBN3
C
Chromosome 8
133.7 mb
133.8 mb
133.9 mb
134 mb
134.1 mb
rs73354644
PHF20L1
SLA
WISP1
TMEM71
TG
NDRG1
D
Chromosome 2
133.2 mb
133.3 mb
133.4 mb
133.5 mb
133.6 mb
133.7 mb
133.8 mb
133.9 mb
134 mb
134.1 mb
134.2 mb
134.3 mb
rs12611515
GPR39
LYPD1
NCKAP5
E
Chromosome 2
162.5 mb
162.6 mb
162.7 mb
162.8 mb
162.9 mb
163 mb
rs113208259
rs113278723
SLC4A10
GCG
DPP4
FAP
F
Chromosome 4
72.8 mb
72.9 mb
73 mb
73.1 mb
73.2 mb
73.3 mb
73.4 mb
NPFFR2
ADAMTS3
G
Chromosome 6
84.7 mb
84.8 mb
84.9 mb
85 mb
MRAP2
KIAA1009
H
Chromosome 16
55.3 mb
55.4 mb
55.5 mb
55.6 mb
55.7 mb
rs243834
IRX6
MMP2
CAPNS2
SLC6A2
LPCAT2

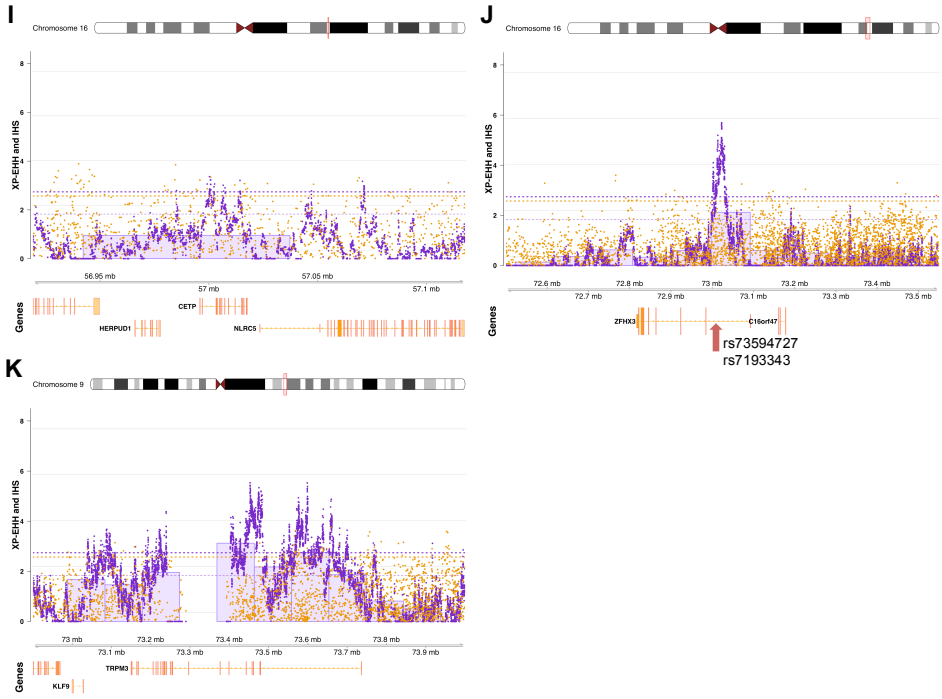


Figure 7 A-K: Genomic context of Regions 3 to 13, candidates of positive selection in the Nama. Each purple bar represents the mean XP-EHH positive score of a 100 kb window, purple dots correspond to the positive XP-EHH score of a single genetic position, yellow dots correspond to the absolute iHS scores of the Nama. Purple dotted lines indicate the 99th percentile thresholds of the positive XP-EHH scores per 100 kb windows (lower line) and the positive XP-EHH score per genetic position (upper line), yellow dotted lines indicate the 99th percentile thresholds of the absolute iHS scores.

A

Variants	Annotated	Not Annotated	Threshold (XP-EHH score)
436 250	248 472	187 778	2.74
%	57.0	43.0	

B

100 kb windows	Annotated	Not Annotated	Threshold (mean XP-EHH)
271	237	34	1.83
%	87.0	13.0	

	Coding Elements	Protein	lincRNA	antisense	pseudogene	processed transcript	Sense intron	misc_RNA	miRNA	sense overl	snRNA	snoRNA	polym pseudo	rRNA
Variants	2812	1944	353	246	148	49	16	13	12	10	8	6	5	2
	%	69.1	12.6	8.7	5.3	1.7	0.6	0.5	0.4	0.4	0.3	0.2	0.2	0.1
Windows	729	356	61	63	120	3	7	26	29	1	19	8	1	1
	%	49	8.4	8.6	16.5	0.41	1	3.6	4	0.14	2.6	1.1	0.14	0.14

Table 1: Summary and annotations of the top 1% scoring variants and 100 kb windows. A) Top, summary of the annotations of the highest scoring XP-EHH 1% variants. Among the 436.250 SNPs, 57% of them have a coding gene annotation. The 1 percent threshold falls at an XP-EHH score of 2.74. Bottom, information for 100 kb windows, using the mean XP-EHH across the window. The 1 percent threshold for 100 kb windows is a mean XP-EHH score of 1.83. B) Types of coding elements in the top 1% selection signals. Top, using SNPs and bottom, using 100 kb windows. Abbreviations: lincRNA, Long intergenic non-coding RNAs; antisense, transcripts that overlap the genomic span (i.e. exon or introns) of a protein-coding locus on the opposite strand; processed transcript, transcript that doesn't contain an open reading frame; miRNA, microRNA precursors; misc_RNA, miscellaneous other RNA; snRNA, small nuclear RNA; snoRNA, small nucleolar RNA; rRNA, ribosomal RNA; sense intronic, long non-coging transcript in introns of a coding gene that does not overlap any exons; sense overlapping, a long non-coding transcript that contains a coding gene in its intron on the same strand; polym pseudo, polymorphic pseudogene.

Trait	p-value	p-adjusted (FDR)	Odds ratio	taSNP hits	taSNP total
Blood Viscosity	3.85e-05	0.011	57	2	6
Coronary Disease	8.99e-05	0.013	41	2	8
Coronary Disease through EC 3.1.1.47 Activity	0.00017	0.016	32	2	10
Skin Pigmentation	0.00048	0.0331	63	1	3
Lipoproteins. HDL	0.00081	0.0448	18	2	17
Lipoproteins. LDL	0.00130	0.0604	38	1	5
Heart Function Tests	0.00190	0.0756	13	2	23
Atrial Fibrillation	0.00252	0.0876	27	1	7

Table 2: Trait enrichment analysis with traseR of the 1% extreme XP-EHH scoring positions. The most significant GWAS traits are reported with their corresponding p-values and q-values (obtained from p-values after FDR correction). Odds ratios were calculated after the number of trait-associated SNP among our selection candidates (taSNP hits) for a specific trait**,** the number of trait-associated SNPs for a trait across whole genome (taSNP total) and the total number of SNPs putatively under selection and in the reference. EC 3.1.1.47 corresponds to 1-Alkyl-2-acetylglycerophosphocholine Esterase.

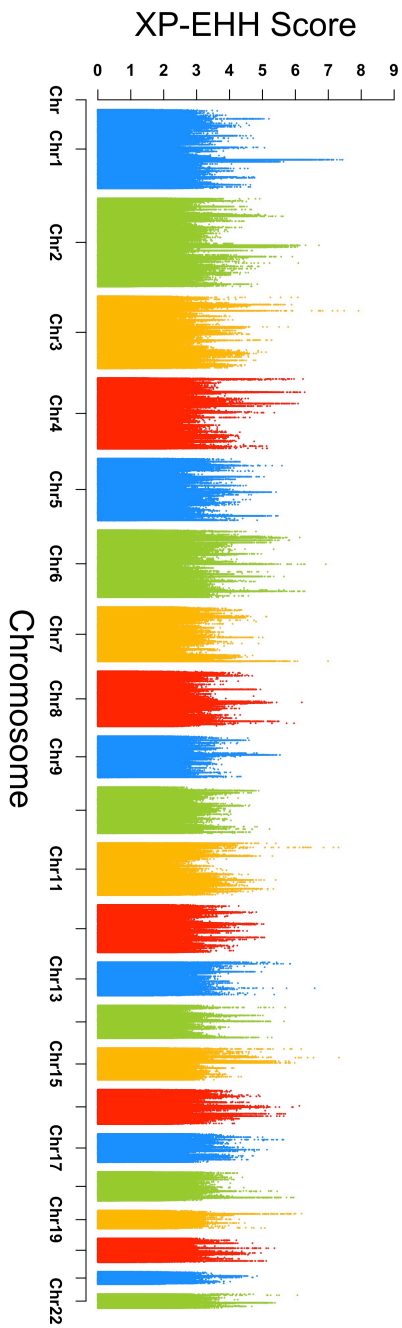
Geneset	Ref	Obs	Exp	Ratio	P-value	FDR	Description
hsa04713	88	25	11.0	2.3	4.60e-05	8.92e-03	Circadian entrainment
hsa04724	105	28	13.1	2.1	5.77e-05	8.92e-03	Glutamatergic synapse Retrograde endocannabinoid signalling
hsa04723	90	23	11.3	2.0	5.10e-04	3.85e-02	
hsa04911	74	20	9.3	2.2	5.37e-04	3.85e-02	Insulin secretion
hsa04720	59	17	7.4	2.3	6.23e-04	3.85e-02	Long-term potentiation

Table 3: KEGG pathway enrichment analysis with WebGestalt using the gene annotations of the 1% extreme XP-EHH scoring positions. The background to perform the enrichment analysis was built after annotating all positions with an XP-EHH positive score. The obtained p-values were then corrected for multiple testing with an FDR.

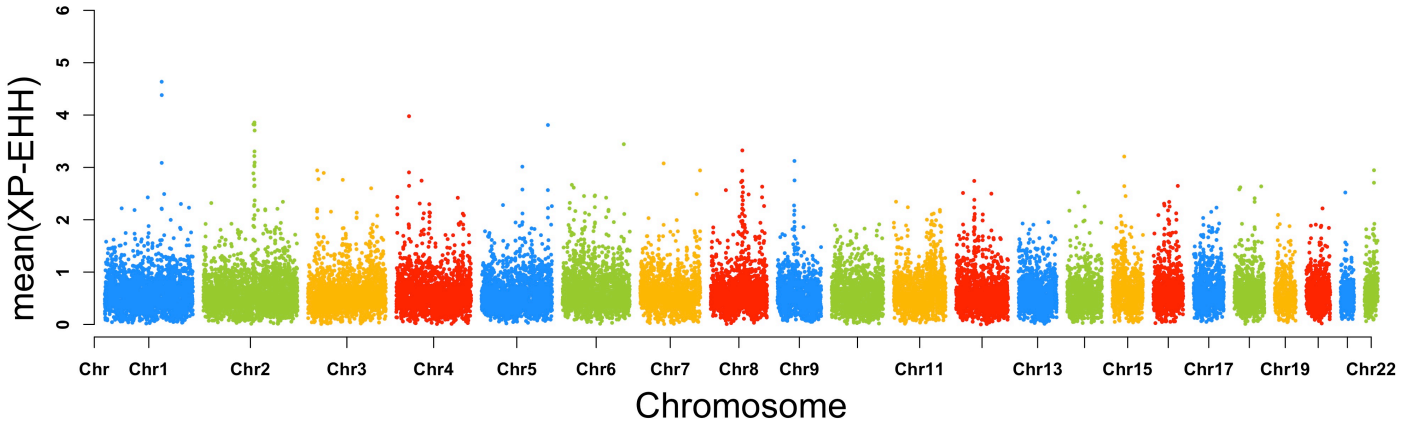
Region	Chromosome	Start window	Stop window	Mean XP-EHH	Max XP-EHH	Rank	Gene
Region-1	chr1	171663805	171763805	4,64	7,43	1	*METTL13, VAMP4*
	chr1	171853805	171953805	4,38	6,42	2	*DNM3*
	chr1	171568805	171668805	2,21	7,01	105	*MYOC*
Region-2	chr2	136014022	136114022	3,86	5,96	4	*ZRANB3*
	chr2	136489022	136589022	3,71	6,00	8	*LCT, UBXN4*
	chr2	136204022	136304022	3,30	5,16	11	*R3HDM1*
	chr2	135729022	135829022	3,22	6,06	12	*MAP3K19, RAB3GAP1*
	chr2	136584022	136684022	3,09	6,00	15	*MCM6, DARS*
	chr2	135444022	135544022	2,77	5,13	29	*TMEM163*
	chr2	135634022	135734022	2,28	3,85	86	*ACMSD, CCNT2*
	chr2	135159022	135259022	2,01	4,37	164	*MGAT5*
Region-3	chr15	48423431	48523431	1,95	5,00	189	*SLC24A5, CTXN2, MYEF2, SLC12A1*
Region-4	chr19	8173367	8273367	2,09	5,96	144	*FBN3*
Region-5	chr8	134064587	134164587	2,63	5,49	44	*TG, SLA*
	chr8	133779587	133879587	2,43	4,07	67	*PHF20L1*
Region-6	chr2	133544022	133644022	3,82	6,13	5	*NCKAP5*
	chr2	133354022	133454022	1,88	3,88	235	*GPR39, LYPD1*
Region-7	chr2	162614022	162714022	2,20	4,85	110	*SLC4A10*
	chr2	162804022	162904022	2,18	5,41	116	*DPP4*
Region-8	chr4	73071487	73171487	2,12	5,15	130	*ADAMTS3*
Region-9	chr6	84851015	84951015	2,46	6,92	61	*KIAA1009*
	chr6	84756015	84856015	1,95	6,29	188	*MRAP2*
Region-10	chr16	55517246	55617246	2,34	4,41	78	*MMP2, CAPNS2, LPCAT2*
	chr16	55327246	55427246	1,92	4,34	205	*IRX6*
Region-11	chr16	56942246	57042246	0,96	3,36	3436	*CETP, HERPUD1*
Region-12	chr16	72997246	73097246	2,12	5,69	131	*ZFHX3*
Region-13	chr9	73368433	73468433	3,12	5,53	13	*TRPM3*

Table 4 Summary of the candidate positive selection regions. For each region, rows correspond to a 100 kb window found in the reported selection regions. For each 100 kb window we report the mean of XP-EHH score, maximum XP-EHH over all the SNPs found inside, the ranking of the window (number one being the highest scoring window) and the genes inside the window. We included regions of interest that either their window mean or maximum XP-EHH score were higher than the 1 percent per window or per SNP threshold respectively (reported in Table 1A).

SUPPLEMENTARY FIGURES



Supplementary Figure 1: Manhattan plot of the XP-EHH analysis of the Nama population. The analysis was done masking the European haplotypes from the Nama genomes and the Zulu population was used as a reference to avoid admixture biases.



Supplementary Figure 2: Manhattan plot of the mean XP-EHH per 100 kb window of the Nama population. Same analysis as in Supplementary Figure 1, just averaging across a 100 kb window.

SUPPLEMENTARY TABLES

Trait	SNP id	Chr	Position	Gene	Annotation	Anc	Der	Zulu	Nama	Gumuz	CHB	CEU
Blood Viscosity	rs10518934	chr15	58088064	*GRINL1A/ MYZAP ALDH1A2/ POLR2M*	Intergenic	A	G	0.15	0.05	0.06	0	0
	rs10518936	chr15	58090523		Intergenic	G	T	0.14	0.05	0.06	0	0
Coronary Disease	rs12740374	chr1	109817590	*CELSR2*	UTR-3'	G	**T**	0.38	0.54	0.22	0.05	0.31
	rs599839	chr1	109822166	*PSRC1*	NearGene3'	G	**A**	0.76	0.77	0.67	0.06	0.35
Coronary Disease through EC 3.1.1.47 Activity	rs599839	chr1	109822166	*PSRC1 SORT1 CELSR2*	NearGene3'	G	A	0.76	0.77	0.67	0.06	0.35
	rs7528419	chr1	109817192		UTR-3	G	A	0.60	0.45	0.64	0.95	0.69
Skin Pigmentation	rs1834640	chr15	48392165	*SLC24A5*	Intergenic	G	A	0.97	0.54	0.97	0.91	0.01
Lipoproteins. HDL	rs7499892	chr16	57006590	*CETP*	Intron	T	C	0.55	0.71	0.61	0.79	0.78
	rs9989419	chr16	56985139		Intergenic	G	A	0.38	0.65	0.56	0.73	0.55
Lipoproteins. LDL	rs646776	chr1	109818530	*CELSR2/SORT1*	NearGene3'	T	C	0.54	0.42	0.53	0.95	0.69
Heart Function Tests	rs7784776	chr7	46620145	*TTC4P1*	Intergenic	A	G	0.35	0.58	0.22	0.83	0.38
Heart Function Tests	rs991014	chr18	42439886	*SETBP1*	Intron	C	T	0.18	0.06	0.03	0.33	0.42
Atrial Fibrillation	rs7193343	chr16	73029160	*ZFHX3*	Intron	T	C	0.12	0.05	0.14	0.64	0.18

Supplementary Table 1: Associated SNPs from traseR enrichment reported categories and derived allele frequencies in 7 worldwide populations.

Geneset	Ref	Obs	Exp	Ratio	P-value	FDR	Description
GO:0007399	1811	301	216.0	1.4	1.16e-10	1.35e-06	nervous system development
GO:0048667	421	93	50.2	1.9	1.66e-09	9.64e-06	cell morphogenesis involved in neuron differentiation
GO:0048812	465	97	55.5	1.7	1.66e-08	4.84e-05	neuron projection morphogenesis
GO:0031175	687	130	82.0	1.6	3.38e-08	6.12e-05	neuron projection development
GO:0007409	349	77	41.6	1.8	4.53e-08	6.21e-05	axonogenesis
GO:0048666	810	146	96.6	1.5	1.14e-07	1.02e-04	neuron development
GO:0061564	379	80	45.2	1.8	1.83e-07	1.52e-04	axon development
GO:0097485	189	47	22.5	2.1	5.66e-07	3.46e-04	neuron projection guidance
GO:0030182	1018	172	121.4	1.4	7.24e-07	4.21e-04	neuron differentiation
GO:0048699	1122	186	133.8	1.4	9.18e-07	4.85e-04	generation of neurons
GO:0022008	1205	196	143.7	1.4	1.72e-06	8.71e-04	neurogenesis
GO:0097503	20	11	2.4	4.6	4.04e-06	1.94e-03	sialylation
GO:0016358	155	36	18.5	1.9	5.60e-05	2.03e-02	dendrite development
GO:0006688	25	11	3.0	3.7	6.06e-05	2.03e-02	glycosphingolipid biosynthetic process
GO:0050954	135	32	16.1	2.0	9.41e-05	2.88e-02	sensory perception of mechanical stimulus
GO:0006928	1499	224	178.8	1.3	1.14e-04	3.40e-02	movement of cell or subcellular component
GO:0007610	461	82	55.0	1.5	1.18e-04	3.42e-02	behavior
GO:0007605	120	29	14.3	2.0	1.37e-04	3.89e-02	sensory perception of sound
GO:0048813	92	24	11.0	2.2	1.45e-04	4.01e-02	dendrite morphogenesis

Supplementary Table 2 Gene Ontology Biological Process enrichment analysis with WebGestalt of the 1% extreme XP-EHH scoring positions.

Geneset	Ref	Obs	Exp	Ratio	P-value	FDR	Description
GO:0005509	583	103	65.5	1.6	1.59e-06	1.63e-03	calcium ion binding
GO:0008373	20	11	2.2	4.9	2.23e-06	1.63e-03	sialyltransferase activity
GO:0005216	361	69	40.6	1.7	5.98e-06	1.94e-03	ion channel activity
GO:0015075	735	121	82.6	1.5	7.59e-06	1.98e-03	ion transmembrane transporter activity
GO:0097367	1846	263	207.5	1.3	1.06e-05	2.06e-03	carbohydrate derivative binding
GO:0003779	337	64	37.9	1.7	1.59e-05	2.91e-03	actin binding
GO:0022838	374	69	42.0	1.6	2.04e-05	3.42e-03	substrate-specific channel activity
GO:0042805	28	12	3.1	3.8	2.11e-05	3.42e-03	actinin binding
GO:0022857	850	134	95.5	1.4	2.28e-05	3.44e-03	transmembrane transporter activity
GO:0005524	1237	184	139.0	1.3	2.46e-05	3.44e-03	ATP binding
GO:0022891	791	126	88.9	1.4	2.48e-05	3.44e-03	substrate-specific transmembrane transporter activity
GO:0051393	21	10	2.4	4.2	3.42e-05	3.49e-03	alpha-actinin binding
GO:0015267	395	70	44.4	1.6	6.80e-05	6.20e-03	channel activity
GO:0022803	396	70	44.5	1.6	7.37e-05	6.52e-03	passive transmembrane transporter activity
GO:0003828	6	5	0.7	7.4	9.70e-05	7.87e-03	alpha-N-acetylneuraminate alpha-2.8-sialyltransferase activity
GO:0005261	258	49	29.0	1.7	1.52e-04	1.17e-02	cation channel activity
GO:0022836	279	52	31.4	1.7	1.59e-04	1.19e-02	gated channel activity
GO:0046914	1107	162	124.4	1.3	1.80e-04	1.28e-02	transition metal ion binding
GO:0072509	150	32	16.9	1.9	2.48e-04	1.68e-02	divalent inorganic cation transmembrane transporter activity
GO:0008324	539	87	60.6	1.4	2.89e-04	1.92e-02	cation transmembrane transporter activity
GO:0008066	26	10	2.9	3.4	3.04e-04	1.96e-02	glutamate receptor activity
GO:0008270	907	135	101.9	1.3	3.09e-04	1.96e-02	zinc ion binding
GO:0046873	379	65	42.6	1.5	3.18e-04	1.97e-02	metal ion transmembrane transporter activity
GO:0005234	19	8	2.1	3.7	6.03e-04	3.39e-02	extracellular-glutamate-gated ion channel activity
GO:0015085	113	25	12.7	2.0	6.46e-04	3.49e-02	calcium ion transmembrane transporter activity
GO:0005262	101	23	11.4	2.0	6.84e-04	3.63e-02	calcium channel activity
GO:0005499	5	4	0.6	7.1	7.24e-04	3.71e-02	vitamin D binding
GO:0051371	8	5	0.9	5.6	7.45e-04	3.72e-02	muscle alpha-actinin binding

Supplementary Table 3 Gene Ontology Molecular Function enrichment analysis with WebGestalt of the 1% extreme XP-EHH scoring positions.

Rank	Chromosome	Start window	End window	Mean XP-EHH	Max XP-EHH	Gene
1	chr1	171663805	171763805	4,64	7,43	*METTL13, VAMP4*
2	chr1	171853805	171953805	4,38	6,42	*DNM3*
4	chr2	136014022	136114022	3,86	5,96	*ZRANB3*
5	chr2	133544022	133644022	3,82	6,13	*NCKAP5*
7	chr5	170403115	170503115	3,81	5,43	*RANBP17*
8	chr2	136489022	136589022	3,71	6,00	*LCT, UBXN4*
11	chr2	136204022	136304022	3,30	5,16	*R3HDM1*
12	chr2	135729022	135829022	3,22	6,06	*MAP3K19, RAB3GAP1*
13	chr15	547884311	548884311	3,21	5,80	*UNC13C*
14	chr9	733684333	734684333	3,12	5,53	*TRPM3*
15	chr2	136584022	136684022	3,09	6,00	*DARS, MCM6*
20	chr5	107513115	107613115	3,01	5,41	*FBXL17*
21	chr22	41689709	41789709	2,95	5,38	*TEF, ZC3H7B*
22	chr3	234749722	235749722	2,94	5,89	*UBE2E2*
23	chr7	158479844	158579844	2,94	6,99	*ESYT2, NCAPG2*
26	chr3	398149722	399149722	2,89	7,92	*MYRIP*
29	chr2	135444022	135544022	2,77	5,13	*TMEM163*
34	chr12	496883111	497883111	2,74	4,85	*PRPH, C1QL4, TROAP, DNAJC22, SPATS2*
35	chr8	816245877	817245877	2,72	5,06	*ZNF704*
36	chr22	41594709	41694709	2,71	5,27	*CHADL, L3MBTL2, AL035681.1, RANGAP1*
38	chr2	136299022	136399022	2,65	4,49	*R3HDM1*
42	chr15	548834311	549834311	2,64	4,98	*UNC13C*
43	chr18	704366300	705366300	2,64	5,95	*NETO1*
44	chr8	134064587	134164587	2,63	5,49	*TG, SLA*
45	chr8	858045877	859045877	2,63	4,53	*RALYL*
46	chr18	147666300	148666300	2,62	4,01	*ANKRD30B*
49	chr18	121066300	122066300	2,58	4,05	*ANKRD62*
50	chr5	107608115	107708115	2,58	5,20	*FBXL17*

Supplementary Table 4 Ranking of XP-EHH scores per 100 kb windows containing protein coding genes. The SNPs have been grouped in regions of 100Kb taking the mean and maximum XP-EHH score to have an overall view of a window.

