“Through the layers of the Ethiopian genome: A survey of human genetic variation based on genome-wide genotyping and re-sequencing data”

Luca Pagani
Emmanuel College

This dissertation is submitted for the degree of
Doctor of Philosophy

Division of Biological Anthropology
Department of Archaeology & Anthropology
University of Cambridge

If the desert were 'home',
if our instincts were forged in the desert
to survive the rigours of the desert;
then it is easier to understand
why greener pastures pall on us;
why possessions exhaust us,
and why Pascal's imaginary man
found his comfortable lodgings
a prison."

Bruce Chatwin, *The Songlines*

Table of contents

i. Declaration

This dissertation describes work undertaken under the supervision of Dr Toomas Kivisild at the Division of Biological Anthropology and advices by Dr. Chris Tyler-Smith from The Wellcome Trust Sanger Institute and Dr. Neil Bradman from UCL, in fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Cambridge. It is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. I hereby declare that the work described here has not been submitted for a degree, diploma, or any other qualification at any other university or institution. I also confirm that this thesis does not exceed the word or page limits specified by the Archaeology and Anthropology Degree Committee.

Luca Pagani

Cambridge, January 8th, 2013

ii. Summary

"Through the layers of the Ethiopian Genome: A survey of human genetic variation based on genome-wide genotyping and re-sequencing data"

PhD candidate: Luca Pagani

Understanding our evolutionary history as a species has since long been one of the most attracting and controversial themes of the scientific investigation. From its geographical position, outstanding fossil record and richness of human diversity, the Horn of Africa and, particularly, the Ethiopian region offers an unmatched opportunity to investigate our origins from a genetic perspective. To carry out a genome-wide survey of this region, 13 out of the estimated 80 extant Ethiopian populations were typed on an Illumina Omni 1M SNP array. The results showed a good concordance between genetic and linguistic stratification and, overall, a complex population structure placing the Ethiopians in between North and Sub Saharan Africans, due to the recent non African gene flow which was dated at around 3000 years ago. Furthermore the SNP array data unveiled putative traces of the out of Africa migrations as well as, in two of the typed populations, signatures of genetic adaptation to high altitude.

To obtain an unbiased, high resolution representation of the Ethiopian genetic landscape, 25 individuals from each of five populations were newly collected and sequenced on an Illumina HiSeq platform. These populations were chosen, from among the ones typed on the SNP array, to represent the main components of Ethiopian genetic diversity. Of the 25 samples per population, 24 were sequenced at low depth to generate a broad list of genetic variants, while one sample from each was sequenced at high depth to provide a higher resolution list of variants peculiar to each analysed population. The 125 Ethiopian genomes thus sequenced, while overall consistent with the genotyping results, described the Ethiopian populations in a less biased way than the SNP array data. Furthermore estimation of past effective population size fluctuations from the individual genomes unveiled a unique pattern in the ancestry of the Ethiopian populations in the early stages of human evolution. These results provide a data resource which can be used in future analyses.

iii. List of figures.

(Figures reported as published papers or in appendices are excluded from this list)

iv. List of tables

(Tables reported as published papers or in appendices are excluded from this list)

v. List of abbrev. used in the thesis

bp	base pair
DNA	deoxyribonucleic acid
ds	double strand
F_{ST}	fixation index
ga	generations ago
GWAS	genome wide association study
H	heterozigosity
IBD	identity by descent
iHS	integrated haplotype score
Indel	insertion/deletion
k	thousands
k	number of clusters
LD	linkage disequilibrium
M	millions
m	migration rate
MDS	multidimensional scaling
mtDNA	mitochondrial DNA
n	nano
N_e	effective population size
OOA	out of Africa
PBS	population branch statistic

PCA	principal component analysis
PCR	polymerase chain reaction
π	average number of pairwise differences
PNG	Papua Nuova Guinea
PSMC	pairwise sequentially Markovian coalescent model
QC	quality checks
qPCR	quantitative PCR
r	correlation coefficient
REC	research ethical committee
S	number of mutations
SFS	site frequency spectrum
SNP	single nucleotide polymorphism
SpO_2	percentage of saturated oxygen
STR	short tandem repeat
T	time
TMRCA	time from the most recent common ancestor
µ	micro
UCL	University College London
WGA	whole genome amplification
WTSI	The Wellcome Trust Sanger Institute
ya	years ago

vi. Thesis structure and publications disclosure

This dissertation describes the genesis, scientific outcomes and future directions of the project I have been working on, during three years spent as a PhD candidate at the Division of Biological Anthropology of the University of Cambridge, under the supervision of Dr. Toomas Kivisild and advices by Dr. Chris Tyler-Smith from The Wellcome Trust Sanger Institute and Dr. Neil Bradman from UCL.

The work is structured in four main chapters and six appendices:

Chapter 1: Introduction and thesis rationale: the overall project is placed in the context of the data and methods available to the genetic anthropology field.

Chapter 2: Genotyping results: the results obtained from the first phase of the project as well as additional information on the quality checks applied to the results generated.

Chapter 3: Sequencing results: the collection of further Ethiopian samples and the preliminary results obtained after their whole genome sequencing.

Chapter 4: General Discussion: the main outcomes of my research and the contribution it makes to the field of genetic anthropology.

Since part of Chapter 2 and the appendices have already been published or written up in manuscript form in the framework of international collaborations, the following paragraphs provide a brief description of my contribution to each of them.

Chapter 2

Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D, Bekele E, Bradman N, Balding DJ and Tyler-Smith C (2012). "Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool." Am J Hum Genet **91**(1): 83-96.

This paper includes a considerable part of the results reported in this chapter. My role included running all the analyses, preparing figures and tables, and coordinating the scientific discussion among the co-authors about the results. I also wrote first draft of the manuscript.

Appendix 1

Colonna V, Pagani L, Xue Y and Tyler-Smith C (2011). "A world in a grain of sand: human history from genetic data." Genome Biol **12**(11): 234.

My contributions to this collaborative effort consisted of writing the section devoted to the Jewish Diaspora, gathering all the genomic datasets reviewed in the manuscript and running the analyses displayed in the four figures and supplementary tables. I also contributed to the general discussion and design of the review and I created the above mentioned figures.

Appendix 2

Pagani L and Ayub Q
"The proportion of human DNA within buccal swab extracts before and after whole genome amplification."
My contribution to this unpublished work involved designing and running the laboratory experiments as well as the statistical analysis of the results. I also wrote the manuscript, integrating the advice of Dr. Qasim Ayub from the Wellcome Trust Sanger Institute, Hinxton, UK.

Appendix 3

Huerta-Sánchez* E, DeGiorgio* M, Pagani* L, Tarekegn A, Ekong R, Antao T, Cardona A, Montgomery HE, Cavalleri GL, Robbins PA, Weale ME, Bradman N, Bekele E, Kivisild T, Tyler-Smith* C, Nielsen* C. *equally contributing authors

"Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations"

This work has been submitted to Current Biology and is currently under review. My contribution to this manuscript was primarily sharing the unpublished genotyping data at an early stage. I also took part in the overall discussion and interpretation of the results; and in providing supportive analyses; the draft of the manuscript was written and the core analyses performed by Dr. Huerta-Sanchez and Dr. DeGiorgio.

Appendix 4

"Search for a set of 22 markers that would capture the main patterns of genetic diversity among Ethiopian populations".

This short report describes the work that I carried out to develop a near-optimal list of markers that can be used by other researchers in future projects on Ethiopian populations.

Appendix 5 and 6 These two appendices reports auxiliary data to the sample collection and SNP calling processing, respectively

vii. Acknowledgements

The first thank you goes to my girlfriend Gloria who followed me on my academic plans, without caring for the years of sun and good food left behind. Without her, my time in Cambridge would have been half as happy.

The next people on the thanks list are my PhD supervisor Toomas Kivisild from the University of Cambridge and Chris Tyler-Smith from The Wellcome Trust Sanger Institute. Their guidance and cheerful comments have made me grow and inspired me throughout these years.

A special thank you goes as well to Neil Bradman from UCL. His help and support was vital to obtain the DNA samples described in this thesis. Meeting him has certainly strengthened my collaboration skills.

Outside academia, the people I would like to thank for having shared, at least in part, their life paths with mine cannot fit in a page or two. However these four years spent abroad made a landmark in my life. I would then like to thank those people who were there before all this started and who are still there, waiting for my return. I would also like to thank all those people who were not there but who now are and, I am sure, will be with me also after this thesis will be submitted. Lastly, since this is the only chance I have to do so, I would like to thank all those people I met in Cambridge and whom I will probably never see again.

When I first arrived in Cambridge, the plan was to stay just for a few months. These later became four years. A thank you is hence due to my family, and especially my brother Filippo, my father Ambrogio and my mother Giusi, for having understood that academia means all this and, perhaps, more.

My work would not have been possible without the help and support of all the Ethiopian collaborators and donors, who kindly agreed to delegate the study of their history to foreign people in a foreign country. Thank you.

I could not end this section without acknowledging the funding I received to make my research project come true and to survive throughout it. I would then thank Domestic Research studentship, Cambridge European Trust, Emmanuel College and Melford Charitable Trust for funding me, and The Wellcome Trust for providing funding to generate the data for my project.

1. Introduction and Thesis rationale

1.1 The genomic era

One of the main consequences of a highly developed neural system in humans is the emergence of self-awareness and the need to understand our origins and history. A powerful tool to investigate our past is provided by the genetic information encoded in the DNA of modern populations. The genetic mutations accumulated through the generations since our origins as a species provide the genetic variation in modern human populations. The mode by which these mutations spread across generations can either be random (genetic drift) or constrained by environmental conditions (natural selection). Both these processes contribute to the observed human genetic diversity. Relating the observed genetic diversity with ancestry estimation and known events from the past can inform us about the demographic and adaptive processes that characterised the genetic history of a given population. Traditionally, due to the technological limitations to accessing such information, only a few genomic loci have been used to ask questions about the origin of our species, the demographic processes that led to the present distribution of diversity on the planet and the selection processes that were associated with environmental differences. The three genomic loci that have traditionally been used to address these questions are mitochondrial DNA (mtDNA), the non-recombining portion of the Y chromosome and the HLA region. In addition, other specific genes of medical interest have been studied since the advent of the first sequencing techniques (Sanger et al. 1977) in limited number of samples.

In the past decade, several technological advances have dramatically increased the throughput of generating genetic data at reduced cost. The ready availability of genomic information has allowed the field of biological anthropology to increase the resolution of studying human genetic diversity, both in terms of number of populations sampled and of surveyed genomic markers. In the genomic era, the investigation focus has indeed shifted from a few predetermined loci to a representative, unbiased set of genomic variants shedding new light onto the human evolutionary processes. The three main categories of DNA analysis approaches that stemmed from such advances can be summarised as follows.

1.1.1 SNP arrays

Current micro-bead based technologies allow for simultaneous genotyping of hundreds of thousands to millions of markers from across the genome. This is achieved through fluorescent hybridization of the template DNA with a set of probes anchored to the array platform and, in

order to detect the zygosity of each analysed marker, the two alleles are marked with different colours. However, this kind of approach has one main limitation when applied to the study of worldwide human populations. The set of variants included in most commercially available SNP arrays (such as the ones marketed by Illumina and Affymetrix) have been compiled from those available from previous studies and selected among the available candidates to fit the requirements of genome wide association studies (GWAS). These studies, aimed at finding markers associated with a given medical condition, typically make use of markers that have moderate to high frequencies in the general population and, in order to provide independencies between tests, use only one representative marker from each block of linkage disequilibrium (LD). Therefore the composition of the commercially available SNP arrays is biased toward moderate to high frequency markers and reduced levels of LD. These two factors, together, reduce the power to describe recent demographic processes (flagged by low frequency variants) and selective sweeps (represented by long blocks of LD). Furthermore, for historical reasons the SNP discovery processes has mostly been focused on non-African and, particularly, European populations. As a consequence, the resulting arrays are optimized to detect variability in certain populations. This effect is usually referred to as “ascertainment bias” and its consequences are summarized in Figure 1.1: when the genetic diversity of population 2 (in red) is assessed using genetic markers that were ascertained in population 1 (in blue), a considerable proportion of variants is missed in population 2. To compensate this effect, however, the most recently designed arrays have been enriched for variants specifically typed in previously understudied populations. The Illumina Omni 1M SNP array described in Chapter 2 belongs to this new generation of platforms. In summary, the SNP array technology allows a cost effective, high density genomic analysis based on a pool of markers selected on the basis of previous knowledge of variable sites. The information obtained can be efficiently used to describe the overall population structure and patterns of admixture as well as performing some selection tests (described in section 1.6). However, due to the biases described above, a comparison of the genetic diversity between different worldwide populations, as well as some selection tests that rely on the full frequency spectrum, can only be performed on sequencing data.

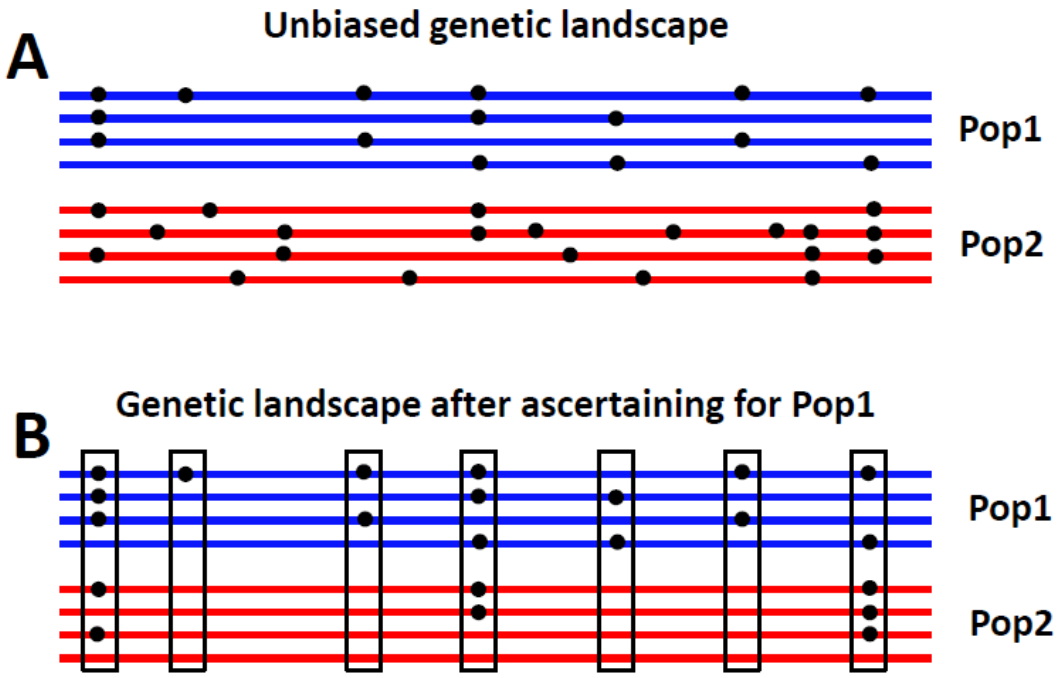

Figure 1.1 Effect of the ascertainment bias documented for most of the commercially available SNP array platform. Each line is a chromosome, each dot is a variant. The usage on population 2 of a set of markers optimized to detect the genetic diversity of population 1 (**B**) underestimates the higher diversity of population 2 (**A**).

1.1.2 Mass spectrometry

A complementary strategy to SNP array typing, which allows a cost-effective reading of a few tens of markers on hundreds of samples is represented by DNA mass spectrometry, as applied in the Sequenom platform (Jurinke et al. 2002) . This approach is often used as an orthogonal way of validation of variants discovered through sequencing, but can also be used as a cost-efficient way of typing small numbers of relevant markers, such as sets of ancestry informative markers (AIMs), in several human populations. The work described in Appendix 4 is intended to test whether, and with what efficiency, an optimized set of a few tens of markers can replicate the patterns of inter-population diversity detected among Ethiopian populations while using high density (Illumina 1M Omni) genome-wide data. An alternative technology for typing on this scale in the absence of a mass spectrometer is multiplex PCR followed by single base extension of the variable position, for example in the ABI PRISM SNaPshot reaction.