Region	Chromosome	Start window	Stop window	Mean XP-EHH	Max XP-EHH	Rank	Gene
Region-1	chr1	171663805	171763805	4,64	7,43	1	*METTL13, VAMP4*
	chr1	171853805	171953805	4,38	6,42	2	*DNM3*
	chr1	171568805	171668805	2,21	7,01	105	*MYOC*
Region-2	chr2	136014022	136114022	3,86	5,96	4	*ZRANB3*
	chr2	136489022	136589022	3,71	6,00	8	*LCT, UBXN4*
	chr2	136204022	136304022	3,30	5,16	11	*R3HDM1*
	chr2	135729022	135829022	3,22	6,06	12	*MAP3K19, RAB3GAP1*
	chr2	136584022	136684022	3,09	6,00	15	*MCM6, DARS*
	chr2	135444022	135544022	2,77	5,13	29	*TMEM163*
	chr2	135634022	135734022	2,28	3,85	86	*ACMSD, CCNT2*
	chr2	135159022	135259022	2,01	4,37	164	*MGAT5*
Region-3	chr15	48423431	48523431	1,95	5,00	189	*SLC24A5, CTXN2, MYEF2, SLC12A1*
Region-4	chr19	8173367	8273367	2,09	5,96	144	*FBN3*
Region-5	chr8	134064587	134164587	2,63	5,49	44	*TG, SLA*
	chr8	133779587	133879587	2,43	4,07	67	*PHF20L1*
Region-6	chr2	133544022	133644022	3,82	6,13	5	*NCKAP5*
	chr2	133354022	133454022	1,88	3,88	235	*GPR39, LYPD1*
Region-7	chr2	162614022	162714022	2,20	4,85	110	*SLC4A10*
	chr2	162804022	162904022	2,18	5,41	116	*DPP4*
Region-8	chr4	73071487	73171487	2,12	5,15	130	*ADAMTS3*
Region-9	chr6	84851015	84951015	2,46	6,92	61	*KIAA1009*
	chr6	84756015	84856015	1,95	6,29	188	*MRAP2*
Region-10	chr16	55517246	55617246	2,34	4,41	78	*MMP2, CAPNS2, LPCAT2*
	chr16	55327246	55427246	1,92	4,34	205	*IRX6*
Region-11	chr16	56942246	57042246	0,96	3,36	3436	*CETP, HERPUD1*
Region-12	chr16	72997246	73097246	2,12	5,69	131	*ZFHX3*
Region-13	chr9	73368433	73468433	3,12	5,53	13	*TRPM3*
Region-14	chr5	170403115	170503115	3,81	5,43	7	*RANBP17*
	chr5	170213115	170313115	2,57	5,47	51	*GABRP*
Region-15	chr22	22024709	22124709	1,82	6,06	284	*PPIL2, YPEL1, MAPK1*
Region-16	chr22	41689709	41789709	2,95	5,38	21	*ZC3H7B, TEF*
Region-17	chr18	70436630	70536630	2,64	5,95	43	*NETO1*
Region-18	chr20	40847955	40947955	1,89	4,95	233	*PTPRT*
Region-19	chr3	125219972	125319972	2,03	5,28	158	*OSBPL11, SNX4*
Region-20	chr1	109818805	109918805	1,67	4,35	427	*PSRC1, SORT1, MYBPHL*
	chr1	109723805	109823805	1,26	4,73	1307	*CELSR2*
Region-21	chr15	58018431	58118431	1,88	5,38	234	*POLR2M*
	chr15	57828431	57928431	1,43	3,44	823	*GCOM1, MYZAP*
	chr15	58303431	58403431	1,08	2,79	2248	*ALDH1A2*
Region-22	chr17	8855457	8955457	1,14	5,00	1860	*NTN1, PIK3R5*

Supplementary Table 5 Summary of the candidate selection regions. For each region, rows correspond to a 100 kb window found in the reported selection regions. For each 100 kb window we report the mean of XP-EHH score and maximum XP-EHH over all the SNPs found inside. We included regions of interest that either their window mean or maximum XP-EHH score were higher than the 1 percent per window or per SNP threshold respectively (reported in Tables 2a and b). We also report the name of the genes found and the ranking of the mean XP-EHH for each window (the number one being the highest scoring window).

Chr	Position	Ref	Alt	Annot	Gene	rsID	Zulu	Nama	Gumuz	Amhara	Wolayta	CEU	CHB	Mean dist
2	133541575	C	T	exonic	*NCKAP5*	rs12611515	0.5385	0.7949	0.2222	0.3182	0.5238	0.5128	0.2821	0.395
19	8203328	G	A	exonic	*FBN3*	rs7246376	0.4744	0.7051	0.1389	0.3409	0.3333	0.5513	0.1282	0.377
1	109479978	G	T	exonic	*CLCC1*	rs168107	0.5897	0.7821	0.6667	0.4091	0.5	0.05128	0.2692	0.368
22	50315363	C	A	exonic	*CRELD2*	rs8139422	0.6538	0.7179	0.6667	0.4091	0.381	0.07692	0	0.353
2	201736166	A	C	exonic	*PPIL3*	rs7562391	0.5385	0.5897	0.3333	0.25	0.2381	0.01282	0.1795	0.331
15	93616975	A	G	exonic	*RGMA*	rs4598860	0.2179	0.141	0.2222	0.4773	0.381	0.9359	0.4872	0.313
18	6977844	A	G	exonic	*LAMA1*	rs671871	0.7821	0.7564	0.6667	0.5227	0.4762	0.1154	0.1667	0.310
6	28270047	C	G	exonic	*PGBD1*	rs6456811	0.6923	0.6667	0.5	0.4545	0.3333	0.1538	0.0641	0.309
6	28264692	C	G	exonic	*PGBD1*	rs3800325	0.6923	0.6667	0.5	0.4545	0.3571	0.1538	0.0641	0.305
11	100226883	T	A	exonic	*CNTN5*	rs1216183	0.6154	0.7436	0.75	0.3864	0.5	0.2692	0.1282	0.304
16	56673227	C	A	exonic	*MT1A*	rs11640851	0.1795	0.141	0.3056	0.4545	0.4524	0.5385	0.7179	0.300
11	124135215	T	G	exonic	*OR8G5*	rs11219544	0.3205	0.4231	0.1944	0.06818	0.07143	0.07692	0.02564	0.297
11	124135677	A	G	exonic	*OR8G5*	rs10893192	0.3205	0.4231	0.1944	0.06818	0.07143	0.07692	0.02564	0.297
5	140222641	A	G	exonic	*PCDHA8*	rs6580012	0.3077	0.1538	0.4444	0.4545	0.5238	0.4103	0.5385	0.293
11	124095647	A	C	exonic	*OR8G2*	rs2466615	0.1154	0.07692	0.1389	0.5	0.381	0.5256	0.4872	0.281
1	9009444	A	C	exonic	*CA6*	rs2274328	0.5256	0.6795	0.25	0.4545	0.2619	0.4487	0.4872	0.275
2	136555659	T	C	exonic	*LCT*	rs2322659	0.2436	0.141	0.3333	0.25	0.3095	0.5	0.8077	0.266
7	158536267	T	C	exonic	*ESYT2*	rs2305473	0.5769	0.6667	0.3333	0.3409	0.5238	0.5	0.141	0.264
7	158536345	A	G	exonic	*ESYT2*	rs2305475	0.5769	0.6667	0.3333	0.3409	0.5238	0.5	0.141	0.264
8	134107312	G	A	exonic	*TG*	rs73354644	0.2821	0.359	0.05556	0.1136	0.119	0	0	0.264
7	53103371	G	T	exonic	*POM121L12*	rs72598684	0.1026	0.05128	0.1667	0.3409	0.2857	0.4103	0.4231	0.237
2	133540605	G	T	exonic	*NCKAP5*	rs13016342	0.1282	0.359	0.08333	0.1818	0.1667	0.05128	0.1282	0.236
22	42276742	G	C	exonic	*SREBF2*	rs2228314	0.5513	0.3333	0.6111	0.6364	0.6905	0.141	0.2692	0.235
8	18257854	T	C	exonic	*NAT2*	rs1801280	0.2692	0.141	0.2778	0.4318	0.5952	0.01282	0.3974	0.232

Supplementary Table 6 Non-synonymous variants within the 1% extreme XP-EHH scores. The allele frequencies are reported for 7 different populations and the mean distance between Nama and the rest. We averaged the difference between the allele frequency of the Nama and the allele frequency of each the populations in the table.

SUPPLEMENTARY TEXT

Regions of interest

Region 1. The first, top-scoring window found among the Nama contains several genes: *METTL13*, *DNM3 and VAMP4*. The top signal is found between *VAMP4* and *METTL13*, with XP-EHH scores uncommonly high along the region (Figure 5). We find histone marks, many differentiated variants (derived allele frequency in Nama 0.5 and 0 in Amhara, Wolayta, Gumuz, CEU, CHB) and several eQTLs such as the highest XP-EHH scoring variant (in rs60956160, decribed in Brain – Cerebellum as an eQTL with a p-value 2e-9 for *DNM3* and in Spleen with a p-value 9.6e-7 for *METTL13*). The intergenic region between *VAMP4* and *METTL13* has clear signs of being a regulatory region where the Nama underwent a selective sweep. There is very limited information about the function of methyltransferase Like 13 (*METTL13*). It has been reported abnormal *METTL13* expression in several human cancers since *METLL13* is the dimethyltransferase of eEF1A which stimulates protein synthesis in cancer cells (1). It is a possible tumour suppressor gene related to bladder cancer since it has been found lower expressions of the gene during late stages of the disease and during tumour progression (2) which in turn is supported by a study where they found that depletion of *METTL13* inhibits proliferation of several cancer lines (1).

Dynamin 3 (DNM3) is a GTPase necessary for endocytosis and neurotransmission localized postsynaptically (3) and it is differentially expressed in the central nervous system (4). It has been associated with obsessive-compulsive disorder without a genome-wide significant p-value (5). Moreover, it has also been shown that *DNM3* is over-expressed in Sézary syndrome patients (6), a rare CD4+ cutaneous T-cell lymphoma.

Vesicle associated membrane protein 4 (VAMP4) is involved in the docking and fusion of synaptic vesicles with the presynaptic membrane. Activity-dependent bulk endocytosis is the prevalent form of synaptic vesicle endocytosis during intense neuronal activity and VAMP4 has been identified as an essential promoter of it (7). It also has been associated in GWAS studies with and hippocampal atrophy (8), but none of the variants described in this

study reached genome-wide significance. Additionally, it has been shown that VAMP4 plays an essential role mediating vesicular eukaryotic interactions at the chlamydial inclusion, an organelle formed by the obligate intracellular pathogen to survive inside the eukaryotic cell (9,10).
All in all, we find that this region is highly differentiated in the Nama compared to other worldwide populations and is the top scoring region that we found in our analysis. Even if the function of the genes inside this block is not completely understood, we have found *VAMP4* and *DNM3* that point to a recently brain specific adaptation that is largely supported by our enrichment analysis presented earlier where we have found *DNM3* in several of the reported neuronal related categories.

Region 2. We observe a very broad region with high XP-EHH and iHS selection scores in chromosome 2, encompassing the area around the lactase (*LCT*) gene and a larger region at its 5' (Figure 6A and B). It is known that 50% of Nama population is lactase persistent (LP) (11) so it is an interesting region to consider. Many studies have focused on the study of LP, especially in Europe and in East African populations from Tanzania and Kenya (12), reporting several candidate variants conferring lactase persistence. Given that the European haplotypes were masked in our dataset, this signal cannot be an artefact generated by the already described strong selection in Europeans around the same area. In fact, we find that the European associated LP derived allele 13910C>T (rs4988235) is found in 5 out of 78 haplotypes (frequency of 0.06) in the raw dataset, and after masking for European ancestry we found a frequency of 0. The main East African mutation 14010G>C (rs145946881), first described in populations from Tanzania and Kenya (12) and found at 55% in Massai population (13), is at an allele frequency of 33% in the Nama in our data, suggesting an event of demic diffusion of pastoralism from eastern Africa (14,15) and having been diffused in many Khoesan populations and been under selection after admixture (11). There are other East African variants that have been reported: 13907*G (rs41525747) described as "marginally" associated with LP, 13915*G (rs41380347; (12) and 14009T>G (rs869051967; (16). None of these variants seem to play a role in LP in the Nama since all of them are in an ancestral state. All the five variants described belong to the regulatory region of *LCT*, located in intron 13 of the neighbour *MCM6* gene. In our

results, there are not signals of selection coming from this particular intron as seen in Figure 6B, but strong signals are found inside *LCT* and in its 5' (as in supplemental material of (14).

Our results indicate a LP in the Nama driven by the Eastern African mutation 14010G>C through an event of positive selection after admixture, but we do not observe a signal of selection in the regulatory region in intron 13 of *MCM6*. Our results, with the highest peaks within *LCT* suggest the existence of adaptive mutations that could be playing an important role for LP among the Nama. We have found in this extensive signal of positive selection an active enhancer (Z-Score of 3.59 H3k27ac in neuron cells in the ENCODE database) in an intron of *LCT* gene (red box in Figure 6B). TRIM28 is a transcription factor that binds in this region. There are two variants that are highly differentiated that could be driving the signal, rs79633114 (frequency in Nama 0.46, Zulu 0.32, Amhara 0.48) and rs12373779 (frequency in Nama 0.59, Zulu 0.32, Amhara 0.48). Both variants show top XP-EHH scores of 5.6 and 6.

Moreover, we have used *haplostrips* (17) to cluster by distance the haplotypes of the 100 kb region around the *LCT* gene (hg19 coordinates 2:136544418-136643577) using a non-admixed Nama haplotype as reference and a set of worldwide populations. In Figure 6C we observe that 69 haplotypes have less than 10 differences with the reference Nama. As expected, 35 are from Nama individuals but 16 from Amhara individuals, 7 from Gumuz, 5 from CEU, 4 from Zulu and 2 from YRI. One should expect a similar number of differences over different populations but in our case, we observe significant differences between populations (Kruskal-Wallis Test chi-squared = 20.438, df = 4, p=0.00041). In particular, that the haplotypes of the Nama individuals share with Amhara a significant lower number of differences in comparison to the Zulu (9.6% of differences in the Amhara and 16,1% in the Zulu over 543 variants, Wilcoxon test p-adjusted = 0.0013) (Figure 6E). Same results are found for the YRI, and CEU comparisons with Amhara (15% and 16% Wilcoxon test p-adjusted = 0.044 and 0.00056). This could be an indicator of an admixture event with an Amhara-like population from Ethiopia, as already mentioned in (6). But the fact that neither the Amhara nor the Gumuz from Ethiopia carry the Eastern African allele of LP (in fact, the allele is not found in any individual) suggests that this allele arose after the initial migration of Afroasiatic populations from Ethiopia into Kenya and Tanzania (16).

Region 3. Another interesting signal is found in a region containing genes *SLC24A5*, *SLC12A1* (Figure 7A). *SLC24A5* is a well-known gene that has been strongly associated with skin pigmentation especially in European populations. Although we see a clear selection signal on the 5' end of *SLC24A5* (with several variants that have been associated to skin pigmentation in Africans (18)), we do not find the European ancestry light skin allele (rs1426654) among the variants under selection with XP-EHH or iHS analyses. Nonetheless, the frequency of the derived allele is quite high in the Nama, 0.46. It is a quite unexpected high allele frequency given the amount of European ancestry in the Nama genome, suggesting that an event of selection in the region happened after admixture as already described in (19).
We also find a clear selection signal in *SLC12A1* (Figure 7A). This gene encodes for a Na-K-Cl cotransporter (NKCC2) expressed in kidney where it reabsorbs sodium, potassium and chloride from urine into the blood (20). Loss-of-function mutations in this gene cause Bartter syndrome type 1, a disease that causes plasma volume reduction, polyuria, hypotension and metabolic alkalosis among others (21). Interestingly, another study found that rare heterozygous mutations in this gene leads to a blood pressure reduction and protects from hypertension, highlighting that a probable combined effect of rare independent mutations could explain blood pressure variation in populations (22).

Region four. Another member of the fibrillin gene family, *FBN3,* is among our candidates of adaptive selection (Figure 7B). *FBN3* is important in maintaining the structure and integrity of the extracellular matrix in connective tissues and contribution to the regulation of TGFß family of growth factors. It is expressed in early development and it is not found among rodents but among cow, sheep, dog, swine, chick, zebrafish and primates (23,24). Genetic variation in the gene has been linked to polycystic ovary syndrome (25) and in a GWAS analysis it has been related to attention deficit hyperactivity disorder and facial morphology (26,27). We find a nonsynonymous substitution (rs7246376) at a high frequency in Nama (0.7), Zulu (0.47), CEU (0.13); no strong consequences are expected (Phred score 6.13). We also find rs57295135, a lung and thyroid eQTL (p-values 6.1e-90 and 1e-120 respectively) that decreases *FBN3* expression with a very high XP-EHH value of

4.35. Interestingly, this variant is at higher frequency in African populations (Nama 0.74, Zulu 0.80, Gumuz 0.90, CEU 0.55). The highest scoring variants of this particular signal (chr19:8207452-8208396) intersect with two cis-regulatory elements (EH37E0480711 and EH37E0480712) that have both high Z-scores of H3K4me3 in myoepithelial of mammary gland female adult (2.61) and H3K27ac in hepatocyte (3.22) indicating active promoters and enhancer marks.

Region five. In next interesting region we find the thyroglobulin (TG) gene (Figure 7C). It is a substrate for the synthesis of thyroid hormones T_3 and T_4, the latter being the main thyroid hormone. The synthesis of T_4 is a two-step process: first, there is an iodination of specific tyrosine residues from TG and secondly a coupling of two doubly iodinated tyrosines to the same TG glycoprotein. Therefore, iodide availability is crucial for this synthesis process. Thyroid hormones are needed in every step of human life; during foetal development it is indispensable for neuronal differentiation and from childhood to adulthood it is required for the regulation of metabolism. This is why a reduced iodine intake can cause irreversible neuronal abnormalities, hypothyroidism and goitre. The TG protein contains 66 tyrosine residues, which 10 to 15 are iodinated depending on iodine intake. A TG protein will form around 3 or 4 thyroid hormones (28). The prevalence of iodine deficiency in Africa was quite high till around 1990. Nowadays, iodine deficiency is no longer a health problem since the introduction of iodized salt. Conversely, there has been an increase of 60% of Grave's disease in South Africa (29) during the last 11 years. The most affected individuals are urban dwellers that recently migrated from iodine deficient areas.
There is an interesting non-synonymous variant with a high Phred-scaled CADD score (11.73) among our selected variants, rs73354644 (A2422T), which has been associated with thyroid dyshormonogenesis, but without clear evidence. We find the highest derived frequency in Nama (0.36), then Zulu (0.28), Gumuz (0.06), CEU and CHB with frequencies of 0. The variant is also found among the top 1% candidates for iHS, supporting the evidence of selection. The variant is located at the end of the protein in the cholinesterase-like domain (ChEL). ChEL domain is crucial since it stabilizes TG and is a molecular carrier for the transport along the secretory pathway of TG. Missense mutations have been found in

the ChEL domain, specifically p.A2215D was found to decrease dramatically TG protein in comparison to TG mRNA and the protein was cornered intracellularly leading to a congenital hypothyroidism (30).

Region 6. There is a high signal in chromosome 2 where two genes are found. *NCKAP5* and close to *GPR39* (Figure 7D). Not much information is available about *NCKAP5* gene function, not even in other organisms. It has a paralogue, *NCKAP5L* that regulates microtubule organization and stabilization (31). *NCKAP5* has been associated in several GWAS with height, body mass index, palmitoleic acid levels, hypersomnia and bipolar disorder (32–36). Interestingly, we find a non-synonymous mutation (rs12611515) at a frequency of 0.80 in the Nama (Table 6) with a very high Phred score of 21; Unfortunately, nothing is known on its possible phenotype implications.
GPR39 is a zinc-dependent G-protein-coupled receptor member of the ghrelin receptor family. Ion zinc is an agonist of GPR39 and may be involved in regulation of body weight, gastrointestinal mobility, hormone secretion and cell death. This gene is mostly active in intestines, prostate and salivary glands. Some studies associate *GPR39* in the pathogenesis of human gastric adenocarcinomas (37). Other studies relate down-regulation of *GPR39* to zinc-deficiency that leads to depressive-like behaviours (38). Since zinc deficiencies can cause a broad number of symptoms and neurological deficiencies this particular region could provide an interesting adaptation in the Nama.

Region 7 We find in the following region two peaks of XP-EHH high scores corresponding clearly to *DDP4* and *SLC4A10* genes (Figure 7E). *DPP4* (also called *CD26*) is expressed on the surface of most cell types and plays an important role in glucose metabolism by regulating the degradation of incretins like Glucagon-like peptide-1 (GLP-1). GLP-1 and glucose-dependent insulinotropic polypeptide (GIP) are hormones secreted by the enteroendocrine cells of the gut in response to the ingestion of nutrients. These incretin hormones, so called because they increase insulin secretion, are key modulators of pancreatic islet hormone secretion and, thus, glucose homeostasis. The glucoregulatory effects of incretins are the basis for new therapies currently being developed for the treatment of type 2 diabetes mellitus (T2DM)

(39). Drugs that inhibit dipeptidyl peptidase-4 (DPP-4), a ubiquitous enzyme that rapidly inactivates both GLP-1 and GIP, increase active levels of these hormones and, in doing so, improve islet function and glycemic control in T2DM (40). This gene is also involved in the negative regulation of lymphocyte trafficking, and its inhibition enhances T cell migration and tumour immunity (41).
SLC4A10 or *NBCn2* gene belongs to a small family of sodium-coupled bicarbonate transporters (NCBTs) that regulate the intracellular pH of neurons, the secretion of bicarbonate ions across the choroid plexus, and the pH of the brain extracellular fluid (42). Among the potential functions of the gene, it has been shown in mice that the knockout (KO) of NBCn2 leads to a slow bicarbonate dependent pH recovery from an intracellular acid load in hippocampal CA3 region cells of mouse. Since a faster pH recovery after neuronal firing leads to a faster recovery of neuronal excitability, SLC4A10 would be crucial to maintain a faster neuronal excitability. Secondly, it has been described that SLC4A10 could be playing a role in cerebrospinal fluid (CSF) secretion because KO mouse exhibit CSF secretion flaws (43). In agreement with the mentioned functions, a case of complex partial epilepsy has been associated with a chromosomal translocation involving SLC4A10 (44) and other cases of chromosomal deletions involving more genes than *SLC4A10* have been also described (45). Other disruptions of *SLC4A10* have been found in autistic patients (46). Moreover, two genome-wide significant variants from *SLC4A10* have been found in GWAS of educational attainment and schizophrenia (47,48). *SLC4A10* is found among many neuronal related categories in our enrichment analysis with WebGestalt. Finally, there are 2 interesting variants in *SLC4A10* in our data (rs113208259, rs113278723) with high Phred scores 10 and 19.6 where Nama has derived allele frequencies of 0.44, Zulu 0.18, Gumuz, Amhara and Wolayta ~0.1 and out of Africans CHB and CEU have a frequency of 0. No phenotypic information exists for any of them.