1.1.3 Whole genome re-sequencing

The last and perhaps most innovative approach described in this section, among the ones brought by the so called “genomics revolution”, is the opportunity to sequence in a quick and cost-effective manner the genomes of hundreds or thousands of samples. This has been made possible by two main factors. Firstly, the availability of a reference sequence for many species, including humans, allowed the development of next-generation “shotgun” re-sequencing technologies. The short sequences (reads) generated by these shotgun technologies are mapped to the reference sequence which acts as a genomic scaffold facilitating the assembly of the tested genome. The second and equally fundamental factor is the explosive development of low-cost, high-throughput, sequencing technologies. Of the many sequencing methods developed during the various stages of the genomic revolution, the Illumina (Bentley 2006) strategy is the most widely used at the time when this thesis is being written. The principle behind this sequencing approach is quite similar to the one adopted by the latest versions of the classic Sanger sequencing (Sanger et al. 1977). The DNA sequence is read when the complementary strand is sequenced *in vitro.* Each added nucleotide has a fluorescent marker that stimulates a light detector only if the nucleotide is successfully added to the nascent strand. In order to unambiguously detect each added base, these modified nucleotides carry a terminator on the deoxy group. However, in contrast to the Sanger method, this terminator can be removed through chemical washing, hence allowing the sequencing reaction to proceed in a controlled stepwise manner. This innovation allows the reading of multiple bases per DNA fragment, creating reads between 30 and 200 bp depending on the used sequencing platform. The pool of reads thus generated is then aligned to the reference sequence, of which perhaps 85-95% is accessible to this approach, depending on the read length and additional factors such as whether or not the reads are paired. The superimposition of multiple reads on the same genomic regions (sequencing depth, or “coverage”) allows an increased confidence in interpreting the resulting sequence. Furthermore, the abundance of reads in each genomic region can be used to infer the zygosity of a given variant. Since the sequencing costs increase linearly with the sequencing depth achieved, it is important to estimate the minimum depth required for a given experiment. In particular, it has been shown that while an average depth of at least 30x (30 reads per each genomic position) is required to call heterozygous sites in a single sample, comparable results can be achieved for the shared variants with as little as 6x if many samples are processed and the variant sites are called together (The 1000 Genomes Project Consortium 2010). While the SNP calling efficiency is around 99% when a high

sequencing depth (30x) is available, the pooling of many low depth samples together can achieve an efficiency of at least 96% with a fivefold reduction of the overall costs (the missing variants are mostly accounted for by singletons, variants that are observed only once in the overall sample) (The 1000 Genomes Project Consortium et al. 2012).

In summary, the full genome re-sequencing, allowed by platforms such as the Illumina HiSeq, provides a cost-effective way to detect the genetic variability in human populations in an almost unbiased way. However, since the costs of re-sequencing at moderate or high depth are still one order of magnitude higher than the SNP array costs, the latter still remains a valuable approach to provide an explorative survey of the genetic diversity from a large number of samples. For the Ethiopian samples described below, a mixed approach was chosen. To inform the choice of the populations to be re-sequenced, a broad set of samples was firstly typed on a SNP array platform. Following this preliminary survey, it was decided to sequence, for each population of interest, one high depth (30x) and 24 low depth (8x) samples. The rationale of this sequencing strategy was to reduce the sequencing costs for the global set while keeping the benefits of analyzing one well-ascertained genome per population. Further details on the sequencing strategy are provided in section1.4.

1.2 . Applications to the field of Biological Anthropology

The deluge of genome-wide genotype and sequence data generated by international consortia (Frazer et al. 2007; Li et al. 2008; The 1000 Genomes Project Consortium et al. 2012) has facilitated, in the last few years, the study from a genetic perspective of classical anthropological questions such as the origin of our species and the main demographic and migratory events that generated the current worldwide human diversity.

1.2.1 The classical genetic contribution to the main themes of biological anthropology

Genetics, since its introduction into the biological anthropology field, has been crucial to shedding light on the scientific debate about the origin of our species. Such debate initially opposed a multiregional (Thorne and Wolpoff 1992) to an African origin (Stringer and Andrews 1988) of *Homo sapiens* (Figure 1.2). These two competing models interpreted the geographical overlap between *Homo erectus* (an ancestral hominin originating about 1.9 million years ago) and *Homo sapiens* fossils within and outside Africa in two different ways. According to the multiregional model (Figure 1.2, left), modern humans evolved from *Homo erectus* in multiple geographical regions and, by means of continuous gene flow between the various groups,

managed to maintain a genetic uniformity. The African origin model (Figure 1.2, right) proposed instead a single African origin of all modern humans, and a recent colonization of the non-African continents with replacement of all archaic hominins.

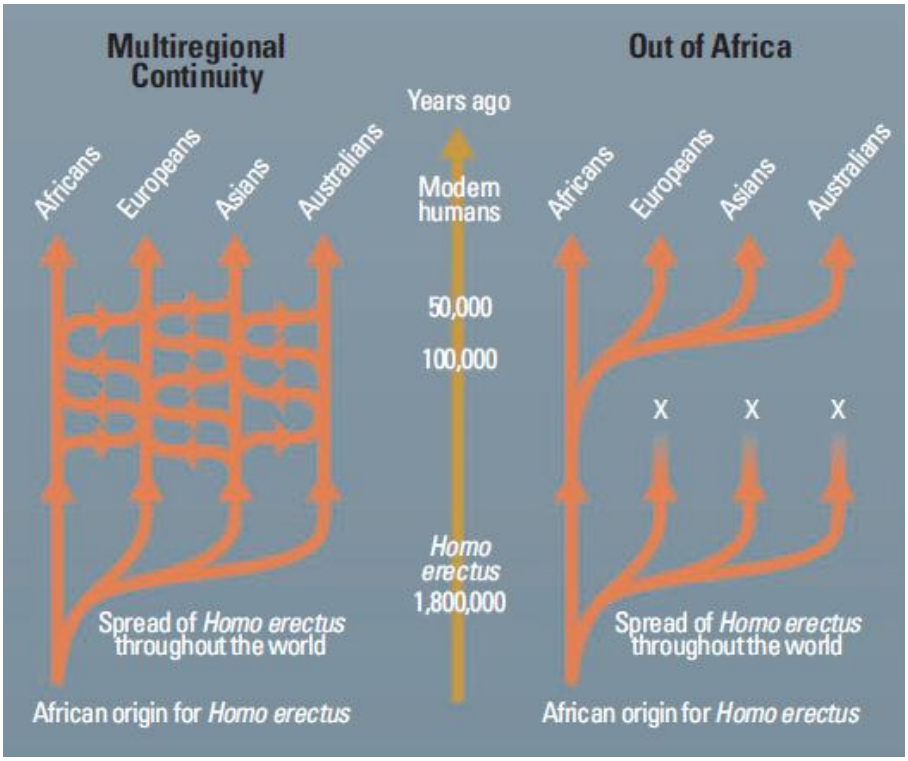

Figure 1.2 Multiregional or African origin of modern humans. The multiregional model proposed by Thorne and Wolpoff in 1992 (left panel) contemplates multiple transitions from Homo erectus to Homo sapiens in multiple geographical regions and continuous gene flow between them. The single African origin model (right panel), introduced by Stinger and Andrews in 1988, proposes instead a unique origin of all modern humans in Africa and their subsequent migration out of Africa. This figure was adapted from (Gibbons 2011).

Both models fit well with the fossil and paleo-climatic records alone. However, the highest genetic diversity observed in Africa and the gradual decline with the geographic distance from that continent (Figure 1.3) could only be explained with a single African origin and subsequent migration out of Africa. While the exact timing (Gravel et al. 2011; Soares et al. 2011) and routes (Lahr and Foley 1994; Campbell and Tishkoff 2008) of such migrations are still debated, the genetic data was important in the debate on the early evolutionary history of our species.

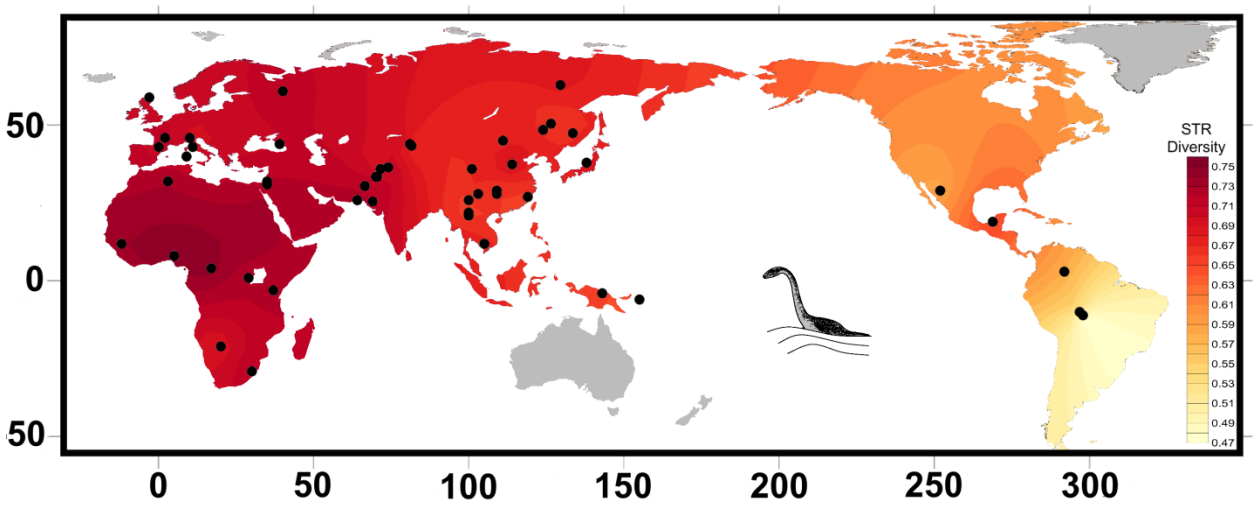

Figure 1.3 Pattern of genetic diversity in worldwide human populations. The distribution of STR diversity in worldwide human populations, adapted from the literature (Colonna et al. 2011), shows a higher diversity in African populations and a decline with the distance from Africa (each black dot represents a sampled population). The observed pattern fits with the proposed single African origin with subsequent migrations out of Africa proposed by Stringer and Andrews in 1988.

The genetic dating of any evolutionary event is ultimately based on the assumption of molecular clock (Pauling 1964). A molecular clock implies that mutations introduced during each meiotic event are randomly distributed across the genome and occur with a rate that is, on average, constant. The mutation rate can be calibrated in several ways, for example by dividing the number of fixed genetic differences between two species by the time since their speciation, inferred from the fossil record. The number of mutations observed between any two human samples can therefore be used to estimate the time since the divergence from their most recent common ancestor. The dating and geographic patterns of human evolutionary events have traditionally been inferred from two main genomic loci: the non-recombining portion of the Y chromosome and the mitochondrial genome (mtDNA). These loci have two main characteristics in common: they do not recombine, hence they are transmitted as whole units across generations, and they are inherited exclusively from one parental lineage (the paternal and the maternal, respectively). These two peculiarities made them widely used in early evolutionary genetic studies, thanks to the relative ease in tracing and dating phylogenetic trees based on these two loci. However, the direct dating and inference of demographic processes have recently been supplemented by simulation-based methods. These computer-based approaches generate simulated sequencing data by running a user-defined evolutionary model. The artificially-generated data are compared with the empirical data, and the degree of similarity

between the two sets provides information about the goodness of the particular evolutionary model. Another strategy to estimate the time from the most recent ancestor (TMRCA) between multiple sequences at a population level, consists in dividing the genome into sets of SNPs in high linkage disequilibrium (LD blocks). Each block is therefore considered as a single recombination unit and the haplotypes observed in a given population as originating from a single, ancestral sequence. The number of observed mutations divided by the total length of all the considered sequences, multiplied by the mutation rate can therefore yield and estimate of the TMRCA of that sequence for a given population. Due to the high noise implied in this method, the TMRCA of a population is usually inferred by the distribution of each sequence TMRCA.

The observed decrease of genetic diversity with the increase in distance from the African continent (Figure 1.3) has led to the formulation of one crucial assumption on the mode of human dispersals. The isolation by distance model, firstly introduced by Wright (Wright 1946), was indeed further developed in light of the available genetic data (Prugnolle et al. 2005; Liu et al. 2006) and used to explain the observed neutral diversity outside Africa. In summary, the molecular clock and isolation by distance models applied to the mtDNA, Y chromosome and a few other genomic loci formed the basis for the early interpretation of the worldwide human genetic diversity. These early studies started shedding light on the African origin of our species and the demographic processes that led to the colonization of the other continents, and formed the basis for future studies of the patterns of admixture between human populations and the natural selection events acting upon them.

1.2.2 Recent genomic contributions

The increase of available genomic data from several human populations (The 1000 Genomes Project Consortium et al. 2012) as well as from extinct hominids (Green et al. 2010; Reich et al. 2010) has allowed researchers to further describe the demographic processes that shaped the observed genetic diversity worldwide. Recent studies have indeed started to unveil the dynamics of important landmarks of human evolutionary history. In particular, the early African origin model proposed by Stringer and colleagues has been supported and refined in the light of recent genetic data. The patterns of linkage disequilibrium (LD) observed across the genomes of multiple African populations were used, as proxy for long term effective population size (N_e), to infer a putative geographic place of origin of human diversity. The analyses performed by Henn and colleagues (Henn et al. 2011) showed that the populations currently inhabiting the

South West African region might have preserved the biggest N_e over time, and suggested the region they inhabit as the putative source of the human genetic diversity investigated (Figure 1.4)

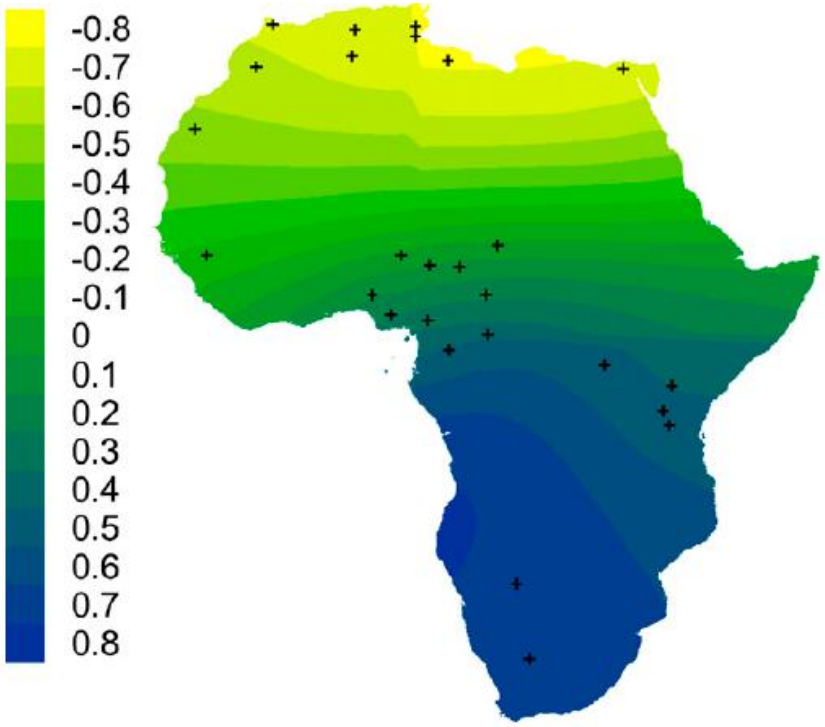

Figure 1.4 Putative South-West African origin of human diversity. Linkage disequilibrium (LD) based analysis identified the populations currently inhabiting the South West portion of the African continent as the ones showing the biggest N_e of all the human populations. The blue-yellow gradient shows the correlation coefficient (r) between the LD pattern and distance from the putative geographical origin of the human diversity. The highest correlation values are obtained when the putative geographical origin was set in South West Africa. Each cross is a sampled population. This figure was adapted from (Henn et al. 2011).

The availability of genomic data from multiple loci increased the reliability of simulation analyses. These analyses (Gravel et al. 2011) allowed modelling of the demographic parameters of the migrations out of Africa and of subsequent population splits (Figure 1.5).

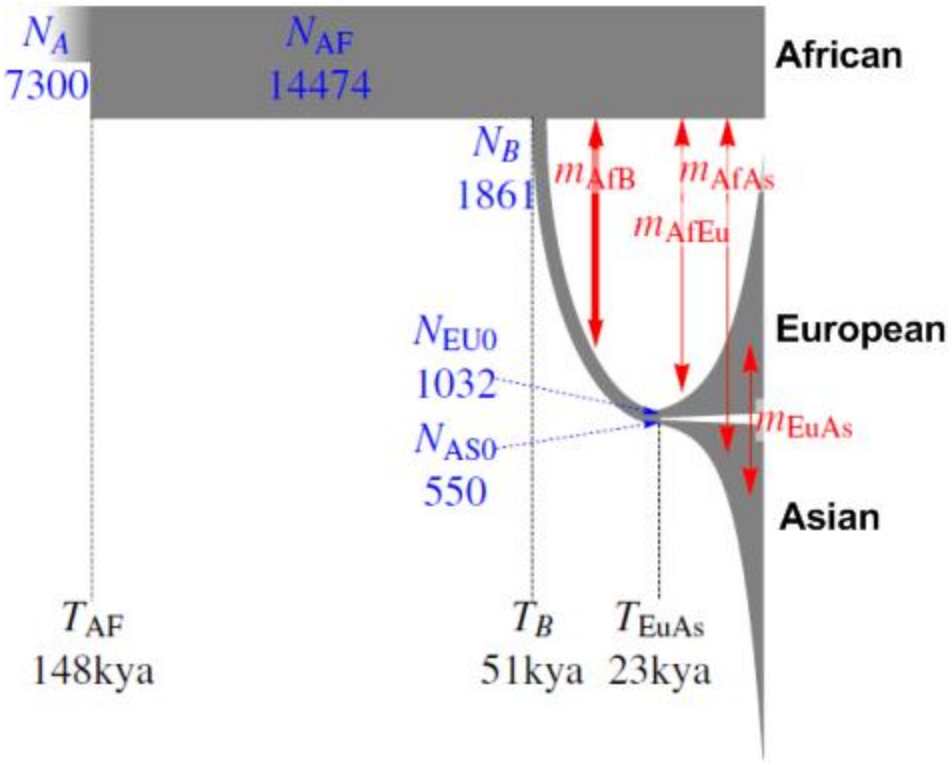

Figure 1.5 Demographic model describing the origin of worldwide human populations. Simulation analyses based on the observed genetic data from three African, European and Asian populations were used to estimate the effective population sizes of the ancestral (N_A) and modern African populations (N_{AF}) as well as the size of the out-of-Africa bottleneck (N_B) and of the resulting European (N_{EU0}) and Asian (N_{AS0}) populations. Furthermore, the split times (T) associated to the above-mentioned population differentiations, as well as the migration rates (m) between those groups, were calculated. The figure was adapted from (Gravel et al. 2011).