Region eight. A specific region in chromosome 4 (Figure 7F) shows high scoring variants where we find *NPFFR2* (or *GPR74*) and *ADAMTS3*. *NPFFR2* is a member of a subfamily of G-protein-coupled neuropeptide receptors. The receptor is activated by neuropeptides A-18 (NPAF) and F-8-amide (NPFF). The main studied function of NPFF is the ability to regulate opioid systems

(49) but it is also known to regulate other physiological processes such as the cardiovascular system, response to stress or reward and food intake (50–54). *NPFFR2* has been described as a novel candidate gene for regulation of body mass index by affecting adipocyte lipolysis, in fact it has been found in a Swedish cohort a *NPFFR2* haplotype that protects against obesity (55). Thus this gene could be related to metabolism via a brain physiological process.
The other gene in this location, *ADAMTS3* belongs to the ADAMTS family (disintegrins and metalloproteinases with thrombospondin motifs), is a member of the procollagen aminopropeptidase subfamily and plays an important role in the collagen synthesis. Collagen is the most abundant protein of the connective tissue in humans that supports tissues, confering tensile strength and elasticity. The collagen precursor, procollagen, is a homo or hetero-trimer flanked by amino and carboxy propeptides that have to be removed to allow the assembly of trimers that will conform collagen fibrils. The amino-propeptide is processed by the proteinases ADAMTS2, 3 and 14 (56). A study reported ADAMTS-3 and 16 high levels in women with recurrent miscarriage in comparison with healthy controls. Such result might be explained by a deficient remodelling of the extracellular matrix of the endometrium and result in an impaired implantation (57). In a GWAS study, a variant (rs9993613) of this gene has been associated with height with a p-value 5e-24 (58).

Region nine. Moreover, we find another region related to body weight (Figure 7G). Melanocortin receptor accessory protein 2 (*MRAP2*) modulates signalling of melanocortin receptors *in vitro* such as melanocortin 4 receptor (MC4R) that regulates food intake and energy expenditure. The expression of *MRAP2* is located especially in brain and plays an important role in the regulation of energy homeostasis (59). A study revealed that mice with deletion of *Mrap2* develop severe obesity at a young age probably because of the subsequently alteration of *Mc4r* and perhaps other MCRs. They also found in humans with severe early onset obesity four rare potentially pathogenic variants in *MRAP2*, suggesting a potential role of *MRAP2* in body weight regulation in humans as well (60). Another study described an association of two nonsynonymous *MRAP2* mutations to obesity (61), supporting the hypothesis of *MRAP2* implication in obese patients.

A second peak in the region corresponds to the *KIAA1009* (*CEP162*) gene, which is required to promote assembly of the transition zone in primary cilia. Acts by specifically recognizing and binding the axonemal microtubule (62).
We have found a region of high scores that contains an experimentally tested enhancer (hs2063 located in chr6:84761493-84763345) that is expressed in hindbrain and the somites of embryos. Moreover, the ENCODE database also reports high Z-scores for active promoter marks (H3K4me3) in neurons (2.95). A specific variant is found in this enhancer region rs73479099 at an allele frequency of 0.35, which is not found in any other population (except Zulu at an allele frequency of 0.14).

Region 10. Another example of region under putatively positive selection related to cardiovascular metabolism is matrix metalloproteinase 2 (*MMP2*) or gelatinase A, which is an enzyme associated with tissue remodelling and repair, cleaving extracellular matrix components such as collagen type IV (Figure 7H). Fibroblasts from the dermis and leukocytes are the major sources of MMP2 although we find MMP2 in other tissues such as vascular tissues and cardiac myocytes (63). MMP2 is involved in many physiological and pathological processes, such as angiogenesis, tissue repair, inflammation, in tumour invasion and metastasis (64). MMP2 is one of the most studied metalloproteinases in cardiovascular research since a study showed that MMP2 was activated after a myocardial ischemia-reperfusion injury (65). The active form of MMP2 targeted intracellular proteins within de cardiac cells inducing myocardium injuries and contractile dysfunction. In fact a study showed that patients undergoing I/R injury during a coronary bypass surgery had a fast increase in myocardial MMP2 and MMP9 activity (66). There is an interesting variant in a splice site, rs243834, which the derived allele G is associated with osteolysis-nodulosis-arthropathy spectrum disorders. Nama has by far the lowest derived allele frequency (0.09) compared to other populations (Zulu: 0.32, Gumuz: 0.58, Amhara and Wolayta: 0.5, CHS: 0.65 and CEU: 0.51). This variant is in the top 0.5% scoring in the Nama and has high Z-scores (4.26) of active enhancer marks H3K27ac and DNase, indicating a clear regulatory activity of this variant.

Region 11. This region in chromosome 16 contains the gene *CETP* encoding a cholesteryl ester transfer protein (Figure 7I). It is an important protein for high density lipoporteins (HDL) metabolism since it enables the transfer of cholesteryl ester from HDL toward triglyceride rich lipoproteins and low density lipoproteins (LDL) and contributing to lower HDL cholesterol (67). It is known from epidemiological studies that LDL cholesterol is a risk factor for coronary heart disease and that HDL cholesterol is cardioprotective (68). Interestingly, CETP inhibitors reduce LDL cholesterol and increase HDL cholesterol. In fact, CETP deficiency is associated with elevated plasma levels of HDL cholesterol and low levels of LDL cholesterol causing a high-density lipoprotein deficiency. Many pharmacological studies have focussed on finding CETP inhibitors for over a decade to treat dyslipidemic coronary heart disease patients but none of them reached enough efficacy. Almost all loss-of-function mutations of *CETP* are found in East Asia but their phenotypic consequences are not clear even if it has been found that HDL is higher in countries such as Japan ad their prevalence of coronary heart disease low (69). Moreover, this region is found in our enrichment results in the trait category of Lipoproteins, LDL.

Region 12. Another region that we have found related to cardiovascular function contains *ZFHX3,* a transcription factor that regulates myogenic and neuronal differentiation (Figure 7J). *ZFHX3* knockdown in atrial myocytes causes dysregulation of calcium homeostasis and increased atrial arrhythmogenesis, which might contribute to the occurrence of atrial fibrillation (70). It has been associated with atrial fibrillation in Europeans (71), and in Chinese Han (72). A study of Icelandic population revealed that the variant rs7193343-T (the ancestral variant) is significantly associated with atrial fibrillation, ischemic stroke and cardio embolic stroke (73). In our data, Nama has the lowest frequency of T-allele (0.05) compared to other populations such as Zulu (0.12), Gumuz (0.14), Amhara (0.07), Wolayta (0.12), CHB (0.64) and CEU (0.18). Another interesting intronic variant, rs73594727, that has a very high Phred score of 20.1 is found in Nama at the highest derived allele frequency (0.5) followed by Zulu (0.37), Gumuz (0.08), Amhara (0.02) and Wolayta (0.17) and CHB and CEU (0). This particular variant is in a very high peak of scores (chr16:73001773-73030279), where we find several cis- regulatory elements (35

elements) indicating a high level of regulatory activity of this region. In particular 6 out of the 35 elements have very high Z-scores (>3.39) in cardiac muscle cells of active enhancer marks H3K27ac. These findings, together with the fact that the region is under strong positive selection and the enrichment in cardiac function categories from enrichment results, indicates a Nama specific adaptation related to heart function.
It is interesting to note that a study found that African Americans had a lower incidence of atrial fibrillation and that an increasing European ancestry was associated with an increased risk in African Americans (74).

Region 13. This region (Figure 7K) contains *TRPM3*, a member of the family of transient receptor potential channels. It is a non-selective calcium permeable channel that is expressed in sensory neurons, pancreatic ß-cells, kidney and vascular smooth muscle (75–77). In sensory neurons, it has been demonstrated that TRPM3 acts as a sensor of nociceptive heat in sensory neurons that is activated from a temperature of 40º C (78) and in mice the Trpm3 knock-out show deficiencies in avoidance of noxious heat (76). In vascular smooth muscle cells from patients, a study showed that TRPM3 is involved in cell contraction, cytokine secretion and that its activity is repressed by cholesterol, a main driver of atherosclerosis (77). This gene mediates calcium entry potentiated by calcium store depletion. When trying to unravel molecular and cellular differences in brain organization between human and nonhuman primates, TRPM3 was found to be human-specific down-regulated in the amygdala, which is located in the brain an plays an essential role in memory, decision-making and emotional reactions (79). Mutations in the gene can also cause mental disabilities (80).
There are several cis-regulatory elements (EH37E1297900 to EH37E1297905) with high XP-EHH values in the region chr9:73478924-73483708 (with very high XP-EHH values, see Figure 7K) that have high Z-scores of active promoter marks (H3K4me3) specially in retinal pigment epithelial cell (>5) and endothelial cell of umbilical vein newborn (>4) and in kidney foetal cells.

Bibliography

1. Liu S, Hausmann S, Carlson SM, Rechem V, Mazur K, Gozani O. METTL13 Methylation of eEF1A Increases Translational Output to Promote Tumorigenesis. Cell. 2019;176:491–504.
2. Zhang Z, Zhang G, Kong C, Zhan B, Dong X, Man X. METTL13 is downregulated in bladder carcinoma and suppresses cell proliferation, migration and invasion. Sci Rep. 2016 Jan 14;6:19261.
3. Lu J, Helton TD, Blanpied TA, Rácz B, Newpher TM, Weinberg RJ, et al. Postsynaptic Positioning of Endocytic Zones and AMPA Receptor Cycling by Physical Coupling of Dynamin-3 to Homer. Neuron. 2007 Sep 20;55(6):874–89.
4. Romeu A, Arola L. Classical dynamin DNM1 and DNM3 genes attain maximum expression in the normal human central nervous system. BMC Res Notes. 2014 Mar 28;7:188.
5. Costas J, Carrera N, Alonso P, Gurriarán X, Segalàs C, Real E, et al. Exon-focused genome-wide association study of obsessive-compulsive disorder and shared polygenic risk with schizophrenia. Transl Psychiatry. 2016 Mar 29;6(3):e768–e768.
6. Booken N, Gratchev A, Utikal J, Weiß C, Yu X, Qadoumi M, et al. Sézary syndrome is a unique cutaneous T-cell lymphoma as identified by an expanded gene signature including diagnostic marker molecules CDO1 and DNM3. Leukemia. 2008 Feb 22;22(2):393–9.
7. Nicholson-Fish JC, Kokotos AC, Gillingwater TH, Smillie KJ, Cousin MA. VAMP4 Is an Essential Cargo Molecule for Activity-Dependent Bulk Endocytosis. Neuron. 2015 Dec 2;88(5):973–84.
8. Mather KA, Armstrong NJ, Wen W, Kwok JB, Assareh AA, Thalamuthu A, et al. Investigating the Genetics of Hippocampal Volume in Older Adults without Dementia. Arking DE, editor. PLoS One. 2015 Jan 27;10(1):e0116920.
9. Fields KA, Hackstadt T. The Chlamydial Inclusion: Escape from the Endocytic Pathway. Annu Rev Cell Dev Biol. 2002 Nov;18(1):221–45.
10. Kabeiseman EJ, Cichos K, Hackstadt T, Lucas A, Moore ER. Vesicle-associated membrane protein 4 and syntaxin 6 interactions at the chlamydial inclusion. Infect Immun. 2013

Sep;81(9):3326–37.
11. Macholdt E, Lede V, Barbieri C, Mpoloka SW, Chen H, Slatkin M, et al. Tracing pastoralist migrations to southern Africa with lactase persistence alleles. Curr Biol. 2014 Apr 14;24(8):875–9.
12. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007 Jan 10;39(1):31–40.
13. Wagh K, Bhatia A, Alexe G, Reddy A, Ravikumar V, Seiler M, et al. Lactase Persistence and Lipid Pathway Selection in the Maasai. Johnson N, editor. PLoS One. 2012 Sep 28;7(9):e44751.
14. Breton G, Schlebusch CM, Lombard M, Sjödin P, Soodyall H, Jakobsson M. Lactase Persistence Alleles Reveal Partial East African Ancestry of Southern African Khoe Pastoralists. Curr Biol. 2014 Apr 14;24(8):852–8.
15. Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, van Helden PD, et al. Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. Genetics. 2016 Sep 1;204(1):303–14.
16. Jones BL, Raga TO, Liebert A, Zmarz P, Bekele E, Danielsen ET, et al. Diversity of Lactase Persistence Alleles in Ethiopia: Signature of a Soft Selective Sweep. Am J Hum Genet. 2013 Sep 5;93(3):538–44.
17. Marnetto D, Huerta-Sánchez E. *Haplostrips* : revealing population structure through haplotype visualization. Price S, editor. Methods Ecol Evol. 2017 Oct 1;8(10):1389–92.
18. Martin AR, Lin M, Granka JM, Myrick JW, Liu X, Sockell A, et al. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. Cell. 2017 Nov 30;171(6):1340-1353.e14.
19. Lin M, Siford RL, Martin AR, Nakagome S, Möller M, Hoal EG, et al. Rapid evolution of a skin-lightening allele in southern African KhoeSan. Proc Natl Acad Sci U S A. 2018;115(52):13324–9.
20. Carmosino M, Rizzo F, Procino G, Zolla L, Timperio AM, Basco D, et al. Identification of moesin as NKCC2-interacting protein and analysis of its functional role in the NKCC2 apical trafficking. Biol Cell. 2012 Nov;104(11):658–76.

21. Markadieu N, Delpire E. Physiology and pathophysiology of SLC12A1/2 transporters. Pflügers Arch - Eur J Physiol. 2014 Jan 6;466(1):91–105.
22. Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet. 2008 May;40(5):592–9.
23. Sabatier L, Miosge N, Hubmacher D, Lin G, Davis EC, Reinhardt DP. Fibrillin-3 expression in human development. Matrix Biol. 2011 Jan;30(1):43–52.
24. Corson GM, Charbonneau NL, Keene DR, Sakai LY. Differential expression of fibrillin-3 adds to microfibril variety in human and avian, but not rodent, connective tissues. Genomics. 2004 Mar;83(3):461–72.
25. Davis MR, Summers KM. Structure and function of the mammalian fibrillin gene family: Implications for human connective tissue diseases. Mol Genet Metab. 2012 Dec;107(4):635–47.
26. Hawi Z, Yates H, Pinar A, Arnatkeviciute A, Johnson B, Tong J, et al. A case–control genome-wide association study of ADHD discovers a novel association with the tenascin R (TNR) gene. Transl Psychiatry. 2018 Dec 18;8(1):284.
27. Lee MK, Shaffer JR, Leslie EJ, Orlova E, Carlson JC, Feingold E, et al. Genome-wide association study of facial morphology reveals novel associations with FREM1 and PARK2. Li Y, editor. PLoS One. 2017 Apr 25;12(4):e0176566.
28. Di Jeso B, Arvan P. Thyroglobulin From Molecular and Cellular Biology to Clinical Endocrinology. Endocr Rev. 2016 Feb 1;37(1):2–36.
29. Okosieme OE. Impact of iodination on thyroid pathology in Africa. J R Soc Med. 2006 Aug;99(8):396–401.
30. Pardo V, Vono-Toniolo J, Rubio IGS, Knobel M, Possato RF, Targovnik HM, et al. The p.A2215D Thyroglobulin Gene Mutation Leads to Deficient Synthesis and Secretion of the Mutated Protein and Congenital Hypothyroidism with Wide Phenotype Variation. J Clin Endocrinol Metab. 2009 Aug;94(8):2938–44.
31. Mori Y, Inoue Y, Tanaka S, Doda S, Yamanaka S, Fukuchi H, et al. Cep169, a Novel Microtubule Plus-End-Tracking Centrosomal Protein, Binds to CDK5RAP2 and Regulates

Microtubule Stability. PLoS One. 2015;10(10):e0140968.
32. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010 Oct 29;467(7317):832–8.
33. Kichaev G, Bhatia G, Loh P-R, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am J Hum Genet. 2019 Jan 3;104(1):65–75.
34. Wu JHY, Lemaitre RN, Manichaikul A, Guan W, Tanaka T, Foy M, et al. Genome-Wide Association Study Identifies Novel Loci Associated With Concentrations of Four Plasma Phospholipid Fatty Acids in the De Novo Lipogenesis Pathway. Circ Cardiovasc Genet. 2013 Apr;6(2):171–83.
35. Khor S-S, Miyagawa T, Toyoda H, Yamasaki M, Kawamura Y, Tanii H, et al. Genome-wide association study of *HLA-DQB1*06:02* negative essential hypersomnia. PeerJ. 2013 Apr 16;1:e66.
36. Wang K-S, Liu X-F, Aragam N. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. Schizophr Res. 2010 Dec;124(1–3):192–9.
37. Alén BO, Leal-López S, Alén MO, Viaño P, García-Castro V, Mosteiro CS, et al. The role of the obestatin/GPR39 system in human gastric adenocarcinomas. Oncotarget. 2016 Feb 2;7(5):5957–71.
38. Młyniec K, Singewald N, Holst B, Nowak G. GPR39 Zn2+-sensing receptor: A new target in antidepressant development? J Affect Disord. 2015 Mar 15;174:89–100.
39. Rufinatscha K, Radlinger B, Dobner J, Folie S, Bon C, Profanter E, et al. Dipeptidyl peptidase-4 impairs insulin signaling and promotes lipid accumulation in hepatocytes. Biochem Biophys Res Commun. 2017 Apr 1;485(2):366–71.
40. Pratley RE, Salsali A. Inhibition of DPP-4: a new therapeutic approach for the treatment of type 2 diabetes. Curr Med Res Opin. 2007 Apr;23(4):919–31.
41. Hollande C, Boussier J, Ziai J, Nozawa T, Bondet V, Phung W, et al. Inhibition of the dipeptidyl peptidase DPP4 (CD26) reveals IL-33-dependent eosinophil-mediated control of tumor growth. Nat Immunol. 2019 Mar 18;20(3):257–64.
42. Parker MD, Boron WF. The Divergence, Actions, Roles, and

Relatives of Sodium-Coupled Bicarbonate Transporters. Physiol Rev. 2013 Apr;93(2):803–959.
43. Jacobs S, Ruusuvuori E, Sipila ST, Haapanen A, Damkier HH, Kurth I, et al. Mice with targeted Slc4a10 gene disruption have small brain ventricles and show reduced neuronal excitability. Proc Natl Acad Sci. 2008 Jan 8;105(1):311–6.
44. Gurnett CA, Veile R, Zempel J, Blackburn L, Lovett M, Bowcock A. Disruption of Sodium Bicarbonate Transporter SLC4A10 in a Patient With Complex Partial Epilepsy and Mental Retardation. Arch Neurol. 2008 Apr 1;65(4):550.
45. Krepischi ACV, Knijnenburg J, Bertola DR, Kim CA, Pearson PL, Bijlsma E, et al. Two distinct regions in 2q24.2-q24.3 associated with idiopathic epilepsy. Epilepsia. 2010 Dec;51(12):2457–60.
46. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. Science. 2007 Apr 20;316(5823):445–9.
47. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018 Aug 23;50(8):1112–21.
48. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014 Jul 22;511(7510):421–7.
49. Bray L, Froment C, Pardo P, Candotto C, Burlet-Schiltz O, Zajac J-M, et al. Identification and Functional Characterization of the Phosphorylation Sites of the Neuropeptide FF $_2$ Receptor. J Biol Chem. 2014 Dec 5;289(49):33754–66.
50. Panula P, Aarnisalo AA, Wasowicz K. Neuropeptide FF, a mammalian neuropeptide with multiple functions. Prog Neurobiol. 1996 Mar 1;48(4–5):461–87.
51. Cador M, Marco N, Stinus L, Simonnet G. Interaction between neuropeptide FF and opioids in the ventral tegmental area in the behavioral response to novelty. Neuroscience. 2002 Mar 12;110(2):309–18.
52. Zhang L, Ip CK, Lee I-CJ, Qi Y, Reed F, Karl T, et al. Diet-

induced adaptive thermogenesis requires neuropeptide FF receptor-2 signalling. Nat Commun. 2018 Dec 9;9(1):4722.
53. Huang EYK, Li JY, Wong CH, Tan PPC, Chen JC. Dansyl-PQRamide, a possible neuropeptide FF receptor antagonist, induces conditioned place preference. Peptides. 2002 Mar;23(3):489–96.
54. Nicklous DM, Simansky KJ. Neuropeptide FF exerts pro- and anti-opioid actions in the parabrachial nucleus to modulate food intake. Am J Physiol Integr Comp Physiol. 2003 Nov;285(5):R1046–54.
55. Dahlman I, Dicker A, Jiao H, Kere J, Blomqvist L, van Harmelen V, et al. A common haplotype in the G-protein-coupled receptor gene GPR74 is associated with leanness and increased lipolysis. Am J Hum Genet. 2007 Jun;80(6):1115–24.
56. Bekhouche M, Colige A. The procollagen N-proteinases ADAMTS2, 3 and 14 in pathophysiology. Matrix Biol. 2015 May 1;44–46:46–53.
57. Pekcan MK, Sarıkaya E, Tokmak A, Alışık M, Alkan A, Özakşit G, et al. ADAMTS-3, -13, -16, and -19 levels in patients with habitual abortion. Kaohsiung J Med Sci. 2017 Jan 1;33(1):30–5.
58. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014 Nov 5;46(11):1173–86.
59. Rouault AAJ, Srinivasan DK, Yin TC, Lee AA, Sebag JA. Melanocortin Receptor Accessory Proteins (MRAPs): Functions in the melanocortin system and beyond. Biochim Biophys Acta - Mol Basis Dis. 2017 Oct 1;1863(10):2462–7.
60. Asai M, Ramachandrappa S, Joachim M, Shen Y, Zhang R, Nuthalapati N, et al. Loss of Function of the Melanocortin 2 Receptor Accessory Protein 2 Is Associated with Mammalian Obesity. Science (80-). 2013 Jul 19;341(6143):275–8.
61. Schonnop L, Kleinau G, Herrfurth N, Volckmar A-L, Cetindag C, Müller A, et al. Decreased melanocortin-4 receptor function conferred by an infrequent variant at the human melanocortin receptor accessory protein 2 gene. Obesity. 2016 Sep;24(9):1976–82.
62. Wang W-J, Tay HG, Soni R, Perumal GS, Goll MG, Macaluso FP, et al. CEP162 is an axoneme-recognition

protein promoting ciliary transition zone assembly at the cilia base. Nat Cell Biol. 2013 Jun 5;15(6):591–601.
63. Xie Y, Mustafa A, Yerzhan A, Merzhakupova D, Yerlan P, N Orakov A, et al. Nuclear matrix metalloproteinases: functions resemble the evolution from the intracellular to the extracellular compartment. Cell Death Discov. 2017 Dec 14;3(1):17036.
64. Yu C-F, Chen F-H, Lu M-H, Hong J-H, Chiang C-S. Dual roles of tumour cells-derived matrix metalloproteinase 2 on brain tumour growth and invasion. Br J Cancer. 2017 Dec 24;117(12):1828–36.
65. DeCoux A, Lindsey ML, Villarreal F, Garcia RA, Schulz R. Myocardial matrix metalloproteinase-2: inside out and upside down. J Mol Cell Cardiol. 2014 Dec;77:64–72.
66. Lalu MM, Pasini E, Schulze CJ, Ferrari-Vivaldi M, Ferrari-Vivaldi G, Bachetti T, et al. Ischaemia–reperfusion injury activates matrix metalloproteinases in the human heart. Eur Heart J. 2005 Jan 1;26(1):27–35.
67. Mabuchi H, Nohara A, Inazu A. Cholesteryl ester transfer protein (CETP) deficiency and CETP inhibitors. Mol Cells. 2014 Nov;37(11):777–84.
68. Assmann G, Schulte H, von Eckardstein A, Huang Y. High-density lipoprotein cholesterol as a predictor of coronary heart disease risk. The PROCAM experience and pathophysiological implications for reverse cholesterol transport. Atherosclerosis. 1996 Jul;124:S11–20.
69. Sheridan C. CETP inhibitors boost “good” cholesterol to no avail. Nat Biotech. 2016 Jan;34(1):5–6.
70. Kao Y-H, Hsu J-C, Chen Y-C, Lin Y-K, Lkhagva B, Chen S-A, et al. ZFHX3 knockdown increases arrhythmogenesis and dysregulates calcium homeostasis in HL-1 atrial myocytes. Int J Cardiol. 2016 May 1;210:85–92.
71. Benjamin EJ, Rice KM, Arking DE, Pfeufer A, van Noord C, Smith A V, et al. Variants in ZFHX3 are associated with atrial fibrillation in individuals of European ancestry. Nat Genet. 2009 Aug 13;41(8):879–81.
72. Liu Y, Ni B, Lin Y, Chen X, Fang Z, Zhao L, et al. Genetic Polymorphisms in ZFHX3 Are Associated with Atrial Fibrillation in a Chinese Han Population. Ai X, editor. PLoS One. 2014 Jul 1;9(7):e101318.
73. Gudbjartsson DF, Holm H, Gretarsdottir S, Thorleifsson G,

Walters GB, Thorgeirsson G, et al. A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. Nat Genet. 2009 Aug 13;41(8):876–8.
74. Marcus GM, Smith LM, Ordovas K, Scheinman MM, Kim AM, Badhwar N, et al. Intracardiac and extracardiac markers of inflammation during atrial fibrillation. Hear Rhythm. 2010 Feb;7(2):149–54.
75. Wagner TFJ, Loch S, Lambert S, Straub I, Mannebach S, Mathar I, et al. Transient receptor potential M3 channels are ionotropic steroid receptors in pancreatic β cells. Nat Cell Biol. 2008 Dec 2;10(12):1421–30.
76. Vriens J, Owsianik G, Hofmann T, Philipp SE, Stab J, Chen X, et al. TRPM3 Is a Nociceptor Channel Involved in the Detection of Noxious Heat. Neuron. 2011 May 12;70(3):482–94.
77. Naylor J, Li J, Milligan CJ, Zeng F, Sukumar P, Hou B, et al. Pregnenolone sulphate- and cholesterol-regulated TRPM3 channels coupled to vascular smooth muscle secretion and contraction. Circ Res. 2010 May 14;106(9):1507–15.
78. Vriens J, Voets T. Sensing the heat with TRPM3. Pflügers Arch - Eur J Physiol. 2018 May 5;470(5):799–807.
79. Sousa AMM, Zhu Y, Raghanti MA, Kitchen RR, Onorati M, Tebbenkamp ATN, et al. Molecular and cellular reorganization of neural circuits in the human lineage. Science (80-). 2017 Nov 24;358(6366):1027–32.
80. Dyment DA, Terhal PA, Rustad CF, Tveten K, Griffith C, Jayakar P, et al. De novo substitutions of TRPM3 cause intellectual disability and epilepsy. Eur J Hum Genet. 2019 Oct 5;27(10):1611–8.

ADAPTIVE SELECTION DRIVES FUNCTIONAL CHANGES IN TRPP3 IN ETHIOPIAN POPULATIONS

Sandra Walsh[1], Mercè Izquierdo-Serra[2], Sandra Acosta[1], Maria Lloret[2], Jaume Bertranpetit[1*], José Manuel Fernández-Fernández[2*]

1. Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Dr. Aiguader, 88, 08003 Barcelona, Catalonia, Spain.
2. Laboratory of Molecular Physiology, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona Spain

* co-corresponding authors: jaume.bertranpetit@upf.edu and jmanuel.fernandez@upf.edu

Abstract

The functional interpretation of signals of positive selection from genome-wide scans is a necessary step towards the validation and understanding of the adaptive mechnisms of human adaptation. Nonetheless, very few studies engage in this arduous task due to the technical considerations and the need of an interdisciplinary approach. Previous work show that the *TRPP3* gene has been a target of positive selection in the Gumuz from Ethiopia. Two nonsynoymous subsitutions were found at high frequency and were predicted to be damaging (rs17112895 and rs7909153). In this study, we perform a functional follow-up study of a candidate of adatpation in order to understand the underlying molecular mechanisms that lead to the increase of reproductive fitness in a population. For that, we performed electrohpysiological and phenotype-genotype analysis of TRPP3. The analysis revealed an impaired function of the channel driven by both substitutions with no effect on membrane expression. We next examined the putative consequences on the sour taste recognition threshold of individuals carrying both substitutions but no effect was detected. All together this study provides a first functional insight into a target of positive selection in the Gumuz.

Introduction

In evolutionary biology the detection and interpretation of the footprints of adaptive selection is a very promising and interesting topic, as it may pinpoint where in the genome a selective sweep has happened, driven by an event of positive (also called adaptive) selection. The interesting point is that it is a fully blind process of recognizing where in the genome adaptation has happened. These studies may be undertaken at several levels, one of the most powerful and interesting being the recognition of population specific adaptations that may be interpreted as phenotypic adaptations (Scheinfeldt and Tishkoff 2013; Fan et al. 2016; Booker, Jackson, and Keightley 2017). There are some well-known cases for which the adaptation has been fully interpreted, from the detection of the signal to the specific genetic change and into the adaptive phenotype, like the lactase persistence (Tishkoff et al. 2007; Enattah et al. 2002), the adaptation to high altitudes (Simonson et al. 2010; Beall et al. 2010; Huerta-Sánchez et al. 2013; Yi et al. 2010), taste (Campbell et al. 2014), diet (Fumagalli et al. 2015) and some others. Thanks to the massive production of genomic data, it has been possible to undertake whole genome scans of positive selection for several populations. Nonetheless, for many cases, the detection of a selection signal is far for being interpreted. First because in many cases there are a large proportion of signals lying in genome regions without functional annotation and likely being related to regulatory elements; second because it is difficult to pinpoint the specific causal variants, due to the lack of precision of the selection footprints; and last because of the general difficulty of making causal relationships between genotype and phenotype, a fact in common with most of GWAS studies.

In a study of selection in several populations from Ethiopia (Walsh et al, in press) several interesting signals were uncovered using two different selection tests, SFselect (Ronen et al. 2013), that detects mainly ancient signals as it is based in changes in the site frequency spectrum, and iHS (Sabeti et al. 2002), a test based in the extension on linkage disequilibrium and that detects recent events that are mainly population-specific. One of the top signals of positive selection of the iHS analysis was a region in the long arm of chromosome 10 (Fig. 1) that contains the *TRPP3* (also called *PKD2L1*) gene. This signal is found specifically in one Ethiopian population, the Gumuz, a Nilo-Saharan-speaking population unaffected by the admixture from the Levant that is typically found

in other Ethiopian populations (Pagani et al. 2012). In fact there is a strong overlapping between the signal and the gene and, what is more interesting, there are two exonic non-synonymous variants in the gene that have a very strong iHS values and that present a very strong population differentiation and a high derived allele frequency in the Gumuz, rs17112895 and rs7909153 both at a frequency of 0.70. In non-African populations, the allele frequency of the derived allele is zero.

TRPP3 belongs to the TRPP subfamily of ion channels that are characterized by large extracellular domains (Su et al. 2018). TRPP3 forms a Ca^{2+} permeable channel in a homotetrameric structure or in a heterotetrameric structure together with PKD1L3. TRPP3 has a voltage-sensing domain formed by the S1 to S4 transmembrane alpha helices, the pore domain is formed by the S5 and S6 helices and the polycystin domain (PD). The variants are located in the PD and in the intracellular linker between the S2 and S3 alpha helices (Fig. 2). This channel is activated by an off-acid response (Inada et al. 2008), voltage and osmolality. The PKD2L1/PKD1L3 complex is expressed in a subset of taste receptor cells in specific taste areas (Ishimaru et al. 2006) and has been identified as a candidate for sour taste in mammals (Huang et al. 2006). In humans, two patients with sour taste ageusia have been reported and neither had detectable PKD2L1 transcripts (Huque et al. 2009). Sour taste is one of the five basic tastes. And although other tastes have a clear evolutionary purpose (sweet indicates carbohydrate rich food, salty taste sodium, bitter potentially poisonous and umami protein rich), sour tasting remains unexplored in humans. One of the main hypotheses of the evolutionary sour tasting function is that it could warn against the acidic ingestion of rotten or immature fruit (DeSimone, Heck, and DeSimone 1981). Even if the tasting function of *TRPP3* is the most widely studied, two other functions have been described. The first one, in zebrafish, *Trpp3* is expressed in cerebrospinal fluid-contacting neurons and required for the responses to pressure and maintenance of spine curvature (Sternberg et al. 2018). Zebrafish that do not express *Trpp3* present an exaggerated spine curvature similar to human kyphosis. Lastly, a recent paper showed that TRPP3 knockout mice promotes mitochondrial calcium overload in cardiomyocytes (Lu et al. 2018). Moreover, a high salt induced diet on knockout mice led to cardiac hypertrophy and dysfunction, revealing the inhibitory effect of TRPP3 on cardiac hypertrophy.

Besides the external phenotype, what is more intriguing is to which extent the two non-synonymous substitutions affect the electrophysiological properties of the ion channel.
In this study we depart from the selection footprint in the genome, recognize the gene *TRPP3* as the selection target, with two non-synonymous variants, and use molecular physiology techniques to recognize the alterations produced by the two substitutions. Finally, in trying to link genotype and external phenotype we try to find differences in sour taste recognition.

Materials and Methods

cDNA constructs

cDNA of the human TRPP3 wild-type (WT) channel, cloned into the plasmid pCMV6-AC-GFP was obtained from OriGene, Maryland, United States. TRPP3 mutant channels (R278Q, R378W and R278Q/R378W), were generated by site-directed mutagenesis of the human WT TRPP3 cDNA (GenScript Corporation, Piscatway, NJ, USA). cDNA of mouse PKD1L3, cloned into pDisplay vector was a gift from Dr. Yong Yu (Department of Biological Sciences, St. John's University, 8000 Utopia Parkway, Queens, New York, 11439, USA). All cDNA clones were sequenced previously to the study in order to confirm their integrity.

Heterologous expression

HEK293 cells were transfected using a linear polyethylenimine (PEI) derivative, the polycation ExGen500 (Fermentas Inc., Hanover, MD, USA) as previously reported (eight equivalents PEI/3.3 µg DNA/dish) (Izquierdo-Serra et al. 2018). Co-transfection was performed using the ratio for GFP-fused TRPP3 (WT or either mutant R278Q, R378W, or R278Q/R378W), and PKD1L3 cDNA constructs of 1:1. Transfected cells were incubated for 24 h and seeded on 35 mm diameter dishes Poly-D-Lysine-coated (Sigma). Experiments were performed 48 h post-transfection at room temperature (22-24°C).

Whole-cell patch-clamp recordings and electrophysiological analysis

Patch electrodes had a resistance of 2 MΩ when filled with a pipette solution containing (in mM): 80 CsMES, 20 NaCl, 2 $MgCl_2$, 10 HEPES and 5 BAPTA (pH 7.4, adjusted using TrisBase). Osmolarity was adjusted to 300 mOsm/l adding D-mannitol. Recordings were performed at two different extracellular pH values, 7.4 or 9. In the first case, the external bath solution contained (in mM): 150 NaCl, 2 $CaCl_2$, 10 HEPES (310 mOsm/l and pH 7.4, adjusted using Tris-base). The basic external bath solution contained (in mM) 150 NaCl, 2 $CaCl_2$ and 10 TAPS (310 mOsm/l and pH 9, adjusted using Tris-base). All chemicals were obtained from Sigma-Aldrich.

TRPP3 activity was assessed using the whole-cell configuration of the patch-clamp technique on GFP-positive HEK293 cells 48 h after

co-transfection of PKD1L3 with cDNA encoding wild-type (WT), R278Q, R378W or R278Q/R378W TRPP3 channels. Whole-cell cationic currents were evoked by the application of 400 ms step pulses from -100 to +160 mV in 10 mV increments followed by a 200 ms postpulse to -100 mV. Besides, spontaneous single-channel currents were also recorded in the whole-cell configuration by using a 40 ms gap-free protocol at several negative holding-potentials (-60 mV, -80 mV and -100 mV). All recordings were obtained with a D-6100 Darmstadt amplifier (List Medical, Germany), sampled at 10 kHz and filtered at 1 kHz. The pClamp8 or pClamp10.5 software (Molecular Devices, Sunnyvale, CA, USA) was used for pulse generation, data acquisition and subsequent analysis. By employing Fetchan and pStat softwares, the amplitude of single-channel currents was measured as the peak-to-peak distance in Gaussian fits of amplitude histograms. Channel activity (NP_o, where N is the number of channels and P_o is the open probability of a single channel) at negative voltages was calculated by dividing the mean current amplitude of recordings lasting 40 s by the single-channel amplitude obtained from the same traces.

Immunofluorescence

Transfected cells were incubated with ConcanavalinA-Rhodamine (Molecular Probes Ref C860) for 20 minutes, rinsed with PBS and fixed with PFA 4%. For immunofluorescence staining cells were fixed 24 or 48 hours post-transfection with paraformaldehyde 4% for 30 minutes. Upon thorough rinsing, cells were treated with a blocking solution of 5% bovine serum albumin (BSA) and 3% foetal bovine serum with or without triton 0,1% (for permeabilization). Anti-GFP antibody was incubated overnight and Alexa-488 secondary antibody was used for detection. DAPI was used for nuclear staining.

Confocal imaging was performed to detect membrane expression of the GFP in the membrane (shown as ConA-Rho staining) using the Leica SP5. One plane was acquired every 3 μm and maximal projection was reconstructed.

Statistical analysis

Data are presented as the means ± S.E.M. Statistical tests included one-tail Student's t-test, one-tail Mann-Whitney U-test, Kruskal-Wallis test followed by Dunn post hoc test, or one-way ANOVA followed by Bonferroni's post hoc test, as appropriate. Differences were considered significant if $P < 0.05$.

Sour taste sensitivity test

A modified Harris-Kalmus test (Harris and Kalmus 1949) was designed to determine the recognition threshold of sour taste of individuals.

Citric acid (CA) or E-330 was used as stimuli since it is a harmless food additive which maximum doses are catalogued as "quantum satis" after BOE-A-2002-3366. The citric acid used for this study was bought from Sigma-Aldrich with reference C0759.

The stimuli used were solutions of CA in Millipore-filtered, deionized water. They are prepared less than a week in advance of use, stored under refrigeration in amber glass bottles and warmed to room temperature before use. Water blanks are Millipore-filtered deionized water stored and handled the same way. The CA concentrations ranged from $5x10^{-2}$ mM to 1 mM in a $5x10^{-2}$ mM step.

During each trial, subjects received a 10 mL sample in a medicine cup. After holding the sample in their mouth for at least 5 seconds the subjects attempted to identify the quality of taste (sweet, sour, bitter, salty, water). Subjects rinsed at least twice with deionized water. The threshold run begins with the lowest concentration. Subjects sampled each concentration once in ascending order until they identify the target quality "sour". This is the suprathreshold quality.

We next ensure that the subjects experienced a reliable taste sensation with a sorting task. Subjects receive 3 samples at the concentration at which they identified the target quality with 3 more blank samples. All 6 samples are presented at the same time in a random position. The subjects are asked to sort 3 cups into "tastes" and 3 cups into "waters". If the subject can correctly sort the 6 samples in 2 consecutive trials, the threshold run ends. If a subject fails to sort correctly the sorting task is repeated at the next higher concentration. The concentration that first allows successive correct sorts is the taste quality recognition threshold.

Sample collection, DNA extraction and genotyping

After having written informed consent from the participants, we obtained a saliva sample from which DNA was extracted using NaCl 6M solution and subsequently precipitating the DNA on isopropanol. Genotyping was performed through Sanger sequencing of PCR fragments. The primers used for the PCR are the following:

Target	Forward primer	Reverse primer
rs17112895 and rs7909153	5'-AAGATGACCACCAGGTCCAG-3'	5'-TTGATACGTGGAGCAGCAAG-3'
rs7909153	5'-GTCAGAAGGAGGGCTTAGGG-3'	5'-CTGATCCGCTATGTCAGCAA-3'
rs17112895	5'-GCACCTCCTGTAGCTGGAAA-3'	5'-CATTGCCTTGGGAGATGAAC-3'

The primers used for sequencing are:

	Sequencing Primers
Seq1	Reverse 5'-ACTGAAACTCAAATGTAAATCTG-3'
Seq2	Forward 5'-CATGCACAAGAGACGGGAAC-3'
Seq3	Reverse 5'-CATTGCCTTGGGAGATGAAC-3'

Ethical approval

The study of sour test sensitivity, with reference number 2018/8294/I, was approved by the ethical committee CEIm-Parc de Salut MAR and written informed consent to participate in the study and for the saliva donation was obtained from all participants.

Results

Impaired function of the TRPP3 double mutant

In order to characterize the functional consequences of the two TRPP3 non-synonymous genetic variants (R238Q and R378W) under strong positive selection in the Gumuz population from Ethiopia, we compare channel activity of the wild-type (WT) TRPP3 channel with that of channels carrying either the double substitution (R238Q/R378W) or the single substitutions (R278Q or R378W). For that purpose, we performed whole-cell patch-clamp recordings on GFP positive HEK293 cells co-transfected with PKD1L3 and GFP-fused TRPP3 channels (TRPP3-GFP). We measured robust channel activity in response to voltage changes in cells expressing WT TRPP3. As previously reported (Shimizu et al. 2009), they exhibited outwardly rectifying currents with large tail currents after repolarization to -100 mV at extracellular pH 7.4 (Fig. 3A). Besides, recordings at extracellular pH 9.0 revealed largest TRPP3 steady-state and tail currents (Fig. 3B), as reported before for HEK293 cells expressing only TRPP3 channels in the absence of PKD1L3 (Shimizu et al. 2011). The average current-voltage relationships for WT steady-state currents and tail currents at each test pulse and pH condition are shown in Fig. 3E-H. Furthermore, we also found that at extracellular pH 9 the deactivation time course of WT TRPP3 tail currents was accelerated when compared to the extracellular pH 7.4 condition, another known effect of alkalinisation on TRPP3 tail currents (Shimizu et al. 2011).

The presence of the double mutation R278Q/R378W reduced the amplitude of both TRPP3 steady-state and tail currents recorded at pH 7.4 when compared to the WT channel (Fig. 3C, 3E and 3G). Moreover, unlike the response of the WT channel to alkalinisation, currents through the double mutant TRPP3 channel were not enhanced at the extracellular pH 9 condition (Fig. 3D, 3F and 3H).

When recorded at pH 7.4, TRPP3 steady-state and tail currents through the R278Q mutant channel were of similar magnitude to those found for the WT channel (Fig. 3E, 3G). However, no significant increase in R278Q TRPP3 current amplitude neither faster deactivation kinetic was observed at pH 9 (Fig. 3F, 3H, Supplementary Fig. 1B, 1D).

The presence of the single substitution R378W also impaired TRPP3 channel activity at pH 7.4, as observed for the double

mutation R278Q/R378W (Fig. 3E, 3G). Yet, R378W TRPP3 steady-state and tail currents were slightly but significantly largest at pH 9 (Fig. 3F, 3H).

Because of the low open probability of TRPP3 at hyperpolarized potentials and the large amplitude of the single channel currents it is possible to record spontaneous single-channel activity in the whole-cell configuration (Shimizu et al. 2009, 2011). Under these experimental conditions, we observed that the alkaline condition (pH 9) only increased the open probability (NPo) of the WT channel (Fig. 4A-C). Besides, as observed for macroscopic TRPP3 steady-state and tail currents, WT and the R278Q channels showed similar activity only at extracellular pH 7.4, whereas the open probability of R378W and R278Q/R378W channels is lower at both neutral (pH 7.4) and alkaline (pH 9) conditions when compared to the WT channel (Fig. 4A-C). As reported before (Shimizu et al. 2011), WT TRPP3 showed similar single-channel conductance at extracellular pH 7.4 and 9, and it was not altered in the mutant TRPP3 channels (Fig. 4D-E).

To test whether the decrease in channel activity produced by the double mutation R278Q/R378W was due to altered trafficking to the plasma membrane, we evaluated the level of co-localization of both, the ectopically-expressed WT and that with the double substitution TRPP3-GFP channels with the plasma membrane marker concanavalin A. We observed a similar pattern for the plasma membrane expression of both channels (Fig. 5). Altogether, in agreement with the extreme Combined Annotation Dependent Depletion (CADD) scores (over 23) of the two genetic variants (R278Q and R378W) that suggest a potentially strong effect at protein level, our data indicate that they, both alone and in combination, exert loss-of-function effects on TRPP3 channels. Thus, substitution R278Q, at the most external α-helix of the extracellular polycystin domain (Hulse et al. 2018), impairs TRPP3 activation by alkaline pH. Substitution R378W, at the intracellular linker between transmembrane domains S2 and S3 of the voltage sensor, strongly reduces channel activity at both neutral (pH 7.4) and alkaline (pH 9) conditions but still show some degree of activation in response to alkalinisation. Double substitution R278Q/ R378W TRPP3 channels show both low activity and impaired modulation by alkaline pH.

Phenotypic analysis

After demonstrating that both substitutions impair the function of TRPP3, we next examined the possible phenotypic differences related to sour tasting between individuals carrying the substitutions and individuals not carrying them. For that, we recruited a total of 44 individuals of Western African (13), Ethiopian (14) and European (18) ancestry, genotyped form both *TRPP3* variants (not all individuals have been genotyped, see Table 1) from a saliva sample, determined their sour recognition threshold and individuals were also asked to fill a qualitative questionnaire related to sour tasting foods. For some of the individuals the saliva pH was also measured. The Ethiopian sampled populations were of Amhara and Oromo ancestry, none of the individuals had a Gumuz ancestry.

All individuals were either, double homozygous for the ancestral alleles (27), double homozygous for the derived alleles (1), or double heterozygous (3) indicating that both loci are in strong linkage disequilibrium. All individuals (4) carrying a derived state of an allele are of African origin, as expected by the general population allele frequencies, that have a zero allele frequencies outside Africa. The limited number of individuals carrying the derived alleles mirrors the low population allele frequencies of the variants within African (Ethiopians and West Africans) populations.

No significant sour recognition differences were observed between the three populations (Fig 6A, Kruskal-Wallis chi-squared = 1.20, df = 2, p-value = 0.549) or between the three different observed genotypes (Fig 6B, Kruskal-Wallis chi-squared = 1.19, df = 2, p-value = 0.552). The limited sample size of individuals carrying the derived alleles is not optimal to apply a statistical test, but at this level we cannot claim any threshold difference related to the two TRPP3 substitutions.

We report a significant positive correlation between saliva pH and recognition threshold (Fig 6C, Pearson's r=0.42, p-value=0.026).