The estimation of demographic parameters from extant genetic diversity has been further expanded by the production of high-resolution genomic data. Methods relying on the genetic data from single genomes such as the Pairwise Sequentially Markovian Coalescent model (PSMC) (Li and Durbin 2011), described in section 1.6.3.1, produced a representation of the changes in effective population size (N_e) over time for individuals belonging to African and non-African populations (Figure 1.6). The curves represented in Figure 1.6 show, for the non-African populations: a reduction of N_e at 60 kya consistent with the out of Africa bottleneck. PSMC also unveiled an ancient increase of N_e around 100-200 kya shared by the ancestors of all the human populations studied. This increase predated the out-of-Africa migration and was interpreted by the authors as the effect of putative ancient population stratification within Africa.

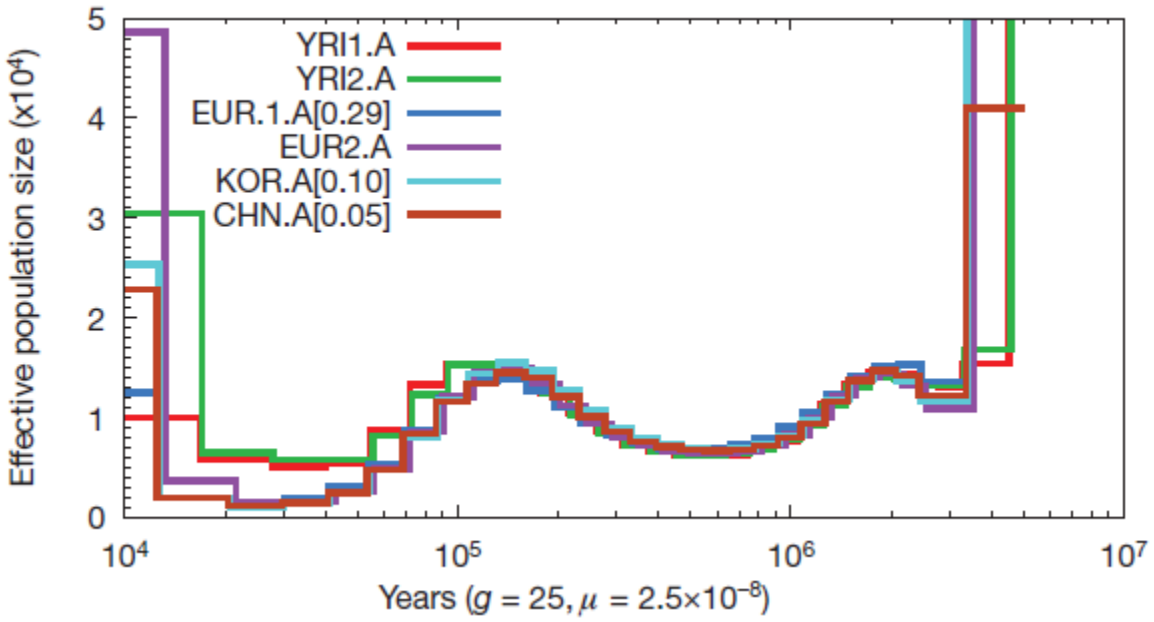

Figure 1.6 PSMC analysis of high resolution genomes from African and non African individuals. The changes of effective population size (N_e) over time (years) in two African (YRI1.A and YRI2.A), two European (EUR.1.A and EUR2.A) and two Asian (KOR.A and CHN.A) samples were analysed applying the Pairwise Sequentially Markovian Coalescent model (PSMC) on high resolution genomic data. The N_e reduction around 60kya confirmed the previously described out-of-Africa population bottleneck, while the increase of N_e around 200kya unveiled previously unknown African population stratification. This figure was adapted from (Li and Durbin 2011).

In addition to the genomes of modern human individuals, the newly developed technologies have allowed the extraction and sequencing of DNA from ancient human (Rasmussen et al. 2010; Keller et al. 2012) and archaic hominin (Green et al. 2010; Reich et al. 2010) specimens. The first whole-genome sequencing of two extinct hominins, Neanderthals (Green et al. 2010) and Denisovans (Reich et al. 2010), discovered a statistically significant excess of allele sharing between these archaic genomes and those of non-African humans, compared with modern African populations. Several models have been proposed to explain the observed data (Figure 1.7). However, the models that best fit the observed data are the ones implying one or multiple admixture events between ancestors of the modern non-Africans and the populations of archaic hominids (Figure 1.7, model 3) (Reich et al. 2010), and the one proposing an ancient African population structure (Figure 1.7, model 4) (Eriksson and Manica 2012). Although the topic is still highly debated, the ancient population structure invoked by Eriksson and Manica to explain the allelic share between human and archaic hominins was observed in empirical data by the PSMC analysis described in Figure 1.6.

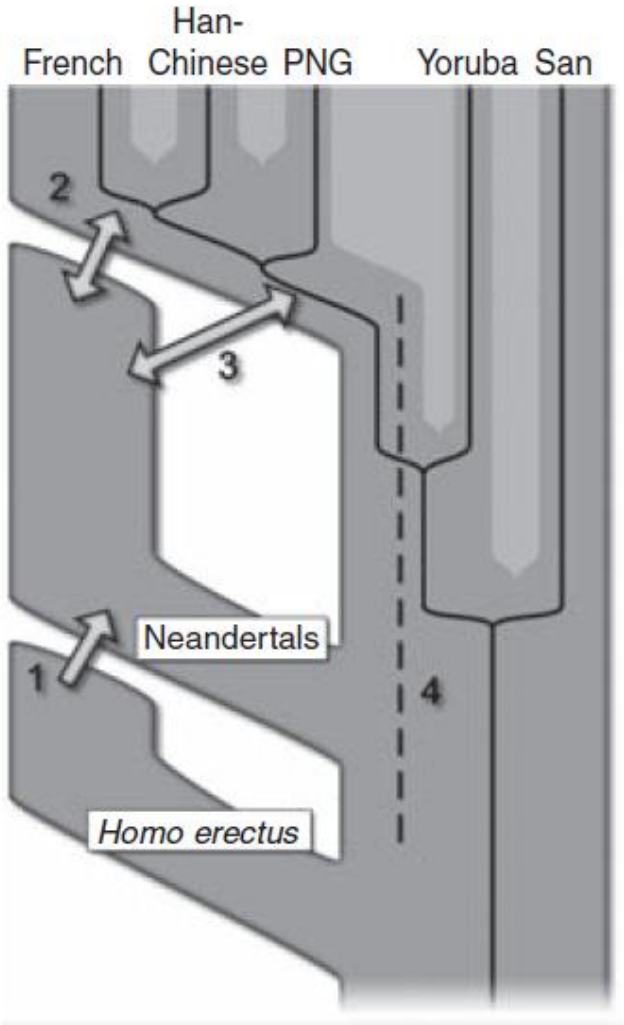

Figure 1.7 Alternative models to explain the excess of allele sharing between Neanderthals and non-African populations. From the genetic and fossil records, several models can be formulated to describe the relationships between modern African (Yoruba and San) and non-African (French, Han-Chinese, Papua New Guinean-PNG) human populations with their extinct relatives. Model 1 proposes ancient gene flow between Neanderthals and the *Homo erectus* encountered after their migration out of Africa. Admixture events might have also occurred between Neanderthals and European populations (2) or between Neanderthals and the ancestors of all the non-African populations (3). An alternative explanation of the observed allele sharing between Neanderthals and non-Africans could involve ancient African population stratification (4). This figure was adapted from (Green et al. 2010).