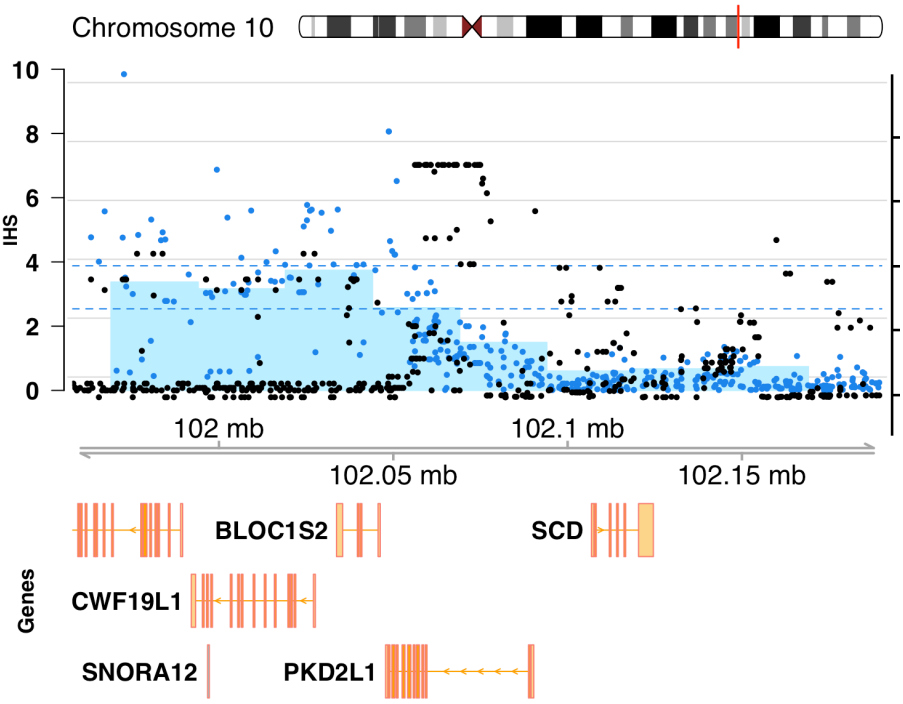

Chromosome 10
10
8
6
4
2
0
IHS
102 mb
102.1 mb
102.05 mb
102.15 mb
Genes
BLOC1S2
SCD
CWF19L1
SNORA12
PKD2L1

Figure 1

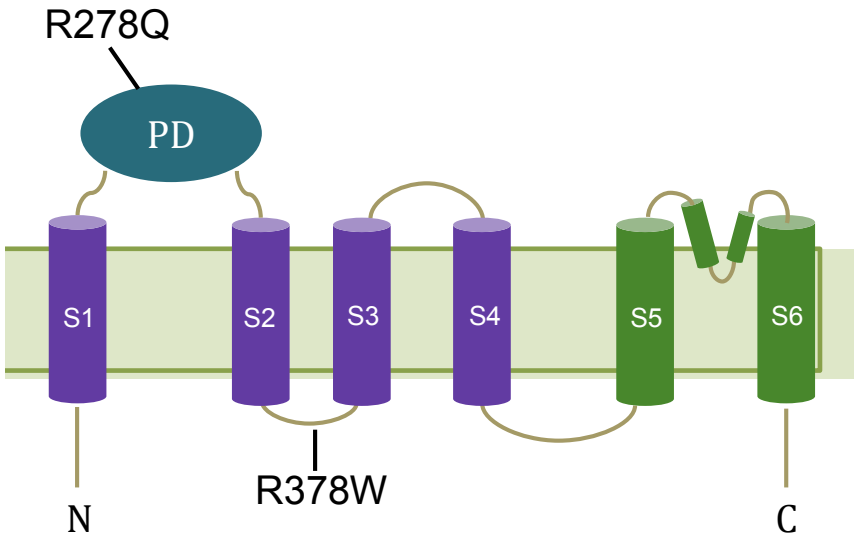


Figure 2

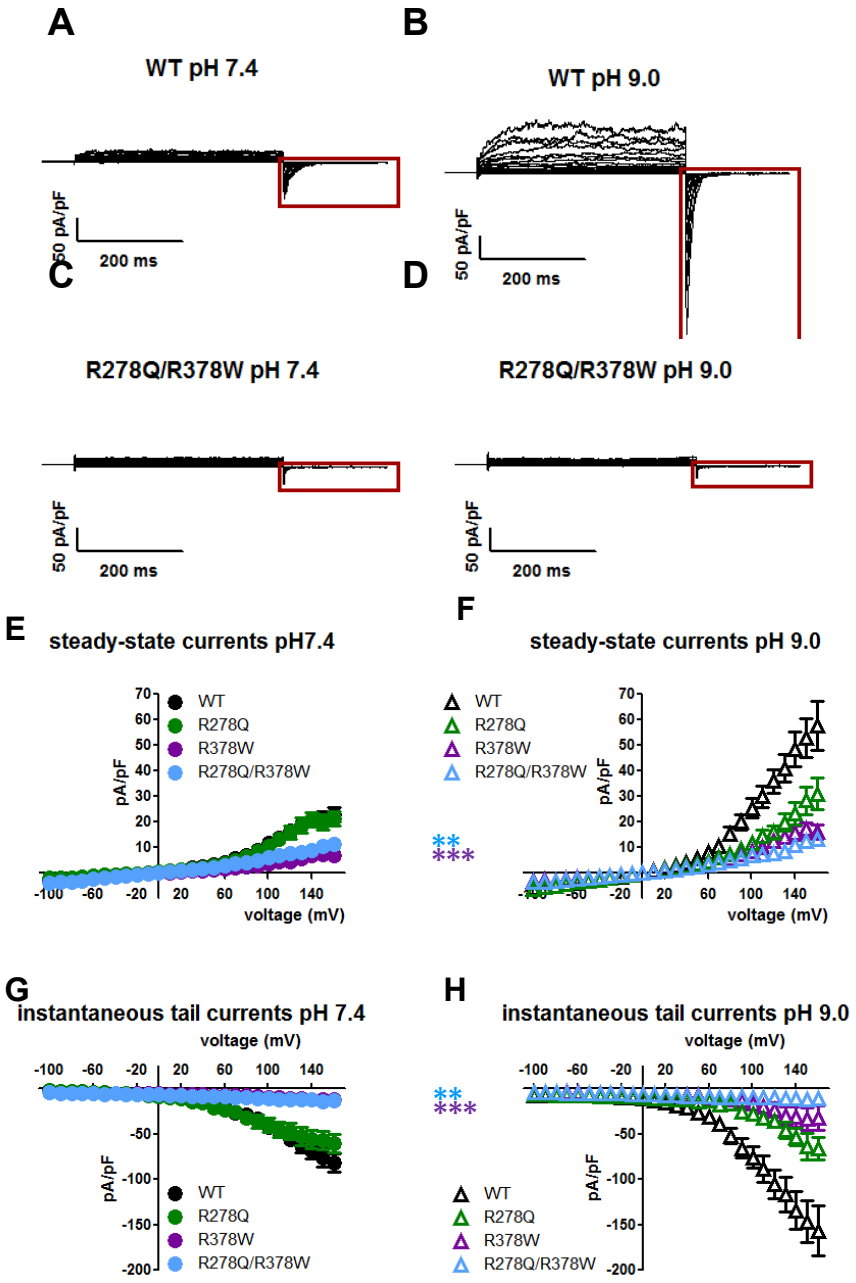

A
WT pH 7.4
50 pA/pF
200 ms
B
WT pH 9.0
50 pA/pF
200 ms
C
R278Q/R378W pH 7.4
50 pA/pF
200 ms
D
R278Q/R378W pH 9.0
50 pA/pF
200 ms
E
steady-state currents pH7.4
WT
R278Q
R378W
R278Q/R378W
pA/pF
voltage (mV)
F
steady-state currents pH 9.0
WT
R278Q
R378W
R278Q/R378W
**
***
pA/pF
voltage (mV)
G
instantaneous tail currents pH 7.4
voltage (mV)
pA/pF
WT
R278Q
R378W
R278Q/R378W
H
instantaneous tail currents pH 9.0
voltage (mV)
**
***
pA/pF
WT
R278Q
R378W
R278Q/R378W

Figure 3

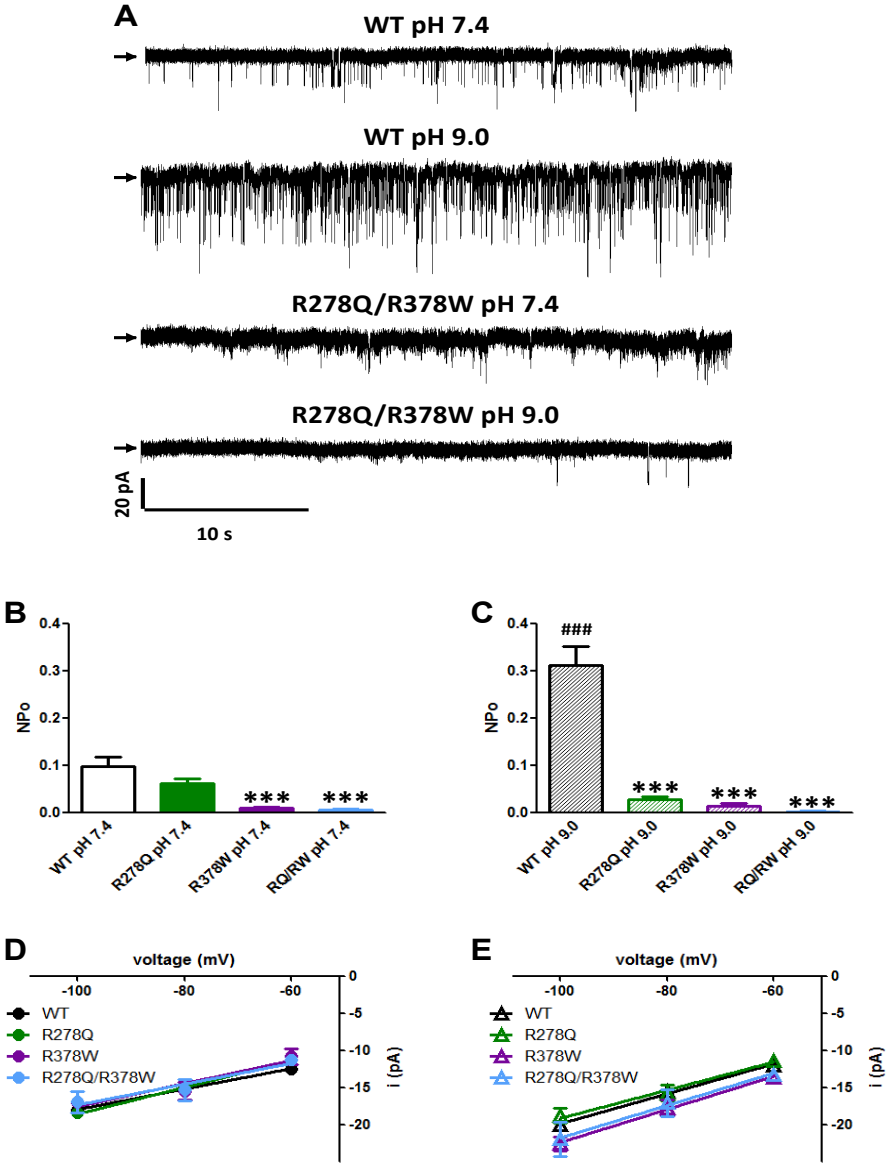

A
WT pH 7.4
WT pH 9.0
R278Q/R378W pH 7.4
R278Q/R378W pH 9.0
20 pA
10 s
B
NPo
0.4
0.3
0.2
0.1
0.0
***
***
WT pH 7.4
R278Q pH 7.4
R378W pH 7.4
RQ/RW pH 7.4
C
NPo
###
***
***
***
WT pH 9.0
R278Q pH 9.0
R378W pH 9.0
RQ/RW pH 9.0
D
voltage (mV)
-100
-80
-60
0
-5
-10
-15
-20
i (pA)
WT
R278Q
R378W
R278Q/R378W
E
voltage (mV)
-100
-80
-60
0
-5
-10
-15
-20
i (pA)
WT
R278Q
R378W
R278Q/R378W

Figure 4

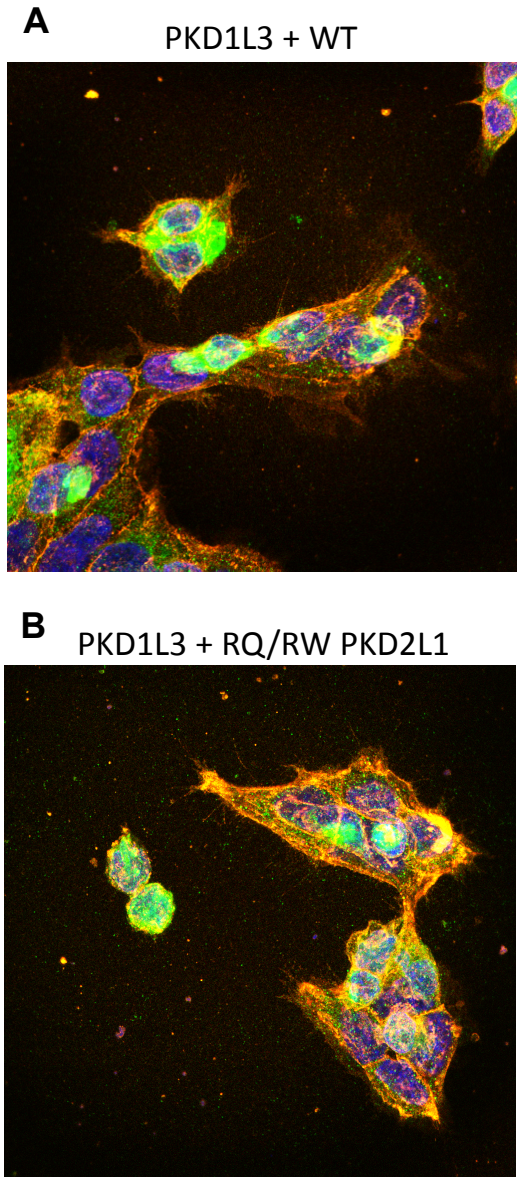


Figure 5

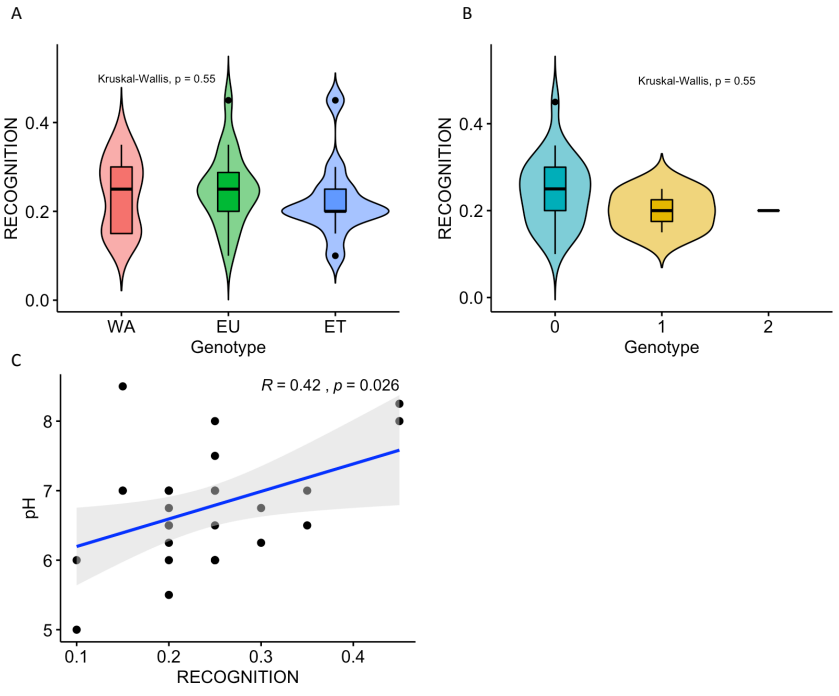

A
Kruskal-Wallis, p = 0.55
RECOGNITION
0.0
0.2
0.4
WA
EU
ET
Genotype
B
Kruskal-Wallis, p = 0.55
RECOGNITION
0.0
0.2
0.4
0
1
2
Genotype
C
R = 0.42 , p = 0.026
pH
5
6
7
8
0.1
0.2
0.3
0.4
RECOGNITION

Figure 6

	Population	Sample	pH	Recognition	R278Q	R278Q	R378W	R378W	Genotype
	WA	S1	-	0.3	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	WA	S2	-	0.15	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	WA	S3	-	0.15	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	WA	S4	-	0.15	-	-	CGG/CGG	R/R	-
	WA	S5	-	0.25	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	WA	S6	-	0.3	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	WA	S7	-	0.35	CAA/CGA	Q/R	-		-
	WA	S8	-	0.35	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	WA	S9	-	0.25	CAA/CGA	Q/R	-	-	-
	WA	S10	-	0.15	CGA/CGA	R/R	-	-	-
*	WA	S11	-	0.3	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	WA	S12	-	0.25	CAA/CGA	Q/R	TGG/CGG	W/R	DHet
	EU	S13	-	0.25	-	-	-	-	-
	EU	S14	7	0.35	-	-	-	-	-
	EU	S15	-	0.3	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	EU	S16	6.25	0.3	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	EU	S17	6.5	0.35	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	EU	S18	7	0.25	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	EU	S19	-	0.15	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	EU	S20	6	0.25	CGA/CGA	R/R	-	-	-
	EU	S21	6.25	0.2	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	EU	S22	6.75	0.2	-	-	-	-	-
	EU	S23	6	0.25	-	-	CGG/CGG	R/R	-
*	ET	S24	7	0.2	CAA/CAA	Q/Q	TGG/TGG	W/W	DDhom
	ET	S25	8.5	0.15	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	ET	S26	6.5	0.2	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
*	ET	S27	5.5	0.2	CAA/CGA	Q/R	TGG/CGG	W/R	Dhet
	ET	S28	8	NA	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	EU	S29	5	0.1	-	-	-	-	-
	EU	S30	7	0.25	CGA/CGA	R/R	CGG/CGG	R/R	ADHom

*	WA	S31	7	0.15	CAA/CGA	Q/R	TGG/CGG	W/R	DHet
	EU	S32	7	0.15	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	EU	S33	7	0.2	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	EU	S34	6	0.25	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	EU	S35	8.25	0.45	CGA/CGA	R/R	-	-	-
	EU	S36	8	0.25	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	ET	S37	6.75	0.3	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	ET	S38	6.5	0.2	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	ET	S39	8	0.45	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	ET	S40	7	0.2	CGA/CGA	R/R	-	-	-
	ET	S41	7.5	0.25	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	ET	S42	6	0.2	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	ET	S43	6	0.1	CGA/CGA	R/R	CGG/CGG	R/R	ADHom
	ET	S44	6.5	0.25	-	-	CGG/CGG	R/R	-

Table 1

Figure Legends

Figure 1. **Genomic context of the region containing TRPP3 from the positive selection analysis (iHS and Fst)**. Each light blue bar represents the absolute mean iHS of a 30 kb window, each blue dot corresponds to the –log10(p-value) of a single variant. The left y-axis corresponds both to the normalized mean absolute iHS score per window and the –log10(p-value) per variant. Blue dotted lines indicate the 99.99th percentile thresholds of the absolute mean iHS per 30 kb windows (lower line) and the –log10(p-value) per variant (upper line) obtained from extensive neutral simulations. Black dots correspond to the Fst values between the Gumuz and CEU population; the right y-axis indicates the scale of Fst values.

Figure 2. Schematic representation of TRPP3 topology in the cellular membrane. TRPP3 has a voltage-sensing domain formed by the S1 to S4 transmembrane alpha helices, the pore domain is formed by the S5 and S6 helices and the polycystin domain (PD). The variants are located in the PD and in the intracellular linker between the S2 and S3 alpha helices.

Figure 3. Loss of TRPP3 function induced by mutations R278Q and R378W.

(**A-D**) Representative whole-cell currents from HEK293 cells co-transfected with cDNAs encoding PKD1L3 and wild-type (WT) or double mutant (R278Q/R378W) TRPP3-GFP and exposed to either extracellular pH 7.4 or pH 9.0, as indicated. Currents were elicited by step pulses from -100 to +160 mV in 10 mV increments with a postpulse to -100 mV, and normalized by cell size (membrane capacitance). Red boxes show tail currents after membrane repolarization to -100 mV. (**E-H**) Average current-voltage (I-V) relationships for steady-state currents (**E, F**) and instantaneous tail currents (**G, H**) at each test pulse in HEK293 cells expressing PKD1L3 and wild-type (WT) or mutants (R278Q, R378W, R278Q/R378W) TRPP3-GFP and exposed to either extracellular pH 7.4 or pH 9.0, as indicated. Area under the I-V curves (AUCs) were calculated for statistical analysis (only values at positive voltages were taken under consideration for steady-state I-V curves). Obtained AUC values for steady-state currents were: WT at pH 7.4 (●, n=14) 1515.8±193.3; WT at pH 9 (△, n=17) 3213.9±444.2; R278Q at pH 7.4 (●, n=9) 1375.6±184.3; R278Q at

pH 9 (△, n=10) 1663.9±291.7; R378W at pH 7.4 (●, n=5) 492±72.5; R378W at pH 9 (△, n=6) 1102.3±176; R278Q/R378W at pH 7.4 (●, n=10) 620.3±100.4; R278Q/R378W at pH 9 (△, n=6) 856.1±132.5. Obtained AUC values for instantaneous tail currents were: WT at pH 7.4 (●, n=14) -6177.3±523.7; WT at pH 9 (△, n=17) -10882.7±1437.6; R278Q at pH 7.4 (●, n=9) -5563.5±1010.1; R278Q at pH 9 (△, n=10) -5034±920.5; R378W at pH 7.4 (●, n=5) -1860±218.1; R378W at pH 9 (△, n=6) -2992±538.2; R278Q/R378W at pH 7.4 (●, n=10) -2005.8±229.2; R278Q/R378W at pH 9 (△, n=6) -1979.2±117.5. $^{\#}P<0.05$, $^{\#\#}P<0.01$ and $^{\#\#\#}P<0.001$ (when compared to the corresponding TRPP3 channel at pH 7.4; one-tail Student's t-test or one-tail Mann-Whitney U-test, as appropriate); *P<0.05, **P<0.01 and ***P<0.001 (when compared to the WT channel at the corresponding extracellular pH condition; Kruskal-Wallis, followed by Dunn post hoc test).

Figure 4. Effect of mutations R278Q and R378W on spontaneous TRPP3 single-channel activity.
(**A**) Representative single-channel activity of wild-type (WT) or double mutant (R278Q/R378W) TRPP3 channels in whole-cell recordings obtained at negative membrane potential (-80 mV) from HEK293 cells co-transfected with cDNAs encoding PKD1L3 and wild-type (WT) or double mutant R278Q/R378W TRPP3-GFP, and exposed to either extracellular pH 7.4 or pH 9.0, as indicated. Arrows indicate the zero current level. (**B, C**). Average channel activity (NPo) in whole-cell recordings obtained at negative membrane potentials (-60, -80 and -100 mV) from HEK293 cells co-transfected with cDNAs encoding PKD1L3 and wild-type (WT) or mutants (R278Q, R378W, R278Q/R378W) TRPP3-GFP, and exposed to either extracellular pH 7.4 or pH 9.0, as indicated. $^{\#\#\#}P<0.0001$ (when compared to WT channels at pH 7.4; one-tail Mann-Whitney U-test); ***P<0.001 (when compared to the WT channel at the corresponding extracellular pH condition; Kruskal-Wallis, followed by Dunn post hoc test). (**D, E**). Average single-channel slope conductances at negative potentials (between -60 mV and -100 mV) obtained in whole-cell recordings from HEK293 cells co-transfected with cDNAs encoding PKD1L3 and wild-type (WT) or mutants (R278Q, R378W, R278Q/R378W) TRPP3-GFP, and exposed to either extracellular pH 7.4 or pH 9.0, as indicated. No

significant differences were found between WT and mutant TRPP3 channels or between pH conditions (one-way ANOVA, P=0.2664). Single-channel conductance values (in pS) for the different TRPP3 channels were: WT at pH 7.4 (●, n=12) 137.8±17.1; WT at pH 9 (△, n=14) 201.7±20.1; R278Q at pH 7.4 (●, n=9) 180±13.7; R278Q at pH 9 (△, n=7) 189.4±35.3; R378W at pH 7.4 (●, n=3) 155.9±35.6; R378W at pH 9 (△, n=7) 222.7±30.4; R278Q/R378W at pH 7.4 (●, n=4) 139.5±46.3; R278Q/R378W at pH 9 (△, n=4) 217.9±73.4.

Figure 5. Substitutions R278Q and R378W do not modify the plasma membrane expression level of heterologously expressed TRPP3 channels.
(**A, B**) Overlaid confocal images of HEK293 cells co-transfected with cDNAs encoding PKD1L3 and wild-type (WT) or double mutant R278Q/R378W TRPP3-GFP (as indicated) showing similar membrane expression (green) of both channels. Plasma membrane was stained with rhodamine-labeled concanavalin A (ConA-Rhod, red).

Figure 6. Genotype-phenotype analysis of the sour recognition thresholds.
(**A**) Sour recognition threshold of the three sampled populations, Ethiopia (ET), West Africa (WA), Europe (EU). No significant threshold differences are observed between populations. (**B**) Genotype analysis of individuals and recognition thresholds. 0 indicates that both alleles are found in ancestral homozygous state, 1 both alleles are heterozygous, 2 both alleles are derived homozygous. No significant differences are found between the different genotypes. (**C**) Positive correlation between saliva pH and sour recognition threshold.

Table Legends

Table 1. Summary of the sampled individuals for the sour taste phenotypic analysis.
The table reports individuals' ancestry, Ethiopian (ET), West African (WA) or European (EU), sample number, sour recognition

threshold, saliva pH and genotype for both variants. Individuals carrying derived alleles are marked with an asterisk (*).

Bibliography

Beall, Cynthia M, Gianpiero L Cavalleri, Libin Deng, Robert C Elston, Yang Gao, Jo Knight, Chaohua Li, et al. 2010. "Natural Selection on EPAS1 (HIF2alpha) Associated with Low Hemoglobin Concentration in Tibetan Highlanders." *Proceedings of the National Academy of Sciences of the United States of America* 107 (25): 11459–64. https://doi.org/10.1073/pnas.1002443107.

Booker, Tom R., Benjamin C. Jackson, and Peter D. Keightley. 2017. "Detecting Positive Selection in the Genome." *BMC Biology*. BioMed Central Ltd. https://doi.org/10.1186/s12915-017-0434-y.

Campbell, Michael C., Alessia Ranciaro, Daniel Zinshteyn, Renata Rawlings-Goss, Jibril Hirbo, Simon Thompson, Dawit Woldemeskel, et al. 2014. "Origin and Differential Selection of Allelic Variation at TAS2R16 Associated with Salicin Bitter Taste Sensitivity in Africa." *Molecular Biology and Evolution* 31 (2): 288–302. https://doi.org/10.1093/molbev/mst211.

DeSimone, J A, G L Heck, and S K DeSimone. 1981. "Active Ion Transport in Dog Tongue: A Possible Role in Taste." *Science (New York, N.Y.)* 214 (4524): 1039–41. https://doi.org/10.1126/science.7302576.

Enattah, Nabil Sabri, Timo Sahi, Erkki Savilahti, Joseph D. Terwilliger, Leena Peltonen, and Irma Järvelä. 2002. "Identification of a Variant Associated with Adult-Type Hypolactasia." *Nature Genetics* 30 (2): 233–37. https://doi.org/10.1038/ng826.

Fan, Shaohua, Matthew E.B. Hansen, Yancy Lo, and Sarah A. Tishkoff. 2016. "Going Global by Adapting Local: A Review of Recent Human Adaptation." *Science*. https://doi.org/10.1126/science.aaf5098.

Fumagalli, Matteo, Ida Moltke, Niels Grarup, Fernando Racimo, Peter Bjerregaard, Marit E Jørgensen, Thorfinn S Korneliussen, et al. 2015. "Greenlandic Inuit Show Genetic Signatures of Diet and Climate Adaptation." *Science (New York, N.Y.)* 349 (6254): 1343–47. https://doi.org/10.1126/science.aab2319.

Harris, H, and H Kalmus. 1949. "The Measurement of Taste Sensitivity to Phenylthiourea." *Annals of Eugenics* 15 (1): 24–31. https://doi.org/10.1111/j.1469-1809.1949.tb02419.x.

Huang, Angela L., Xiaoke Chen, Mark A. Hoon, Jayaram

Chandrashekar, Wei Guo, Dimitri Tränkner, Nicholas J.P. Ryba, and Charles S. Zuker. 2006. "The Cells and Logic for Mammalian Sour Taste Detection." *Nature* 442 (7105): 934–38. https://doi.org/10.1038/nature05084.

Huerta-Sánchez, Emilia, Michael DeGiorgio, Luca Pagani, Ayele Tarekegn, Rosemary Ekong, Tiago Antao, Alexia Cardona, et al. 2013. "Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations." *Molecular Biology and Evolution* 30 (8): 1877–88. https://doi.org/10.1093/molbev/mst089.

Hulse, Raymond E., Zongli Li, Rick K. Huang, Jin Zhang, and David E. Clapham. 2018. "Cryo-EM Structure of the Polycystin 2-L1 Ion Channel." *ELife* 7 (July). https://doi.org/10.7554/eLife.36931.

Huque, Taufiqul, Beverly J. Cowart, Luba Dankulich-Nagrudny, Edmund A. Pribitkin, Douglas L. Bayley, Andrew I. Spielman, Roy S. Feldman, Scott A. Mackler, and Joseph G. Brand. 2009. "Sour Ageusia in Two Individuals Implicates Ion Channels of the ASIC and PKD Families in Human Sour Taste Perception at the Anterior Tongue." *PLoS ONE* 4 (10). https://doi.org/10.1371/journal.pone.0007347.

Inada, Hitoshi, Fuminori Kawabata, Yoshiro Ishimaru, Tohru Fushiki, Hiroaki Matsunami, and Makoto Tominaga. 2008. "Off-Response Property of an Acid-Activated Cation Channel Complex PKD1L3-PKD2L1." *EMBO Reports* 9 (7): 690–97. https://doi.org/10.1038/embor.2008.89.

Ishimaru, Yoshiro, Hitoshi Inada, Momoka Kubota, Hanyi Zhuang, Makoto Tominaga, and Hiroaki Matsunami. 2006. "Transient Receptor Potential Family Members PKD1L3 and PKD2L1 Form a Candidate Sour Taste Receptor." *Proceedings of the National Academy of Sciences* 103 (33): 12569–74. https://doi.org/10.1073/pnas.0602702103.

Izquierdo-Serra, Mercè, Antonio F Martínez-Monseny, Laura López, Julia Carrillo-García, Albert Edo, Juan Darío Ortigoza-Escobar, Óscar García, et al. 2018. "Stroke-Like Episodes and Cerebellar Syndrome in Phosphomannomutase Deficiency (PMM2-CDG): Evidence for Hypoglycosylation-Driven Channelopathy." *International Journal of Molecular Sciences* 19 (2). https://doi.org/10.3390/ijms19020619.

Lu, Zongshi, Yuanting Cui, Xing Wei, Gangyi Yang, Daoyan Liu, and Zhiming Zhu Correspondence. 2018. "Deficiency of

PKD2L1 (TRPP3) Exacerbates Pathological Cardiac Hypertrophy by Augmenting NCX1-Mediated Mitochondrial Calcium Overload." *CellReports* 24: 1639–52. https://doi.org/10.1016/j.celrep.2018.07.022.

Pagani, Luca, Toomas Kivisild, Ayele Tarekegn, Rosemary Ekong, Chris Plaster, Irene Gallego Romero, Qasim Ayub, et al. 2012. "Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool." *American Journal of Human Genetics* 91: 83–96. https://doi.org/10.1016/j.ajhg.2012.05.015.

Ronen, Roy, Nitin Udpa, Eran Halperin, and Vineet Bafna. 2013. "Learning Natural Selection from the Site Frequency Spectrum." *Genetics* 195 (September): 181–93. https://doi.org/10.1534/genetics.113.152587.

Sabeti, Pardis C., David E. Reich, John M. Higgins, Haninah Z. P. Levine, Daniel J. Richter, Stephen F. Schaffner, Stacey B. Gabriel, et al. 2002. "Detecting Recent Positive Selection in the Human Genome from Haplotype Structure." *Nature* 419 (6909): 832–37. https://doi.org/10.1038/nature01140.

Scheinfeldt, Laura B., and Sarah A. Tishkoff. 2013. "Recent Human Adaptation: Genomic Approaches, Interpretation and Insights." *Nature Reviews Genetics* 14 (10): 692–702. https://doi.org/10.1038/nrg3604.

Shimizu, Takahiro, Taiga Higuchi, Takuto Fujii, Bernd Nilius, and Hideki Sakai. 2011. "Bimodal Effect of Alkalization on the Polycystin Transient Receptor Potential Channel, PKD2L1." *Pflugers Archiv : European Journal of Physiology* 461 (5): 507–13. https://doi.org/10.1007/s00424-011-0934-5.

Shimizu, Takahiro, Annelies Janssens, Thomas Voets, and Bernd Nilius. 2009. "Regulation of the Murine TRPP3 Channel by Voltage, PH, and Changes in Cell Volume." *Pflügers Archiv - European Journal of Physiology* 457 (4): 795–807. https://doi.org/10.1007/s00424-008-0558-6.

Simonson, Tatum S, Yingzhong Yang, Chad D Huff, Haixia Yun, Ga Qin, David J Witherspoon, Zhenzhong Bai, et al. 2010. "Genetic Evidence for High-Altitude Adaptation in Tibet." *Science (New York, N.Y.)* 329 (5987): 72–75. https://doi.org/10.1126/science.1189406.

Sternberg, Jenna R, Andrew E Prendergast, Lucie Brosse, Yasmine Cantaut-Belarif, Olivier Thouvenin, Adeline Orts-Del'Immagine, Laura Castillo, et al. 2018. "Pkd2l1 Is Required

for Mechanoception in Cerebrospinal Fluid-Contacting Neurons and Maintenance of Spine Curvature." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-06225-x.

Su, Qiang, Feizhuo Hu, Yuxia Liu, Xiaofei Ge, Changlin Mei, Shengqiang Yu, Aiwen Shen, et al. 2018. "Cryo-EM Structure of the Polycystic Kidney Disease-like Channel PKD2L1." *Nature Communications* 9 (1): 1–12. https://doi.org/10.1038/s41467-018-03606-0.

Tishkoff, Sarah A, Floyd A Reed, Alessia Ranciaro, Benjamin F Voight, Courtney C Babbitt, Jesse S Silverman, Kweli Powell, et al. 2007. "Convergent Adaptation of Human Lactase Persistence in Africa and Europe." *Nature Genetics* 39 (1): 31–40. https://doi.org/10.1038/ng1946.

Yi, Xin, Yu Liang, Emilia Huerta-Sanchez, Xin Jin, Zha Xi Ping Cuo, John E. Pool, N. Xu, et al. 2010. "Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude." *Science* 329 (5987): 75–78. https://doi.org/10.1126/science.1190371.

III. DISCUSSION

The work in this thesis was made possible thanks to the efforts of the scientific community to make genome data and analysis tools as a public resource for research. More importantly, this thesis wouldn't have been possible without the attempt of scientists to overcome the inadequate representation of African populations in genetic studies.

The transcendence of Africa in human evolutionary history is clear from linguistic, archaeological, paleoanthropological and genetic studies. Strong evidence points to Africa as the cradle of Humankind and genetic studies show a complex genetic structure and the highest levels of genetic diversity of African populations. The complex geographical, ecological and climatic diversity of the current African continent and major paleoclimatic shifts (e.g. the "Green Sahara" in the early Holocene) through time are other factors that make African populations extremely interesting in the study of human adaptation. Thus, African populations are excellent for understanding the complex evolutionary history of human populations.

This thesis is focused on two key locations for human evolutionary history, eastern and southern Africa.

Adaptation in Ethiopian populations

The first work of this thesis focused on Ethiopian populations, mainly Afro-asiatic and Nilo-Saharan groups. As previous studies suggested, our results show that Eastern African populations exhibit a match between genetic and linguistic structures. This is further supported by the genome-wide scan of positive selection where closely related populations (Afro-asiatic) share the highest number of selection candidates, whereas the Afro-asiatic and the Nilo-Saharan (Gumuz) share the least. We unravelled putative genetic adaptations by applying two methodologies to capture a wide time-depth of putative positive selection events and took special care with the possible effects of admixture. The positive selection scan captured, in all studied populations, signals of adaptation in genes related to folate metabolism and UV radiation. We also found signals related to defence against pathogens, which is a very common type of signal since organisms have to constantly adapt to new pathogens.

Although we covered a total of five populations that included three branches of Afro-asiatic and one representative of Nilo-Saharan linguistic family, the genetic diversity in Ethiopia is far from being completely sampled and studied. In our study we couldn't address the adaptation to high-altitude, which is one of the most studied cases of adaptive selection in Ethiopians since the Amhara and Oromo samples were not populations living at high-altitude.

Adaptation in the Nama, a KhoeSan population

The aim of the second work of this thesis is to get insights into the KhoeSan population. The KhoeSan, the indigenous populations from southern Africa, are one of the most interesting modern human populations. Some KhoeSan groups still practice a hunter-gatherer lifestyle and belong to the deepest extant human lineage. Their genetic structure does not correlate with language, subsistence strategy or culture, reflecting a complex demographic history. In particular, this thesis focuses on the Nama, a semi-nomadic pastoralist population. Recent migrations from Bantu-speaking individuals, East Africans and Europeans have shaped the Nama genomes. We implemented XP-EHH in European-masked Nama genomes using the Zulu population as a reference to detect and interpret signals of positive selection. We have found strong positive selection signals in the *LCT* loci that show high similarity with the Amhara haplotypes, which is an indicator of the migration from East Africa that brought pastoralism practices into southern Africa. Moreover, we have also found *SLC24A5* as a candidate of adaptive selection. A recent study pointed the European light skin allele as being under recent strong positive selection. Since we have worked with the masked European component of the genomes we did not find such signal but our results indicate a selective sweep at the 5' region of this gene. Dietary and environmental shifts during human evolutionary history are at the base of other selection signals that we find related to diet and metabolism such as the *TG* gene and categories related to lipid metabolism in the enrichment analysis.

Functional validation of a positive selection candidate

The third work of this thesis consists in a functional follow-up study of a putative target of adaptive selection in the Gumuz population

from Ethiopia. The signal of adaptation falls in a region containing *TRPP3*, a gene encoding an ion channel expressed in diverse tissues (taste buds, central nervous system, cardiomyocytes among others). Two nonsynonymous mutations were found at high-derived allele frequency and predicted to be damaging. The electrophysiological analysis (patch-clamp) revealed significant differences between wild type *TRPP3* and the double mutant *TRPP3* which activity is impaired by both mutations. What we couldn't demonstrate was which specific phenotype is increasing the reproductive fitness of the individuals. We tested the effect of both substitutions in sour taste recognition. Our results suggest that there are no differences between populations in the recognition of sour taste. We did not detect any difference between genotypes either even if the sample size of the individuals carrying derived mutations was very low (mainly because the variants segregate at very low frequency outside the Gumuz population). In consequence, *TRPP3* is not likely to be the primary actor in sour taste recognition since individuals carrying substitutions that strongly alter the function of the channel have the same recognition thresholds as the rest. Nonetheless, it remains to be tested whether the aversive pathway or the detection threshold of sour taste remain unaltered. Finally, other possible phenotypes involving other tissues where *TRPP3* has been demonstrated to have an essential role represent other potential adaptive targets to be studied.

Future research

There are still many questions to be answered about African evolutionary history. A more comprehensive sampling of present day African populations (more samples and additional populations) is required to unravel the full spectrum of genetic diversity of humans. An increased availability of ancient genomes would have the potential to cover the gaps about African pre-farming populations since the current African genetic landscape mirrors recent migratory events throughout the continent. These advances in African population history will enable researchers to construct new hypothesis and build better demographic models of human history.

Although many studies focus on the detection of footprints of adaptive selection, the specific mechanisms by which the biological function provides an increased reproductive fitness remain

unknown. The need for an interdisciplinary approach, integrating different layers of *omics* data and follow-up functional studies are essential to validate the signals of positive selection and have a full understanding of the mechanisms of human adaptation.

REFERENCES

Alexander, David H, John Novembre, and Kenneth Lange. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research* 19 (9): 1655–64. https://doi.org/10.1101/gr.094052.109.

Alkorta-Aranburu, Gorka, Cynthia M. Beall, David B. Witonsky, Amha Gebremedhin, Jonathan K. Pritchard, and Anna Di Rienzo. 2012. “The Genetic Architecture of Adaptations to High Altitude in Ethiopia.” Edited by Harmit S. Malik. *PLoS Genetics* 8 (12): e1003110. https://doi.org/10.1371/journal.pgen.1003110.

Alter, Andrea, Louis de Léséleuc, Nguyen Van Thuc, Vu Hong Thai, Nguyen Thu Huong, Nguyen Ngoc Ba, Cynthia Chester Cardoso, et al. 2010. “Genetic and Functional Analysis of Common MRC1 Exon 7 Polymorphisms in Leprosy Susceptibility.” *Human Genetics* 127 (3): 337–48. https://doi.org/10.1007/s00439-009-0775-x.

Ameur, Adam, Stefan Enroth, Åsa Johansson, Ghazal Zaboli, Wilmar Igl, Anna C.V. Johansson, Manuel A. Rivas, et al. 2012. “Genetic Adaptation of Fatty-Acid Metabolism: A Human-Specific Haplotype Increasing the Biosynthesis of Long-Chain Omega-3 and Omega-6 Fatty Acids.” *The American Journal of Human Genetics* 90 (5): 809–20. https://doi.org/10.1016/j.ajhg.2012.03.014.

Arciero, Elena, Simone Andrea Biagini, Yuan Chen, Yali Xue, Donata Luiselli, Chris Tyler-Smith, Luca Pagani, and Qasim Ayub. 2015. “Genes Regulated by Vitamin D in Bone Cells Are Positively Selected in East Asians.” Edited by Arnar Palsson. *PLOS ONE* 10 (12): e0146072. https://doi.org/10.1371/journal.pone.0146072.

Atkinson, Quentin D. 2011. “Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa.” *Science* 332 (6027): 346–49. https://doi.org/10.1126/science.1199295.

Bae, Christopher J., Katerina Douka, and Michael D. Petraglia. 2017. “On the Origin of Modern Humans: Asian Perspectives.” *Science*. American Association for the Advancement of Science.

https://doi.org/10.1126/science.aai9067.
Bamshad, Michael, and Stephen P. Wooding. 2003. "Signatures of Natural Selection in the Human Genome." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg999.
Barnard, Alan. 1992. *Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan Peoples*. Cambridge: Cambridge University Press.
Behar, Doron M., Richard Villems, Himla Soodyall, Jason Blue-Smith, Luisa Pereira, Ene Metspalu, Rosaria Scozzari, et al. 2008. "The Dawn of Human Matrilineal Diversity." *American Journal of Human Genetics* 82 (5): 1130–40. https://doi.org/10.1016/j.ajhg.2008.04.002.
Beltrame, Marcia Holsbach, Meagan A. Rubel, and Sarah A. Tishkoff. 2016. "Inferences of African Evolutionary History from Genomic Data." *Current Opinion in Genetics and Development*. Elsevier Ltd. https://doi.org/10.1016/j.gde.2016.10.002.
Bhatia, Gaurav, Arti Tandon, Nick Patterson, Melinda C. Aldrich, Christine B. Ambrosone, Christopher Amos, Elisa V. Bandera, et al. 2014. "Genome-Wide Scan of 29,141 African Americans Finds No Evidence of Directional Selection since Admixture." *American Journal of Human Genetics* 95 (4): 437–44. https://doi.org/10.1016/j.ajhg.2014.08.011.
Bien, Stephanie A., Genevieve L. Wojcik, Chani J. Hodonsky, Christopher R. Gignoux, Iona Cheng, Tara C. Matise, Ulrike Peters, Eimear E. Kenny, and Kari E. North. 2019. "The Future of Genomic Studies Must Be Globally Representative: Perspectives from PAGE." *Annual Review of Genomics and Human Genetics* 20 (1): 181–200. https://doi.org/10.1146/annurev-genom-091416-035517.
Bindea, Gabriela, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf Herman Fridman, Franck Pagès, Zlatko Trajanoski, and Jérôme Galon. 2009. "ClueGO: A Cytoscape Plug-in to Decipher Functionally Grouped Gene Ontology and Pathway Annotation Networks." *Bioinformatics* 25 (8): 1091–93. https://doi.org/10.1093/bioinformatics/btp101.
Breton, Gwenna, Carina M. Schlebusch, Marlize Lombard,

Per Sjödin, Himla Soodyall, and Mattias Jakobsson. 2014. “Lactase Persistence Alleles Reveal Partial East African Ancestry of Southern African Khoe Pastoralists.” *Current Biology* 24 (8): 852–58. https://doi.org/10.1016/J.CUB.2014.02.041.

Burgdorf, S., A. Kautz, V. Bohnert, P. A. Knolle, and C. Kurts. 2007. “Distinct Pathways of Antigen Uptake and Intracellular Routing in CD4 and CD8 T Cell Activation.” *Science* 316 (5824): 612–16. https://doi.org/10.1126/science.1137971.

Calafell, Francesc, and David Comas. 2014. “Genetics and the Reconstruction of African Population History,” 379–400.

Campbell, Michael C., and Sarah A. Tishkoff. 2008. “African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping.” *Annual Review of Genomics and Human Genetics* 9 (1): 403–33. https://doi.org/10.1146/annurev.genom.9.081307.164258.

———. 2010. “The Evolution of Human Genetic and Phenotypic Variation in Africa.” *Current Biology* 20 (4): R166–73. https://doi.org/10.1016/J.CUB.2009.11.050.

Cann, Howard M, Claudia de Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, et al. 2002. “A Human Genome Diversity Cell Line Panel.” *Science (New York, N.Y.)* 296 (5566): 261–62. http://www.ncbi.nlm.nih.gov/pubmed/11954565.

Cavalli-Sforza, L. L., E. Minch, and J. L. Mountain. 1992. “Coevolution of Genes and Languages Revisited.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (12): 5620–24. https://doi.org/10.1073/pnas.89.12.5620.

Chahal, Harvind S., Yuan Lin, Katherine J. Ransohoff, David A. Hinds, Wenting Wu, Hong-Ji Dai, Abrar A. Qureshi, et al. 2016. “Genome-Wide Association Study Identifies Novel Susceptibility Loci for Cutaneous Squamous Cell Carcinoma.” *Nature Communications* 7 (July): 12048. https://doi.org/10.1038/ncomms12048.

Choudhury, Ananyo, Michèle Ramsay, Scott Hazelhurst, Shaun Aron, Soraya Bardien, Gerrit Botha, Emile R.

Chimusa, et al. 2017. “Whole-Genome Sequencing for an Enhanced Understanding of Genetic Variation among South Africans.” *Nature Communications* 8 (1): 2062. https://doi.org/10.1038/s41467-017-00663-9.

Clark, Andrew G, Stephen Glanowski, Rasmus Nielsen, Paul D Thomas, Anish Kejariwal, Melissa A Todd, David M Tanenbaum, et al. 2003. “Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios.” *Science (New York, N.Y.)* 302 (5652): 1960–63. https://doi.org/10.1126/science.1088821.

Clarkson, Chris, Zenobia Jacobs, Ben Marwick, Richard Fullagar, Lynley Wallis, Mike Smith, Richard G. Roberts, et al. 2017. “Human Occupation of Northern Australia by 65,000 Years Ago.” *Nature* 547 (7663): 306–10. https://doi.org/10.1038/nature22968.

Coop, Graham, Joseph K. Pickrell, John Novembre, Sridhar Kudaravalli, Jun Li, Devin Absher, Richard M. Myers, Luigi Luca Cavalli-Sforza, Marcus W. Feldman, and Jonathan K. Pritchard. 2009. “The Role of Geography in Human Adaptation.” Edited by Mikkel H. Schierup. *PLoS Genetics* 5 (6): e1000500. https://doi.org/10.1371/journal.pgen.1000500.

Crevecoeur, I., A. Brooks, I. Ribot, E. Cornelissen, and P. Semal. 2016. “Late Stone Age Human Remains from Ishango (Democratic Republic of Congo): New Insights on Late Pleistocene Modern Human Diversity in Africa.” *Journal of Human Evolution*. https://doi.org/10.1016/j.jhevol.2016.04.003.