The genomic approach applied to the anthropology field also yielded a number of successful case studies, some of which are reviewed in the paper reported in Appendix 1. From that review it is clear how powerful a genetic approach can be in informing us about our past as a species, starting from samples as small as a bundle of hair, or using the information available from

modern populations to reconstruct their demographic history. In particular, the recent success in describing the Jewish Diaspora from a genetic perspective (Behar et al. 2010), the relationship between the first Greenland settlers and the modern human populations in that region (Rasmussen et al. 2010) and how human-associated species such as hair lice (Toups et al. 2011) can inform us about our evolutionary past were reviewed there. However, this work also emphasized the lack of sampling from several world regions that are believed to be crucial to the understanding of our evolutionary history. Among these, the most striking gap of knowledge concerned the populations living in the Horn of Africa. This lack of data stands in contrast with the high interest that the area has attracted from archaeological and physical anthropology perspectives.

1.3 Ethno-linguistic background of the Horn of Africa

The reasoning behind the need for expanding the knowledge of the genetic landscape of the Horn of Africa can be best understood by first looking at the broader African picture. Since the publication of the first archaeological and genetic evidence of a recent African origin of all human populations (Stringer and Andrews 1988), the continent has been key to understanding our origins as a species. In particular, the genetic diversity observed in Sub-Saharan African groups, shown to be higher than in other human populations (Campbell and Tishkoff 2008), has been brought out as compelling evidence of an African human origin. The most intriguing questions about the human evolutionary path are concerned with the location within Africa of this origin and the dynamics of the emergence of the anatomically modern humans, as well as the development of their complex behaviour, and the migrations that led them to colonize the continents outside Africa. Together with the Lake Turkana region, the Horn of Africa shows abundant early fossil and archaeological evidence of *Homo* ancestors dating back to 3-4 million years ago, and it is the site of the earliest anatomically modern human fossils, dated to ~200,000 years ago (White et al. 2003; McDougall et al. 2005; Campbell and Tishkoff 2008). The ancient biological presence of humans in the area is further supplemented by archaeological and historical evidence of diversity in material culture, and richness of ethno-linguistic groups living in the area today (Kaplan 1971; Hansberry 1974; Levine 1974; Pankhurst 1998; Phillipson 1998), (Figure 1.8 A). Furthermore, the Horn of Africa has a pivotal role in the two hypothesized out-of-Africa (OOA) routes via present day Yemen (southern route) (Cavalli-Sforza et al. 1994; Lahr and Foley 1994) or Sinai (northern route) (Campbell and Tishkoff 2008). This evidence, combined, quite intriguingly suggests the possibility that the extant populations of the Horn of Africa may still retain unique genetic elements descended from the first human

populations and, in particular, from the ancestral populations that gave rise to the Out of Africa migration.

Ethiopia, the broadest and most populous country in the region, includes ~80 ethnic groups (Levine 1974). The languages spoken fall into two main linguistic families (Nilotic and Afro-Asiatic further subdivided into Semitic, Cushitic and Omotic), emphasizing, with their richness, the diversity and complex history of the area (Blench 2006). Furthermore, the presence in the area of three out of the six branches of the Afro-Asiatic family, have led to speculations about its local origin (Ehret 1995; Kitchen et al. 2009). Events of admixture with surrounding, non-African populations during the last few millennia have been documented from a genetic perspective using both mtDNA (Passarino et al. 1998; Kivisild et al. 2004; Poloni et al. 2009) and the Y chromosome (Passarino et al. 1998; Lucotte and Smets 1999; Semino et al. 2002; Lovell et al. 2005) and from a linguistic point of view (Kitchen et al. 2009)(Figure 1.8 B). After correcting for these recent confounding events, analyses of Ethiopian genetic diversity could provide a chance to shed light to our early evolutionary history. In addition the genetic characteristics of the pre-OOA populations, together with the opportunity to clarify the histories of adaptation to a very diverse environment could be retrieved by studying the Ethiopian populations. Remarkably, despite a few studies that have already produced genomic data for a limited set of Ethiopian populations, (Tishkoff et al. 2009; Behar et al. 2010; Alkorta-Aranburu et al. 2012; Scheinfeldt et al. 2012), none of them have focused on the demographic history and human diversity in the area.

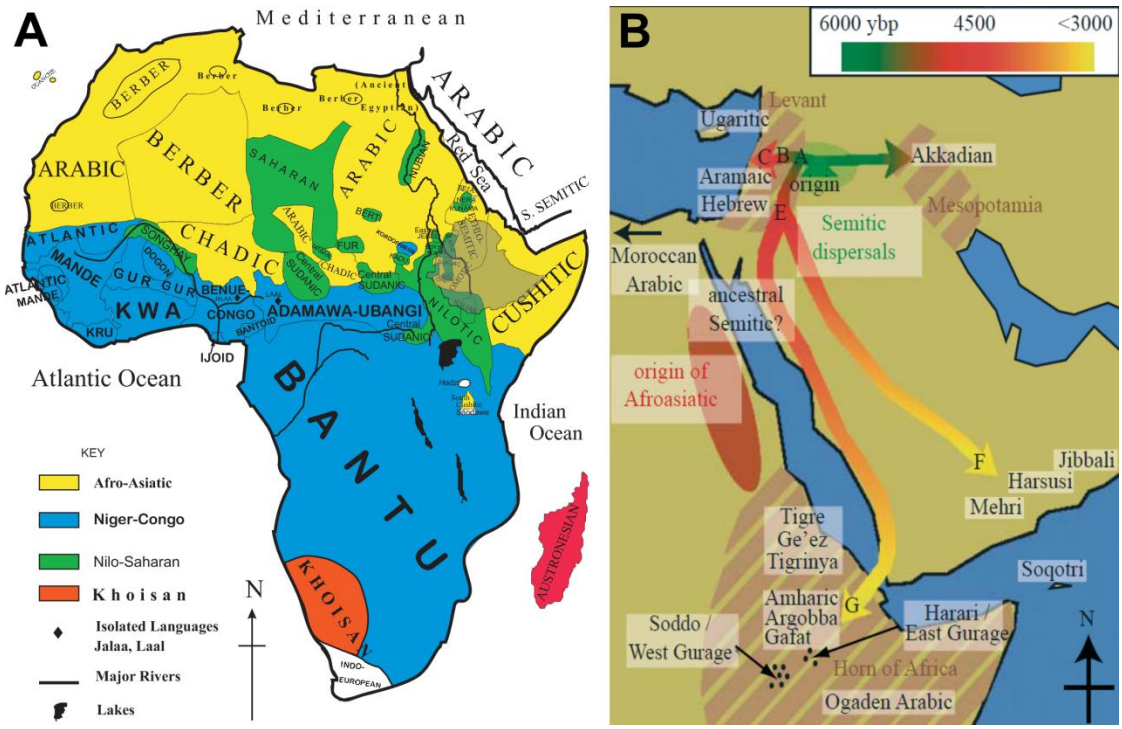

Figure 1.8 Ethiopian linguistic diversity. Ethiopia, from a linguistic point of view, represents one of the most diverse regions in Africa, with the presence of three Afro-Asiatic (Semitic, Omotic, and Cushitic) and several Nilotic linguistic families (**A,** *adapted from Mallam Dendo cartographic service*). Furthermore the presence of Semitic languages in the area can be seen as the result of historical migrations from the Levant (**B,** *adapted from Kitchen et al. 2009*).

1.4 Thesis rationale

The rationale for this thesis stems from the relevance to our evolutionary history of the Horn of Africa region, described in the previous section The main aim of this thesis is to characterize the genetic diversity and population structure in Ethiopia in the context of other African and non-African populations. Three main categories of questions can be asked when dealing with an area as diverse and as stratified as Ethiopia. Before addressing questions about older events in the demographic history of Ethiopian populations, it is important to understand the extent to which recent migrations and population introgressions have shaped the Ethiopian genetic landscape. Demographic processes that have occurred in the last few thousands of years, if not controlled for, can entirely alter the interpretation of analyses aimed at describing the long-term genetic history of the area. The second class of questions concerns the relationship of the current inhabitants of the region with the people who migrated out of Africa around 60 kya.

Given its geographic location, it is indeed sensible to assume a possible role of the area in the processes that led us as a species to spread out of Africa and colonize the rest of the continents. The third and last question one could ask when dealing with Ethiopia is, given the unique chronological series of fossil records available from the Horn of Africa, whether the people currently living in the area still preserve genetic signatures of the evolutionary processes that might have shaped our species in the area, particularly an higher genetic diversity and long term effective population size and shorter linkage disequilibrium blocks when compared with the surrounding populations.