Darwin, Charles, and Alfred Wallace. 1858. “On the Tendency of Species to Form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection.” *Journal of the Proceedings of the Linnean Society of London. Zoology* 3 (9): 45–62. https://doi.org/10.1111/j.1096-3642.1858.tb02500.x.

Day, M. H., M. D. Leakey, and C. Magori. 1980. “A New Hominid Fossil Skull (L.H. 18) from the Ngaloba Beds, Laetoli, Northern Tanzania.” *Nature* 284 (5751): 55–56. https://doi.org/10.1038/284055a0.

DeSimone, J A, G L Heck, and S K DeSimone. 1981. “Active Ion Transport in Dog Tongue: A Possible Role in Taste.” *Science (New York, N.Y.)* 214 (4524): 1039–41.

https://doi.org/10.1126/science.7302576.
Diamond, Jared, and Peter Bellwood. 2003. “Farmers and Their Languages: The First Expansions.” *Science*. https://doi.org/10.1126/science.1078208.
Dobon, Begoña, Hisham Y. Hassan, Hafid Laayouni, Pierre Luisi, Isis Ricaño-Ponce, Alexandra Zhernakova, Cisca Wijmenga, et al. 2015. “The Genetics of East African Populations: A Nilo-Saharan Component in the African Genetic Landscape.” *Scientific Reports* 5 (1): 9996. https://doi.org/10.1038/srep09996.
Dubois, Aurelie, Catherine François, Veronique Descamps, Carole Fournier, Czeslaw Wychowski, Jean Dubuisson, Sandrine Castelain, and Gilles Duverlie. 2009. “Enhanced Anti-HCV Activity of Interferon Alpha 17 Subtype.” *Virology Journal* 6 (1): 70. https://doi.org/10.1186/1743-422X-6-70.
East, Lucy, and Clare M Isacke. 2002. “The Mannose Receptor Family.” *Biochimica et Biophysica Acta - General Subjects*. Elsevier. https://doi.org/10.1016/S0304-4165(02)00319-7.
Elaldi, Nazif, Meral Yilmaz, Binnur Bagci, Izzet Yelkovan, Gokhan Bagci, Mustafa Gokhan Gozel, Aynur Engin, Mehmet Bakir, and Ilyas Dokmetas. 2016. “Relationship between *IFNA1* , *IFNA5* , *IFNA10* , and *IFNA17* Gene Polymorphisms and Crimean-Congo Hemorrhagic Fever Prognosis in a Turkish Population Range.” *Journal of Medical Virology* 88 (7): 1159–67. https://doi.org/10.1002/jmv.24456.
Enard, David, Philipp W Messer, and Dmitri A Petrov. 2014. “Genome-Wide Signals of Positive Selection in Human Evolution.” *Genome Research* 24 (6): 885–95. https://doi.org/10.1101/gr.164822.113.
Enattah, Nabil Sabri, Timo Sahi, Erkki Savilahti, Joseph D. Terwilliger, Leena Peltonen, and Irma Järvelä. 2002. “Identification of a Variant Associated with Adult-Type Hypolactasia.” *Nature Genetics* 30 (2): 233–37. https://doi.org/10.1038/ng826.
Fagny, Maud, Etienne Patin, David Enard, Luis B. Barreiro, Lluis Quintana-Murci, and Guillaume Laval. 2014. “Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data

Sets.” *Molecular Biology and Evolution* 31 (7): 1850–68. https://doi.org/10.1093/molbev/msu118.
Fan, Shaohua, Matthew E.B. Hansen, Yancy Lo, and Sarah A. Tishkoff. 2016. “Going Global by Adapting Local: A Review of Recent Human Adaptation.” *Science*. https://doi.org/10.1126/science.aaf5098.
Ferrer-Admetlla, Anna, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. 2014. “On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure.” *Molecular Biology and Evolution* 31 (5): 1275–91. https://doi.org/10.1093/molbev/msu077.
Filippo, Cesare de, Koen Bostoen, Mark Stoneking, and Brigitte Pakendorf. 2012. “Bringing Together Linguistic and Genetic Evidence to Test the Bantu Expansion.” *Proceedings of the Royal Society B: Biological Sciences* 279 (1741): 3256–63. https://doi.org/10.1098/rspb.2012.0318.
Filippo, Cesare De, Felix M Key, Silvia Ghirotto, Andrea Benazzo, Juan R Meneu, Antje Weihmann, Genís Parra, Eric D Green, and Aida M Andrés. 2016. “Recent Selection Changes in Human Genes under Long-Term Balancing Selection.” *Molecular Biology and Evolution* 33 (6): 1435–47. https://doi.org/10.1093/molbev/msw023.
Fumagalli, Matteo, Ida Moltke, Niels Grarup, Fernando Racimo, Peter Bjerregaard, Marit E Jørgensen, Thorfinn S Korneliussen, et al. 2015. “Greenlandic Inuit Show Genetic Signatures of Diet and Climate Adaptation.” *Science (New York, N.Y.)* 349 (6254): 1343–47. https://doi.org/10.1126/science.aab2319.
Gallego Llorente, M, E R Jones, A Eriksson, V Siska, K W Arthur, J W Arthur, M C Curtis, et al. 2015. “Ancient Ethiopian Genome Reveals Extensive Eurasian Admixture throughout the African Continent.” *Science (New York, N.Y.)* 350 (6262): 820–22. https://doi.org/10.1126/science.aad2879.
Garud, Nandita R., Philipp W. Messer, Erkan O. Buzbas, and Dmitri A. Petrov. 2015. “Recent Selective Sweeps in North American Drosophila Melanogaster Show Signatures of Soft Sweeps.” *PLoS Genetics* 11 (2): 1–32. https://doi.org/10.1371/journal.pgen.1005004.
Gautier, Mathieu, Alexander Klassmann, and Renaud Vitalis.

2017. “Rehh 2.0: A Reimplementation of the R Package Rehh to Detect Positive Selection from Haplotype Structure.” *Molecular Ecology Resources* 17 (1): 78–90. https://doi.org/10.1111/1755-0998.12634.

Gazi, Umut, and Luisa Martinez-Pomares. 2009. “Influence of the Mannose Receptor in Host Immune Responses.” *Immunobiology* 214 (7): 554–61. https://doi.org/10.1016/J.IMBIO.2008.11.004.

Gibbons, Ann. 2017. “World’s Oldest *Homo Sapiens* Fossils Found in Morocco.” *Science*, June. https://doi.org/10.1126/science.aan6934.

Gibbs, Richard A., John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli Yu, Houcan Zhang, Changqing Zeng, et al. 2003. “The International HapMap Project.” *Nature* 426 (6968): 789–96. https://doi.org/10.1038/nature02168.

Gibbs, Richard A., Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Granka, J. M., B. M. Henn, C. R. Gignoux, J. M. Kidd, C. D. Bustamante, and M. W. Feldman. 2012. “Limited Evidence for Classic Selective Sweeps in African Populations.” *Genetics* 192 (3): 1049–64. https://doi.org/10.1534/genetics.112.144071.

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, et al. 2011. “Demographic History and Rare Allele Sharing among Human Populations.” *Proceedings of the National Academy of Sciences* 108 (29): 11983–88. https://doi.org/10.1073/pnas.1019276108.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. “A Draft Sequence of the Neandertal Genome.” *Science* 328 (5979): 710–22. https://doi.org/10.1126/science.1188021.

Grove, Matt, Henry Lamb, Helen Roberts, Sarah Davies, Mike Marshall, Richard Bates, and Dei Huws. 2015. “Climatic Variability, Plasticity, and Dispersal: A Case Study from Lake Tana, Ethiopia.” *Journal of Human Evolution* 87

(October): 32–47. https://doi.org/10.1016/j.jhevol.2015.07.007.
Grun, R., J. S. Brink, N. A. Spooner, L. Taylor, C. B. Stringer, R. G. Franciscus, and A. S. Murray. 1996. “Direct Dating of Florisbad Hominid [2].” *Nature*. https://doi.org/10.1038/382500a0.
Grün, Rainer, Chris Stringer, Frank McDermott, Roger Nathan, Naomi Porat, Steve Robertson, Lois Taylor, Graham Mortimer, Stephen Eggins, and Malcolm McCulloch. 2005. “U-Series and ESR Analyses of Bones and Teeth Relating to the Human Burials from Skhul.” *Journal of Human Evolution* 49 (3): 316–34. https://doi.org/10.1016/j.jhevol.2005.04.006.
Gurdasani, Deepti, Tommy Carstensen, Fasil Tekola-Ayele, Luca Pagani, Ioanna Tachmazidou, Konstantinos Hatzikotoulas, Savita Karthikeyan, et al. 2015. “The African Genome Variation Project Shapes Medical Genetics in Africa.” *Nature* 517 (7534): 327–32. https://doi.org/10.1038/nature13997.
Haller, Benjamin C, and Philipp W Messer. 2017. “SLiM 2: Flexible, Interactive Forward Genetic Simulations.” *Molecular Biology and Evolution* 34 (1): 230–40. https://doi.org/10.1093/molbev/msw211.
Hamblin, Martha T., and Anna Di Rienzo. 2000. “Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus.” *American Journal of Human Genetics* 66 (5): 1669–79. https://doi.org/10.1086/302879.
Hammer, Michael F., August E. Woerner, Fernando L. Mendez, Joseph C. Watkins, and Jeffrey D. Wall. 2011. “Genetic Evidence for Archaic Admixture in Africa.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (37): 15123–28. https://doi.org/10.1073/pnas.1109300108.
Han, Jiali, Graham A. Colditz, and David J. Hunter. 2007. “Polymorphisms in the MTHFR and VDR Genes and Skin Cancer Risk.” *Carcinogenesis* 28 (2): 390–97. https://doi.org/10.1093/carcin/bgl156.
Harvati, Katerina, Carolin Röding, Abel M. Bosman, Fotios A. Karakostis, Rainer Grün, Chris Stringer, Panagiotis Karkanas, et al. 2019. “Apidima Cave Fossils Provide

Earliest Evidence of Homo Sapiens in Eurasia." *Nature*, July. https://doi.org/10.1038/s41586-019-1376-z.
Harvati, Katerina, Chris Stringer, Rainer Grün, Maxime Aubert, Philip Allsworth-Jones, and Caleb Adebayo Folorunso. 2011. "The Later Stone Age Calvaria from Iwo Eleru, Nigeria: Morphology and Chronology." *PLoS ONE* 6 (9). https://doi.org/10.1371/journal.pone.0024024.
Hattori, Takeshi, Satoshi Konno, Nobuyuki Hizawa, Akira Isada, Ayumu Takahashi, Kaoruko Shimizu, Kenichi Shimizu, et al. 2009. "Genetic Variants in the Mannose Receptor Gene (MRC1) Are Associated with Asthma in Two Independent Populations." *Immunogenetics* 61 (11–12): 731–38. https://doi.org/10.1007/s00251-009-0403-x.
Hattori, Takeshi, Satoshi Konno, Ayumu Takahashi, Akira Isada, Kaoruko Shimizu, Kenichi Shimizu, Natsuko Taniguchi, et al. 2010. "Genetic Variants in Mannose Receptor Gene (MRC1) Confer Susceptibility to Increased Risk of Sarcoidosis." *BMC Medical Genetics* 11 (1): 151. https://doi.org/10.1186/1471-2350-11-151.
Hauser, Marc D., Charles Yang, Robert C. Berwick, Ian Tattersall, Michael J. Ryan, Jeffrey Watumull, Noam Chomsky, and Richard C. Lewontin. 2014. "The Mystery of Language Evolution." *Frontiers in Psychology*. Frontiers Research Foundation. https://doi.org/10.3389/fpsyg.2014.00401.
He, Shanshan, Zhen Zhao, Yongfei Yang, Douglas O'Connell, Xiaowei Zhang, Soohwan Oh, Binyun Ma, et al. 2015. "Truncating Mutation in the Autophagy Gene UVRAG Confers Oncogenic Properties and Chemosensitivity in Colorectal Cancers." *Nature Communications* 6 (1): 7839. https://doi.org/10.1038/ncomms8839.
Henn, Brenna M., Christopher Gignoux, Alice A. Lin, Peter J. Oefner, Peidong Shen, Rosaria Scozzari, Fulvio Cruciani, Sarah A. Tishkoff, Joanna L. Mountain, and Peter A. Underhill. 2008. "Y-Chromosomal Evidence of a Pastoralist Migration through Tanzania to Southern Africa." *Proceedings of the National Academy of Sciences* 105 (31): 10693–98. https://doi.org/10.1073/PNAS.0801184105.
Henn, Brenna M., Christopher R. Gignoux, Matthew Jobin,

Julie M. Granka, J. M. Macpherson, Jeffrey M. Kidd, Laura Rodríguez-Botigué, et al. 2011. “Hunter-Gatherer Genomic Diversity Suggests a Southern African Origin for Modern Humans.” *Proceedings of the National Academy of Sciences* 108 (13): 5154–62. https://doi.org/10.1073/PNAS.1017511108.

Henn, Brenna M, Teresa E Steele, and Timothy D Weaver. 2018. “Clarifying Distinct Models of Modern Human Origins in Africa.” *Current Opinion in Genetics and Development*. Elsevier Current Trends. https://doi.org/10.1016/j.gde.2018.10.003.

Hermisson, Joachim, and Pleuni S. Pennings. 2017. “Soft Sweeps and beyond: Understanding the Patterns and Probabilities of Selection Footprints under Rapid Adaptation.” *Methods in Ecology and Evolution* 8 (6): 700–716. https://doi.org/10.1111/2041-210X.12808.

Hermisson, Joachim, and Pleuni S Pennings. 2005. “Soft Sweeps: Molecular Population Genetics of Adaptation from Standing Genetic Variation.” *Genetics* 169 (4): 2335–52. https://doi.org/10.1534/genetics.104.036947.

Hershkovitz, Israel, Gerhard W. Weber, Rolf Quam, Mathieu Duval, Rainer Grün, Leslie Kinsley, Avner Ayalon, et al. 2018. “The Earliest Modern Humans Outside Africa.” *Science* 359 (6374): 456–59. https://doi.org/10.1126/science.aap8369.

Hofer, T., N. Ray, D. Wegmann, and L. Excoffier. 2009. “Large Allele Frequency Differences between Human Continental Groups Are More Likely to Have Occurred by Drift During Range Expansions than by Selection.” *Annals of Human Genetics* 73 (1): 95–108. https://doi.org/10.1111/j.1469-1809.2008.00489.x.

Huang, Angela L., Xiaoke Chen, Mark A. Hoon, Jayaram Chandrashekar, Wei Guo, Dimitri Tränkner, Nicholas J.P. Ryba, and Charles S. Zuker. 2006. “The Cells and Logic for Mammalian Sour Taste Detection.” *Nature* 442 (7105): 934–38. https://doi.org/10.1038/nature05084.

Hublin, Jean-Jacques, Abdelouahed Ben-Ncer, Shara E. Bailey, Sarah E. Freidline, Simon Neubauer, Matthew M. Skinner, Inga Bergmann, et al. 2017. “New Fossils from Jebel Irhoud, Morocco and the Pan-African Origin of Homo Sapiens.” *Nature* 546 (7657): 289–92.

https://doi.org/10.1038/nature22336.
Huerta-Sánchez, Emilia, Michael DeGiorgio, Luca Pagani, Ayele Tarekegn, Rosemary Ekong, Tiago Antao, Alexia Cardona, et al. 2013a. “Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations.” *Molecular Biology and Evolution* 30 (8): 1877–88. https://doi.org/10.1093/molbev/mst089.
———. 2013b. “Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations.” *Molecular Biology and Evolution* 30 (8): 1877–88. https://doi.org/10.1093/molbev/mst089.
Huque, Taufiqul, Beverly J. Cowart, Luba Dankulich-Nagrudny, Edmund A. Pribitkin, Douglas L. Bayley, Andrew I. Spielman, Roy S. Feldman, Scott A. Mackler, and Joseph G. Brand. 2009. “Sour Ageusia in Two Individuals Implicates Ion Channels of the ASIC and PKD Families in Human Sour Taste Perception at the Anterior Tongue.” *PLoS ONE* 4 (10). https://doi.org/10.1371/journal.pone.0007347.
Ingman, Max, Henrik Kaessmann, Svante Pääbo, and Ulf Gyllensten. 2000. “Mitochondrial Genome Variation and the Origin of Modern Humans.” *Nature* 408 (6813): 708–13. https://doi.org/10.1038/35047064.
Innan, Hideki, and Yuseob Kim. 2004. “Pattern of Polymorphism after Strong Artificial Selection in a Domestication Event.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (29): 10667–72. https://doi.org/10.1073/pnas.0401720101.
Ishimaru, Yoshiro, Hitoshi Inada, Momoka Kubota, Hanyi Zhuang, Makoto Tominaga, and Hiroaki Matsunami. 2006. “Transient Receptor Potential Family Members PKD1L3 and PKD2L1 Form a Candidate Sour Taste Receptor.” *Proceedings of the National Academy of Sciences* 103 (33): 12569–74. https://doi.org/10.1073/pnas.0602702103.
Jablonski, Nina G, and George Chaplin. 2010. “Human Skin Pigmentation as an Adaptation to UV Radiation.” *Proceedings of the National Academy of Sciences* 107 (Supplement_2): 8962–68. https://doi.org/10.1073/pnas.0914628107.

Jakobsson, Mattias, Sonja W. Scholz, Paul Scheet, J. Raphael Gibbs, Jenna M. VanLiere, Hon Chung Fung, Zachary A. Szpiech, et al. 2008. “Genotype, Haplotype and Copy-Number Variation in Worldwide Human Populations.” *Nature* 451 (7181): 998–1003. https://doi.org/10.1038/nature06742.

Jeong, Choongwon, Gorka Alkorta-Aranburu, Buddha Basnyat, Maniraj Neupane, David B. Witonsky, Jonathan K. Pritchard, Cynthia M. Beall, and Anna Di Rienzo. 2014. “Admixture Facilitates Genetic Adaptations to High Altitude in Tibet.” *Nature Communications* 5 (February): 3281. https://doi.org/10.1038/ncomms4281.

Jin, Wenfei, Shuhua Xu, Haifeng Wang, Yongguo Yu, Yiping Shen, Bailin Wu, and Li Jin. 2012. “Genome-Wide Detection of Natural Selection in African Americans Pre- and Post-Admixture.” *Genome Research* 22 (3): 519–27. https://doi.org/10.1101/gr.124784.111.

Jobling, Mark A., and Chris Tyler-Smith. 2017. “Human Y-Chromosome Variation in the Genome-Sequencing Era.” *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg.2017.36.

Jones, Patrice, Mark Lucock, Martin Veysey, and Emma Beckett. 2018. “The Vitamin D^{-}Folate Hypothesis as an Evolutionary Model for Skin Pigmentation: An Update and Integration of Current Ideas.” *Nutrients* 10 (5). https://doi.org/10.3390/nu10050554.

Jouganous, Julien, Will Long, Aaron P. Ragsdale, and Simon Gravel. 2017. “Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation.” *Genetics* 206 (3): 1549–67. https://doi.org/10.1534/genetics.117.200493.

Kwiatkowski, Dominic P. 2005. “How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria.” *The American Journal of Human Genetics* 77 (2): 171–92. https://doi.org/10.1086/432519.

Lee, K, I Chiu, RLP Santos-Cortez, S Basit, S Khan, Z Azeem, PB Andrade, SS Kim, W Ahmad, and SM Leal. 2013. “Novel *OTOA* Mutations Cause Autosomal Recessive Non-Syndromic Hearing Impairment in Pakistani Families.” *Clinical Genetics* 84 (3): 294–96.

https://doi.org/10.1111/cge.12047.
Lewis, M. Paul. 2009. *Ethnologue: Languages of the World*. SIL international DallasTX.
Li, Mingcai, Xiaojin Liu, Yanchun Zhou, and Shao Bo Su. 2009. “Interferon- s: The Modulators of Antivirus, Antitumor, and Immune Responses.” *Journal of Leukocyte Biology* 86 (1): 23–32. https://doi.org/10.1189/jlb.1208761.
Li, Mingzhou, Shilin Tian, Long Jin, Guangyu Zhou, Ying Li, Yuan Zhang, Tao Wang, et al. 2013. “Genomic Analyses Identify Distinct Patterns of Selection in Domesticated Pigs and Tibetan Wild Boars.” *Nature Genetics* 45 (12): 1431–38. https://doi.org/10.1038/ng.2811.
Li, Sen, Carina Schlebusch, and Mattias Jakobsson. 2014. “Genetic Variation Reveals Large-Scale Population Expansion and Migration during the Expansion of Bantu-Speaking Peoples.” *Proceedings of the Royal Society B: Biological Sciences* 281 (1793): 20141448. https://doi.org/10.1098/rspb.2014.1448.
Lin, Meng, Rebecca L Siford, Alicia R Martin, Shigeki Nakagome, Marlo Möller, Eileen G Hoal, Carlos D Bustamante, Christopher R Gignoux, and Brenna M Henn. 2018. “Rapid Evolution of a Skin-Lightening Allele in Southern African KhoeSan.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (52): 13324–29. https://doi.org/10.1073/pnas.1801948115.
Liu, Wu, María Martinón-Torres, Yan-jun Cai, Song Xing, Hao-wen Tong, Shu-wen Pei, Mark Jan Sier, et al. 2015. “The Earliest Unequivocally Modern Humans in Southern China.” *Nature* 526 (7575): 696–99. https://doi.org/10.1038/nature15696.
Lohmueller, Kirk E., Carlos D. Bustamante, and Andrew G. Clark. 2011. “Detecting Directional Selection in the Presence of Recent Admixture in African-Americans.” *Genetics* 187 (3): 823–35. https://doi.org/10.1534/genetics.110.122739.
Lorente-Galdos, Belen, Oscar Lao, Gerard Serra-Vidal, Gabriel Santpere, Lukas F K Kuderna, Lara R Arauna, Karima Fadhlaoui-Zid, et al. 2019. “Whole-Genome Sequence Analysis of a Pan African Set of Samples

Reveals Archaic Gene Flow from an Extinct Basal Population of Modern Humans into Sub-Saharan Populations." *Genome Biology* 20 (1): 77. https://doi.org/10.1186/s13059-019-1684-5.
Macholdt, Enrico, Vera Lede, Chiara Barbieri, Sununguko W Mpoloka, Hua Chen, Montgomery Slatkin, Brigitte Pakendorf, and Mark Stoneking. 2014. "Tracing Pastoralist Migrations to Southern Africa with Lactase Persistence Alleles." *Current Biology : CB* 24 (8): 875–79. https://doi.org/10.1016/j.cub.2014.03.027.
Malaspinas, Anna Sapfo, Michael C. Westaway, Craig Muller, Vitor C. Sousa, Oscar Lao, Isabel Alves, Anders Bergström, et al. 2016. "A Genomic History of Aboriginal Australia." *Nature*. Nature Publishing Group. https://doi.org/10.1038/nature18299.
Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016. "The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations." *Nature* 538 (7624): 201–6. https://doi.org/10.1038/nature18964.
Martin, Alicia R., Solomon Teferra, Marlo Möller, Eileen G. Hoal, and Mark J. Daly. 2018. "The Critical Needs and Challenges for Genetic Architecture Studies in Africa." *Current Opinion in Genetics and Development*. Elsevier Ltd. https://doi.org/10.1016/j.gde.2018.08.005.
Martinez-Pomares, Luisa. 2012. "The Mannose Receptor." *Journal of Leukocyte Biology* 92 (6): 1177–86. https://doi.org/10.1189/jlb.0512231.
Mathieson, Iain, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Bastien Llamas, Joseph Pickrell, Harald Meller, Manuel A Rojo Guerra, and Johannes Krause. n.d. "Eight Thousand Years of Natural Selection in Europe."
McDougall, Ian, Francis H. Brown, and John G. Fleagle. 2005. "Stratigraphic Placement and Age of Modern Humans from Kibish, Ethiopia." *Nature* 433 (7027): 733–36. https://doi.org/10.1038/nature03258.
Melé, Marta, Asif Javed, Marc Pybus, Pierre Zalloua, Marc Haber, David Comas, Mihai G. Netea, et al. 2012. "Recombination Gives a New Insight in the Effective Population Size and the History of the Old World Human Populations." *Molecular Biology and Evolution* 29 (1):

25–30. https://doi.org/10.1093/molbev/msr213.
Metz, Jack. 2007. “Folic Acid Metabolism and Malaria.” *Food and Nutrition Bulletin* 28 (4_suppl4): S540–49. https://doi.org/10.1177/15648265070284S407.
Meyer, Matthias, Martin Kircher, Marie Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, et al. 2012. “A High-Coverage Genome Sequence from an Archaic Denisovan Individual.” *Science* 338 (6104): 222–26. https://doi.org/10.1126/science.1224344.
Mijares, Armand Salvador, Florent Détroit, Philip Piper, Rainer Grün, Peter Bellwood, Maxime Aubert, Guillaume Champion, Nida Cuevas, Alexandra De Leon, and Eusebio Dizon. 2010. “New Evidence for a 67,000-Year-Old Human Presence at Callao Cave, Luzon, Philippines.” *Journal of Human Evolution* 59 (1): 123–32. https://doi.org/10.1016/j.jhevol.2010.04.008.
Mills, Melinda C., and Charles Rahal. 2019. “A Scientometric Review of Genome-Wide Association Studies.” *Communications Biology* 2 (1). https://doi.org/10.1038/s42003-018-0261-x.
Montinaro, Francesco, George B J Busby, Miguel Gonzalez-Santos, Ockie Oosthuitzen, Erika Oosthuitzen, Paolo Anagnostou, Giovanni Destro-Bisol, Vincenzo L Pascali, and Cristian Capelli. 2017. “Complex Ancient Genetic Structure and Cultural Transitions in Southern African Populations.” *Genetics* 205 (1): 303–16. https://doi.org/10.1534/genetics.116.189209.
Muir, Andrew J., Sanjeev Arora, Gregory Everson, Robert Flisiak, Jacob George, Reem Ghalib, Stuart C. Gordon, et al. 2014. “A Randomized Phase 2b Study of Peginterferon Lambda-1a for the Treatment of Chronic HCV Infection.” *Journal of Hepatology* 61 (6): 1238–46. https://doi.org/10.1016/j.jhep.2014.07.022.
Nielsen, Rasmus, Joshua M. Akey, Mattias Jakobsson, Jonathan K. Pritchard, Sarah Tishkoff, and Eske Willerslev. 2017. “Tracing the Peopling of the World through Genomics.” *Nature* 541 (7637): 302–10. https://doi.org/10.1038/nature21347.
Olivieri, Anna, Alessandro Achilli, Maria Pala, Vincenza Battaglia, Simona Fornarino, Nadia Al-Zahery, Rosaria

Scozzari, et al. 2006. “The MtDNA Legacy of the Levantine Early Upper Palaeolithic in Africa.” *Science (New York, N.Y.)* 314 (5806): 1767–70. https://doi.org/10.1126/science.1135566.
Owers, Katharine A, Per Sjödin, Carina M Schlebusch, Pontus Skoglund, Himla Soodyall, and Mattias Jakobsson. 2017. “Adaptation to Infectious Disease Exposure in Indigenous Southern African Populations.” *Proceedings of the Royal Society B: Biological Sciences* 284 (1852). https://doi.org/10.1098/rspb.2017.0226.
Pagani, Luca, Toomas Kivisild, Ayele Tarekegn, Rosemary Ekong, Chris Plaster, Irene Gallego Romero, Qasim Ayub, et al. 2012. “Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool.” *American Journal of Human Genetics* 91: 83–96. https://doi.org/10.1016/j.ajhg.2012.05.015.
Pagani, Luca, Daniel John Lawson, Evelyn Jagoda, Alexander Mörseburg, Anders Eriksson, Mario Mitt, Florian Clemente, et al. 2016. “Genomic Analyses Inform on Migration Events during the Peopling of Eurasia.” *Nature* 538 (7624): 238–42. https://doi.org/10.1038/nature19792.
Pagani, Luca, Stephan Schiffels, Deepti Gurdasani, Petr Danecek, Aylwyn Scally, Yuan Chen, Yali Xue, et al. 2015. “Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians.” *American Journal of Human Genetics* 96 (6): 986–91. https://doi.org/10.1016/j.ajhg.2015.04.019.
Pakendorf, Brigitte. 2014. “Coevolution of Languages and Genes.” *Current Opinion in Genetics and Development*. Elsevier Ltd. https://doi.org/10.1016/j.gde.2014.07.006.
Patin, Etienne, Marie Lopez, Rebecca Grollemund, Paul Verdu, Christine Harmant, Hélène Quach, Guillaume Laval, et al. 2017. “Dispersals and Genetic Adaptation of Bantu-Speaking Populations in Africa and North America.” *Science* 356 (6337): 543–46. https://doi.org/10.1126/SCIENCE.AAL1988.
Pennings, Pleuni S., and Joachim Hermisson. 2006. “Soft Sweeps III: The Signature of Positive Selection from

Recurrent Mutation." *PLoS Genetics* 2 (12): 1998–2012. https://doi.org/10.1371/journal.pgen.0020186.
Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, D. Absher, B. S. Srinivasan, et al. 2009. "Signals of Recent Positive Selection in a Worldwide Sample of Human Populations." *Genome Research* 19 (5): 826–37. https://doi.org/10.1101/gr.087577.108.
Pickrell, Joseph K, Nick Patterson, Po-Ru Loh, Mark Lipson, Bonnie Berger, Mark Stoneking, Brigitte Pakendorf, and David Reich. 2014. "Ancient West Eurasian Ancestry in Southern and Eastern Africa." *Proceedings of the National Academy of Sciences of the United States of America* 111 (7): 2632–37. https://doi.org/10.1073/pnas.1313787111.
Pierron, Denis, Margit Heiske, Harilanto Razafindrazaka, Veronica Pereda-Loth, Jazmin Sanchez, Omar Alva, Amal Arachiche, et al. 2018. "Strong Selection during the Last Millennium for African Ancestry in the Admixed Population of Madagascar." *Nature Communications* 9 (1): 1–9. https://doi.org/10.1038/s41467-018-03342-5.
Poole, Lisa A., and David Cortez. 2017. "Functions of SMARCAL1, ZRANB3, and HLTF in Maintaining Genome Stability." *Critical Reviews in Biochemistry and Molecular Biology*. Taylor & Francis. https://doi.org/10.1080/10409238.2017.1380597.
Poznik, G. David, Yali Xue, Fernando L. Mendez, Thomas F. Willems, Andrea Massaia, Melissa A. Wilson Sayres, Qasim Ayub, et al. 2016. "Punctuated Bursts in Human Male Demography Inferred from 1,244 Worldwide Y-Chromosome Sequences." *Nature Genetics* 48 (6): 593–99. https://doi.org/10.1038/ng.3559.
"Prevention of Neural Tube Defects: Results of the Medical Research Council Vitamin Study." 1991. *The Lancet* 338 (8760): 131–37. https://doi.org/10.1016/0140-6736(91)90133-A.
Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9. https://doi.org/10.1038/ng1847.
Pritchard, Jonathan K., Joseph K. Pickrell, and Graham

Coop. 2010. “The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation.” *Current Biology* 20 (4): R208–15. https://doi.org/10.1016/j.cub.2009.11.055.
Pritchard, Jonathan K., and Anna Di Rienzo. 2010. “Adaptation - Not by Sweeps Alone.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2880.
Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *The American Journal of Human Genetics* 81 (3): 559–75. https://doi.org/10.1086/519795.
Pybus, Marc, Pierre Luisi, Giovanni Marco Dall’Olio, Manu Uzkudun, Hafid Laayouni, Jaume Bertranpetit, and Johannes Engelken. 2015. “Hierarchical Boosting: A Machine-Learning Framework to Detect and Classify Hard Selective Sweeps in Human Populations.” *Bioinformatics* 31 (24): btv493. https://doi.org/10.1093/bioinformatics/btv493.
Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.
Quintana-Murci, Lluís, Ornella Semino, Hans-J. Bandelt, Giuseppe Passarino, Ken McElreavey, and A. Silvana Santachiara-Benerecetti. 1999. “Genetic Evidence of an Early Exit of Homo Sapiens Sapiens from Africa through Eastern Africa.” *Nature Genetics* 23 (4): 437–41. https://doi.org/10.1038/70550.
Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. 2005. “Support from the Relationship of Genetic and Geographic Distance in Human Populations for a Serial Founder Effect Originating in Africa.” *Proceedings of the National Academy of Sciences* 102 (44): 15942–47. https://doi.org/10.1073/pnas.0507611102.
Robinson, John D., Alec J. Coffman, Michael J. Hickerson, and Ryan N. Gutenkunst. 2014. “Sampling Strategies for Frequency Spectrum-Based Population Genomic Inference.” *BMC Evolutionary Biology* 14 (1).

https://doi.org/10.1186/s12862-014-0254-4.
Ronen, Roy, Nitin Udpa, Eran Halperin, and Vineet Bafna. 2013. “Learning Natural Selection from the Site Frequency Spectrum,” 230–33.
Rotimi, Charles, Akin Abayomi, Alash’le Abimiku, Victoria May Adabayeri, Clement Adebamowo, Ezekiel Adebiyi, Adebowale D. Ademola, et al. 2014. “Research Capacity. Enabling the Genomic Revolution in Africa.” *Science*. American Association for the Advancement of Science. https://doi.org/10.1126/science.1251546.
Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. “Positive Natural Selection in the Human Lineage.” *Science*. https://doi.org/10.1126/science.1124309.
Sabeti, Pardis C., David E. Reich, John M. Higgins, Haninah Z. P. Levine, Daniel J. Richter, Stephen F. Schaffner, Stacey B. Gabriel, et al. 2002. “Detecting Recent Positive Selection in the Human Genome from Haplotype Structure.” *Nature* 419 (6909): 832–37. https://doi.org/10.1038/nature01140.
Sabeti, Pardis C., Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, et al. 2007. “Genome-Wide Detection and Characterization of Positive Selection in Human Populations.” *Nature* 449 (7164): 913–18. https://doi.org/10.1038/nature06250.
Scheinfeldt, Laura B, Sameer Soi, Simon Thompson, Alessia Ranciaro, Dawit Woldemeskel, William Beggs, Charla Lambert, et al. 2012a. “Genetic Adaptation to High Altitude in the Ethiopian Highlands.” *Genome Biology* 13 (1): R1. https://doi.org/10.1186/GB-2012-13-1-R1.
———. 2012b. “Genetic Adaptation to High Altitude in the Ethiopian Highlands.” *Genome Biology* 13 (1): R1. https://doi.org/10.1186/gb-2012-13-1-r1.
Schiffels, Stephan, and Richard Durbin. 2014. “Inferring Human Population Size and Separation History from Multiple Genome Sequences.” *Nature Genetics* 46 (8): 919–25. https://doi.org/10.1038/ng.3015.
Schlebusch, Carina M., and Mattias Jakobsson. 2018. “Tales of Human Migration, Admixture, and Selection in Africa.” *Annual Review of Genomics and Human Genetics* 19

(1): annurev-genom-083117-021759. https://doi.org/10.1146/annurev-genom-083117-021759.
Schlebusch, Carina M., Pontus Skoglund, Per Sjödin, Lucie M. Gattepaille, Michael G.B. Blum, Himla Soodyall, and Mattias Jakobsson. 2012. “Genomic Variation in Seven Khoe-San” 1187 (October): 374–79. https://doi.org/10.1126/science.1227721.
Schlebusch, Carina M, Helena Malmström, Torsten Günther, Per Sjödin, Alexandra Coutinho, Hanna Edlund, Arielle R Munters, et al. 2017. “Southern African Ancient Genomes Estimate Modern Human Divergence to 350,000 to 260,000 Years Ago.” *Science* 358 (6363): 652–55. https://doi.org/10.1126/science.aao6266.
Schrider, Daniel R., and Andrew D. Kern. 2016. “S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning.” *PLoS Genetics* 12 (3). https://doi.org/10.1371/journal.pgen.1005928.
Schuette, Verena, Maria Embgenbroich, Thomas Ulas, Meike Welz, Jonas Schulte-Schrepping, Astrid M. Draffehn, Thomas Quast, et al. 2016. “Mannose Receptor Induces T-Cell Tolerance via Inhibition of CD45 and up-Regulation of CTLA-4.” *Proceedings of the National Academy of Sciences* 113 (38): 10649–54. https://doi.org/10.1073/pnas.1605885113.
Semo, Armando, Magdalena Gayà-Vidal, Cesar Fortes-Lima, Bérénice Alard, Sandra Oliveira, João Almeida, António Prista, et al. 2019. “Mozambican Genetic Variation Provides New Insights into the Bantu Expansion.” *BioRxiv*, July, 697474. https://doi.org/10.1101/697474.
Shanahan, Timothy M., Nicholas P. Mckay, Konrad A. Hughen, Jonathan T. Overpeck, Bette Otto-Bliesner, Clifford W. Heil, John King, Christopher A. Scholz, and John Peck. 2015. “The Time-Transgressive Termination of the African Humid Period.” *Nature Geoscience* 8 (2): 140–44. https://doi.org/10.1038/ngeo2329.
Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Beno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11): 2498–2504. https://doi.org/10.1101/gr.1239303.

Sherman, Rachel M., Juliet Forman, Valentin Antonescu, Daniela Puiu, Michelle Daya, Nicholas Rafaels, Meher Preethi Boorgula, et al. 2019. “Assembly of a Pan-Genome from Deep Sequencing of 910 Humans of African Descent.” *Nature Genetics* 51 (1): 30–35. https://doi.org/10.1038/s41588-018-0273-y.

Shi, Yufang, Mark R. Walter, Sidney Pestka, Devanand Sarkar, Paul B. Fisher, and Christopher D. Krause. 2004. “ I Nterleukin -10 and R Elated C Ytokines and R Eceptors .” *Annual Review of Immunology* 22 (1): 929–79. https://doi.org/10.1146/annurev.immunol.22.012703.104622.

Shriver, Mark D., Giulia C. Kennedy, Esteban J. Parra, Heather A. Lawson, Vibhor Sonpar, Jing Huang, Joshua M. Akey, and Keith W. Jones. 2004. “The Genomic Distribution of Population Substructure in Four Populations Using 8,525 Autosomal SNPs.” *Human Genomics* 1 (4): 274–86. https://doi.org/10.1186/1479-7364-1-4-274.

Skoglund, Pontus, Jessica C. Thompson, Mary E. Prendergast, Alissa Mittnik, Kendra Sirak, Mateja Hajdinjak, Tasneem Salie, et al. 2017. “Reconstructing Prehistoric African Population Structure.” *Cell* 171 (1): 59-71.e21. https://doi.org/10.1016/J.CELL.2017.08.049.

Smith, Eric Alden. 2010. “Communication and Collective Action: Language and the Evolution of Human Cooperation.” *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2010.03.001.

Smith, John Maynard, and John Haigh. 1974. “The Hitch-Hiking Effect of a Favourable Gene.” *Genetical Research* 23 (1): 23–35. https://doi.org/10.1017/S0016672300014634.

Stojanowski, Christopher M. 2014. “Iwo Eleru’s Place among Late Pleistocene and Early Holocene Populations of North and East Africa.” *Journal of Human Evolution*. https://doi.org/10.1016/j.jhevol.2014.02.018.

Storz, J. F., Bret A Payseur, and Michael W Nachman. 2004. “Genome Scans of DNA Variability in Humans Reveal Evidence for Selective Sweeps Outside of Africa.” *Molecular Biology and Evolution* 21 (9): 1800–1811.

https://doi.org/10.1093/molbev/msh192.
Stringer, C. B., R. Grün, H. P. Schwarcz, and P. Goldberg. 1989. “ESR Dates for the Hominid Burial Site of Es Skhul in Israel.” *Nature* 338 (6218): 756–58. https://doi.org/10.1038/338756a0.
Stringer, Chris. 2016. “The Origin and Evolution of Homo Sapiens.” *Philosophical Transactions of the Royal Society B: Biological Sciences*. Royal Society of London. https://doi.org/10.1098/rstb.2015.0237.
Su, Qiang, Feizhuo Hu, Yuxia Liu, Xiaofei Ge, Changlin Mei, Shengqiang Yu, Aiwen Shen, et al. 2018. “Cryo-EM Structure of the Polycystic Kidney Disease-like Channel PKD2L1.” *Nature Communications* 9 (1): 1–12. https://doi.org/10.1038/s41467-018-03606-0.
Tajima, F. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123 (3): 585–95.
Tang, Kun, Kevin R. Thornton, and Mark Stoneking. 2007. “A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome.” *PLoS Biology* 5 (7): 1587–1602. https://doi.org/10.1371/journal.pbio.0050171.
Tishkoff, Sarah A., Floyd A. Reed, Françoise R. Friedlaender, Christopher Ehret, Alessia Ranciaro, Alain Froment, Jibril B. Hirbo, et al. 2009. “The Genetic Structure and History of Africans and African Americans.” *Science* 324 (5930): 1035–44. https://doi.org/10.1126/science.1172257.
Tishkoff, Sarah A, Floyd A Reed, Alessia Ranciaro, Benjamin F Voight, Courtney C Babbitt, Jesse S Silverman, Kweli Powell, et al. 2007. “Convergent Adaptation of Human Lactase Persistence in Africa and Europe.” *Nature Genetics* 39 (1): 31–40. https://doi.org/10.1038/ng1946.
Udpa, Nitin, Roy Ronen, Dan Zhou, Junbin Liang, Tsering Stobdan, Otto Appenzeller, Ye Yin, et al. 2014. “Whole Genome Sequencing of Ethiopian Highlanders Reveals Conserved Hypoxia Tolerance Genes.” *Genome Biology* 15 (2): R36. https://doi.org/10.1186/gb-2014-15-2-r36.
Underhill, Peter A, and Toomas Kivisild. 2007. “Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations.” https://doi.org/10.1146/annurev.genet.41.110306.130407

.
Uren, Caitlin, Minju Kim, Alicia R Martin, Dean Bobo, Christopher R Gignoux, Paul D van Helden, Marlo Möller, Eileen G Hoal, and Brenna M Henn. 2016. “Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries.” *Genetics* 204 (1): 303–14. https://doi.org/10.1534/genetics.116.187369.
Vernot, Benjamin, and Joshua M. Akey. 2014. “Resurrecting Surviving Neandeltal Linages from Modern Human Genomes.” *Science* 343 (February): 1017–21. https://doi.org/10.5061/dryad.5t110.Supplementary.
Vicente, Mário, Mattias Jakobsson, Peter Ebbesen, and Carina M Schlebusch. 2019. “Genetic Affinities among Southern Africa Hunter-Gatherers and the Impact of Admixing Farmer and Herder Populations.” Edited by Evelyne Heyer. *Molecular Biology and Evolution* 36 (9): 1849–61. https://doi.org/10.1093/molbev/msz089.
Visser, Mijke, Robert-Jan Palstra, and Manfred Kayser. 2014. “Human Skin Color Is Influenced by an Intergenic DNA Polymorphism Regulating Transcription of the Nearby BNC2 Pigmentation Gene.” *Human Molecular Genetics* 23 (21): 5750–62. https://doi.org/10.1093/hmg/ddu289.
Vitti, Joseph J, Sharon R Grossman, and Pardis C Sabeti. 2013. “Detecting Natural Selection in Genomic Data.” https://doi.org/10.1146/annurev-genet-111212-133526.
Voight, Benjamin F., Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K. Pritchard. 2006. “A Map of Recent Positive Selection in the Human Genome.” *PLoS Biology* 4 (3): 0446–58. https://doi.org/10.1371/journal.pbio.0040072.
Wack, Andreas, Ewa Terczyńska-Dyla, and Rune Hartmann. 2015. “Guarding the Frontiers: The Biology of Type III Interferons.” *Nature Immunology*. Nature Publishing Group. https://doi.org/10.1038/ni.3212.
Wang, K., M. Li, and H. Hakonarson. 2010. “ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data.” *Nucleic Acids Research* 38 (16): e164–e164. https://doi.org/10.1093/nar/gkq603.
Weir, B S, and C Clark. 1984. “Estimating F-Statistics for the Analysis of Population Structure.” *Cockerham Source: Evolution*. Vol. 38.
Westaway, K. E., J. Louys, R. Due Awe, M. J. Morwood, G. J.

Price, J. X. Zhao, M. Aubert, et al. 2017. “An Early Modern Human Presence in Sumatra 73,000-63,000 Years Ago.” *Nature* 548 (7667): 322–25. https://doi.org/10.1038/nature23452.
White, Tim D., Berhane Asfaw, David DeGusta, Henry Gilbert, Gary D. Richards, Gen Suwa, and F. Clark Howell. 2003. “Pleistocene Homo Sapiens from Middle Awash, Ethiopia.” *Nature* 423 (6941): 742–47. https://doi.org/10.1038/nature01669.
Wilde, Sandra, Adrian Timpson, Karola Kirsanow, Elke Kaiser, Manfred Kayser, Martina Unterländer, Nina Hollfelder, et al. 2014. “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 (13): 4832–37. https://doi.org/10.1073/pnas.1316513111.
Yang, Yongfei, Christine Quach, and Chengyu Liang. 2016. “Autophagy Modulator Plays a Part in UV Protection.” *Autophagy* 12 (9): 1677–78. https://doi.org/10.1080/15548627.2016.1196319.
Yelmen, Burak, Mayukh Mondal, Davide Marnetto, Ajai K Pathak, Francesco Montinaro, Irene Gallego Romero, Toomas Kivisild, Mait Metspalu, and Luca Pagani. 2019. “Ancestry-Specific Analyses Reveal Differential Demographic Histories and Opposite Selective Pressures in Modern South Asian Populations.” Edited by Evelyne Heyer. *Molecular Biology and Evolution* 36 (8): 1628–42. https://doi.org/10.1093/molbev/msz037.
Yi, Xin, Yu Liang, Emilia Huerta-Sanchez, Xin Jin, Zha Xi Ping Cuo, John E. Pool, N. Xu, et al. 2010. “Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude.” *Science* 329 (5987): 75–78. https://doi.org/10.1126/science.1190371.
Zhang, Xing, Xiang Li, Wanjiang Zhang, Liliang Wei, Tingting Jiang, Zhongliang Chen, Chunping Meng, et al. 2013. “The Novel Human MRC1 Gene Polymorphisms Are Associated with Susceptibility to Pulmonary Tuberculosis in Chinese Uygur and Kazak Populations.” *Molecular Biology Reports* 40 (8): 5073–83. https://doi.org/10.1007/s11033-013-2610-7.
Zwaenepoel, I., M. Mustapha, M. Leibovici, E. Verpy, R.

Goodyear, X. Z. Liu, S. Nouaille, et al. 2002. “Otoancorin, an Inner Ear Protein Restricted to the Interface between the Apical Surface of Sensory Epithelia and Their Overlying Acellular Gels, Is Defective in Autosomal Recessive Deafness DFNB22.” *Proceedings of the National Academy of Sciences* 99 (9): 6240–45. https://doi.org/10.1073/pnas.082515999.

APPENDIX

1. List of publications

Dobon, B., Rossell, C., **Walsh, S**. *et al.* Is there adaptation in the human genome for taste perception and phase I biotransformation?. *BMC Evol Biol* **19,** 39 (2019) doi:10.1186/s12862-019-1366-7

Walsh, S., Pagani, L., Xue, Y., *et al.* Positive selection in admixed populations from Ethiopia. *BMC Genet*. 2019, in press.

2. List of manuscripts in preparation

Walsh' S., Atkinson, E., Dobon, B. *et al*. Positive selection analysis of a KhoeSan population.

Walsh' S., Izquierdo-Serra' M., Acosta. S., *et al.* Adaptive selection drives functional lchanges in *TRPP3* in Ethiopian populations.