To address these three categories of questions it was decided to first assess Ethiopian genetic diversity through a SNP array survey on a broad set of populations, and then to analyse the whole genome sequences of a subset of these.

Because of the huge cultural and historical diversities that characterize the Ethiopian populations, the first problem when dealing with these populations is that the overall genetic diversity of the area could be split between deeply stratified groups. Therefore, the available socio-cultural evidence from the literature (Kaplan 1971; Hansberry 1974; Levine 1974; Pankhurst 1998; Phillipson 1998; Freeman and Pankhurst 2003; Blench 2006) had to be taken into account to include as much information as possible of this diversity in the analysis. To keep the total number of samples analysed within feasibility constraints, still allowing for at least 20 individuals to be included from each group to provide enough power for downstream analyses, it was decided to select 24 individuals each from around 10 ethnic groups. These samples were selected from the 6000 buccal swab extracts from around 60 Ethiopian populations made available by the collaboration with Dr. Neil Bradman from University College London.

To first assess the overall Ethiopian genetic diversity it was decided to type, on an Illumina Omni 1M SNP array, 24 samples from each of the chosen populations (a description of these groups in provided in the next section). This array was chosen from the ones available because of its enrichment in African variants (The 1000 Genomes Project Consortium 2010) which might reduce the otherwise well-documented ascertainment bias towards non-African populations. Furthermore, the choice of such platform was consistent with the need to generate data compatible with the next generations of SNP arrays in the Illumina Omni series.

Based on the preliminary results from the genotyping phase, a sample collection campaign was organized to obtain more samples with sufficient DNA quantity from the five populations that best summarized the observed genetic diversity. These freshly collected blood samples were

needed to generate 24 low depth (8x) and 1 high depth (30x) whole genome sequences from each of the populations for a total of 125 Ethiopian genome sequences. These two levels of analyses were designed to be complementary. In fact, although the SNP array genotypes are thought to provide a representative picture of the genomic variability and clarify which populations are more recently admixed and which ones may still be preserving traces of ancient haplotypes, whole genome sequencing of selected populations is required to discover new variants of all classes, including structural rearrangements, and provide a less biased understanding of genetic variation. In a way, these results would make it possible to quantitatively assess the extent of ascertainment biases in tag-SNP based methods as applied on populations of high genetic diversity - potentially higher than anywhere else in the world.

1.5 The choice of Ethiopian populations for the study

Two main criteria were used to select the Ethiopian populations for the SNP array analyses: their geographic affiliation and representation of all the spoken linguistic groups. Geography is one of the main drivers of genetic diversity within and among human populations (Prugnolle et al. 2005; Handley et al. 2007)**,** while the presence of four deeply split, mutually unintelligible linguistic groups in Ethiopia (Semitic, Cushitic, Omotic and Nilotic) could reinforce a deep genetic stratification between them. In addition to these two criteria, the Amhara and Oromo populations were included in the analysis because, together, they make up 60% of the total census size of the country. The Ari Blacksmiths were included as another Omotic speaking population along with Ari Cultivators because of their supposedly former hunter-gatherer life-style (Freeman and Pankhurst 2003). Finally, two populations from little-studied neighbouring regions, South Sudan and Somalia, were included in the analyses. The final list of 13 Ethiopian and 2 surrounding chosen populations included: Semitic speaking Amhara, Gurage and Tygray; Omotic speaking Ari Cultivators, Ari Blacksmiths, Dorze and Wolayta, Cushitic speaking Afar, Agew, Ethiopian Somali, Oromo, and Somali from Somalia; Nilotic speaking Anuak, Gumuz and South Sudanese. A map showing the geographical location and the linguistic affiliation of each population is provided in Figure 1.9. Further details of the socio-cultural features of the sampled populations, (Pagani et al. 2012) are reported in Table 1.1. All the samples included in the analyses came from healthy adult donors who were recruited by Dr. Neil Bradman from UCL in Ethiopian villages traditionally inhabited by each chosen populations (Figure 1.9). In order to take part to the study, the donors were asked to report their ethnicity along with the ones of their parents and grandparents. Only donors whose ethnicity matched the ones of all their relatives

were processed. Furthermore, at the time of collection, each donor was asked to confirm to be unrelated with any of the other donors from a given village.

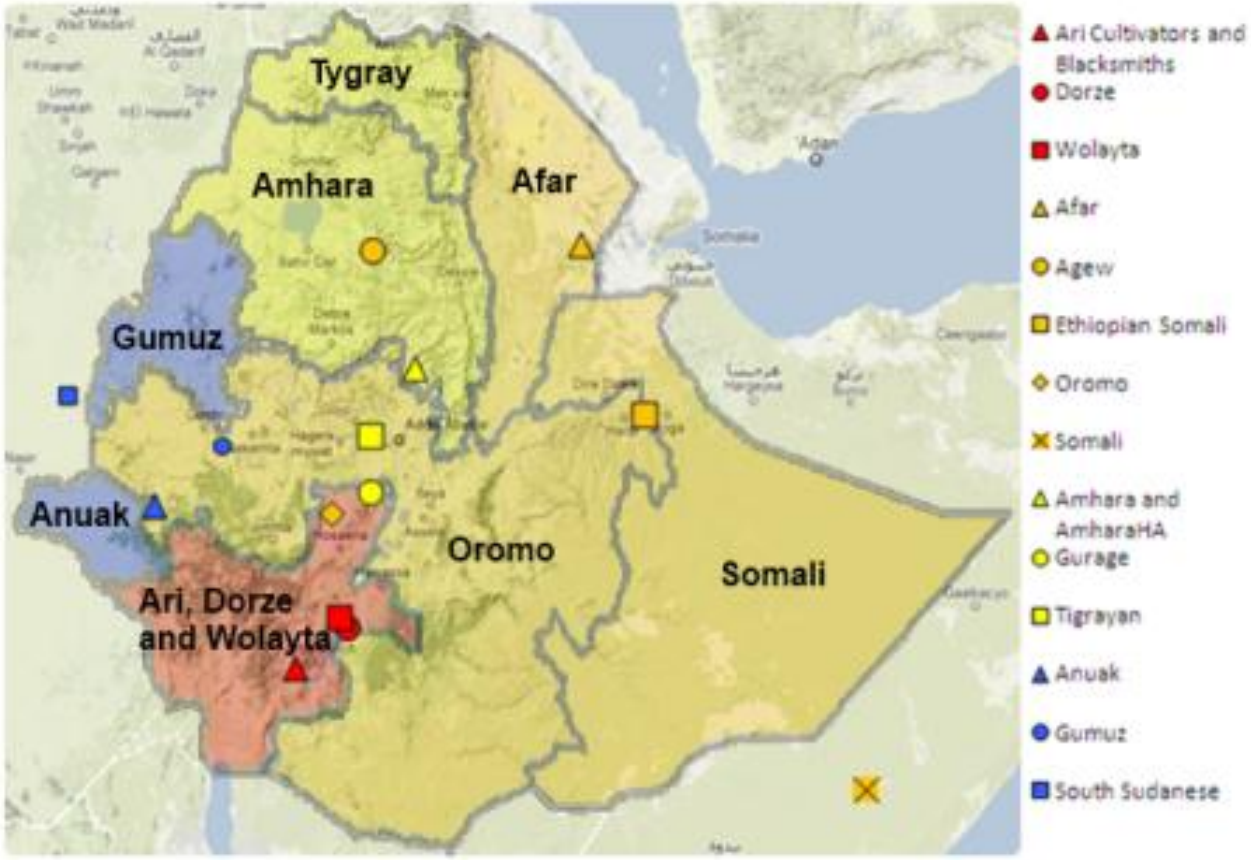

Figure 1.9 Geographical distribution of the Ethiopian populations included in the analysis. The populations are colour coded according to the linguistic families: Omotic in red, Cushitic in orange, Semitic in yellow and Nilotic in blue. This colour coding will be consistent throughout the thesis. The approximated area inhabited by the major groups are outlined and coloured according to the languages spoken.

Table 1.1 Sample size, location and sociological features of the genotyped populations *(adapted from Pagani et al. 2012).*

Pop	N	Lat	Long	Elev	Geo. Location	Ling. Group	Language	1998 Census	Endo/Exogamous	Patri/ Matrilocal	Mono/ Polygamous	Patri/ Matrilineal	High/ Lowland	Food Production
Agew	24	11	38	2063	Wag Hemra Zone	Cushitic	Xamtan	143369	Endo	Patrilocal	Poly	Patrilineal	Highland	Agriculturalist
Afar	24	12	41	379	Afar Region	Cushitic	Afar	979367	Endo	Patrilocal	Poly	Patrilineal	Lowland	Pastoralist
Amhara	26	10	39	2088	Amhara Region	Semitic	Amharic	17372913	Endo	Patrilocal	Mono	Patrilineal	Highland	Agriculturalist
Anuak	24	8	34	500	Gambella	Nilotic	Anuak	45646	Endo	Patrilocal	Poly	Patrilineal	Lowland	Mixed Farming
Ari Blacksmith	24	6	37	1348	South Omo	Omotic	Ari	158857	Endo	Patrilocal	Poly	Patrilineal	Highland	Agriculturalist
Ari Cultivator	24	6	37	1348	South Omo	Omotic	Ari	158857	Endo	Patrilocal	Poly	Patrilineal	Highland	Agriculturalist
Dorze	24	6	38	2779	Gamo Gofa Zone	Omotic	Dorze	20782	Endo	Patrilocal	Poly	Patrilineal	Highland	Agriculturalist
Ethiopian Somali	24	9	42	1543	Somali Region	Cushitic	Somali	3334113	Endo	Patrilocal	Poly	Patrilineal	Lowland	Pastoralist
Gumuz	24	NA	NA	NA	Beni-Shangul Gumuz	Nilo-Saharan	Gumuz	120424	Endo	Patrilocal	Poly	Patrilineal	Lowland	Pastoralist
Gurage	24	8	38	2048	Gurage Zone	Semitic	Gurage	NA	Endo	Patrilocal	Mono	Patrilineal	Highland	Agriculturalist
Oromo	24	8	37	1758	Oromia Region	Cushitic	Oromo	Ca. 17000000	Endo	Patrilocal	Mono/Poly	Patrilineal	Highland	Agriculturalist/Mixed Farming/Pastoralist
Somali	24	NA	NA	NA	NA	Cushitic	Somali	NA	NA	NA	NA	NA	NA	NA
South Sudanese	24	NA	NA	NA	NA	Nilotic	NA	NA	NA	NA	NA	NA	NA	NA
Tygray	24	9	38	1696	Tigray Region	Semitic	Tigrayan	3224875	Endo	Patrilocal	Mono	Patrilineal	Highland	Agriculturalist
Wolayta	24	6	37	1737	Wolayta Zone	Omotic	Wolayta	1231673	Endo	Patrilocal	Mono	Patrilineal	Highland	Agriculturalist

1.6 Methods commonly used in evolutionary genetics to interpret sequence and genotype data

In order to make biological sense of the growing amount of genomic data available for worldwide human populations, several approaches have been developed. These approaches allow researchers to explore the demographic as well as selection events that characterized the evolutionary history of a given population, in addition to the classical summary statistics used in population genetics to measure genetic diversity, such as pairwise difference π, number of segregating sites *S*, heterozygosity *H* and fixation index F_{ST}. The methods used in this thesis can be divided into three main groups: methods to reduce and summarise the complexity of the observed data; methods to detect signatures of selection; methods to detect and study demographic processes.

1.6.1 Methods to reduce the complexity of the observed data.

1.6.1.1 Multi-Dimensional Scaling (MDS) and the Principal Component Analysis (PCA)

The genetic information obtained from multiple loci in multiple individuals can be structured in a (*m x n)* matrix where each row *m* is a sampled individual and each column *n* is a typed locus. In order to reduce the complexity of such matrix, especially when hundreds of thousands of markers are available, a series of methods called Multi-Dimensional Scaling (MDS) (Tzeng et al. 2008) can be applied. Of these, the most widely used for dealing with genomic data is Principal Component Analysis (PCA) (Price et al. 2006). PCA is designed to compute a number of vectors (eigenvectors) of length *m,* each representing a portion of the matrix information. These vectors are computed by first summarizing the information contained in the matrix vectors which are not linearly independent. The eigenvectors are then weighted and ranked according to the proportion of variance they contributed to the original matrix. As a result the first eigenvector describes the axis of greatest variance, the second eigenvector describes the axis orthogonal to the first one, that explains the most variance and so on. The highest ranking vectors can be used to scatter the analysed samples in a reduced dimensional space. Although not representative of the total information, the plotting of the first few such vectors can be an effective way of exploring putative stratification in the data.

1.6.1.2 STRUCTURE-like methods

In contrast to PCA, where no stratification of individual samples from a population is imposed a priori, a complementary set of methods can analyse the data explicitly looking for patterns of admixture within each sample. In particular, each row *m* of the input matrix is interpreted as the

sum of *k* ancestral components. The *k* (an integer number) is defined a priori and, when applied to a set of samples, can be seen as the number of ancestral populations that is needed to explain the observed genetic structure in a given population The genetic composition of each sample is compared to the one of the $a_1, a_2, \ldots, a_k$ inferred ancestral populations and described as a linear combination of them.. Following the publication of STRUCTURE (Pritchard et al. 2000), two other *STRUCTURE-like* methods have been developed: FRAPPE(Tang et al. 2005) and ADMIXTURE (Alexander et al. 2009). With these methods, each individual sample, with no prior information on its population of origin, is calculated to have a composition of $p_1, p_2, \ldots, p_k$ ancestral components. One limitation of this set of methods is the arbitrary choice of the parameter *k.* Although approaches to estimate the optimal *k* exist, prior information on the historical and cultural background of the studied samples can potentially bias the researchers toward the choice of one *k* over another in cases of complex population histories and limited numbers of populations sampled. One further limitation of STRUCTURE-like methods is that, while they are designed to measure the proportion of ancestral components in each individual, they do not provide information about the genomic coordinates that have been derived from one or another specific ancestry component. Therefore other methods have to be used to extract the positional information about any particular ancestry component.

1.6.1.3 Chromosome painting methods

To improve the understanding of the admixture patterns and to assign genomic coordinates to the ancestry components of each studied sample, several approaches have been developed, amongst which the most widely used are Hapmix (Price et al. 2009), fineSTRUCTURE (Lawson et al. 2012), PCAdmix (Brisbin et al. 2012) and Saber (Johnson et al. 2011). The general principle of all of these methods is to compare each test genome (recipient) to a panel of phased chromosomes from populations (donors) acting as proxies of the putative ancestral sources. The output is, per locus, a continuous or discrete probability that a chromosome segment belongs to one of the assumed ancestral populations. While these methods generally perform well on simulated data, the main limitation in empirical studies is the availability or the knowledge of the ancestral populations. The closer the donor haplotypes are to the actual sources, the more accurate is the outcome of the so called “chromosome painting” of the recipient sample. Depending on the assumptions made by each method, some, like SABER, are particularly suited to estimate the number of ancestry switches in each haploid genome (which is a good estimator of the number of generations since the admixture event) while others, like fineSTRUCTURE, are better at tracing the specific contribution of each donor sample to the

genetic make-up of the recipient samples. The strategy behind PCAdmix instead relies on the same principle as PCA. Both recipient and donor haplotypes are subdivided into chromosome segments or windows, each containing an equal number of SNPs. All detected haplotypes from a given window are then processed through PCA and the recipient haplotype is assigned to one of the presumed donor populations, based on its distance from the clusters representing the ancestral populations. A similar approach, inspired by the work of Henn and colleagues (Henn et al. 2011), was applied to the Ethiopian genotype data before the PCAdmix package became available (Brisbin et al. 2012), as described in Chapter 2.

1.6.2 Methods commonly used to detect signatures of selection

Understanding natural selection and its way of shaping the genetic diversity in human populations has long been a key interest in genetics and biological anthropology. The mechanisms through which human populations have adapted to diverse environmental niches can indeed inform both about our history as a species and the biological pathways underlying a certain physiological function. Some of the genes involved in the skin pigmentation, lactase persistence, pathogen resistance and high altitude adaptation have been discovered through classical genetic markers and confirmed by selection scans on worldwide human populations (Voight et al. 2006; Xue et al. 2006; Beall 2007; Beall et al. 2010; Yi et al. 2010; Alkorta-Aranburu et al. 2012). Furthermore, understanding of the mechanisms that regulate medically-relevant phenotypes, such as pathogen resistance or adaptation to hypoxia, can be a powerful tool in applied research. As a consequence, a number of methods to assess deviation from neutrality have been devised in the past years. Many of these methods share a common strategy, since they aim at finding traces of natural selection by looking for one or more of the signatures it might have left on the genetic pool of the studied population. These signatures can be briefly summarized as: extreme differentiation, when the allele under selective pressure rises to frequencies much higher than in the surrounding populations; extended homozygosity, when the haplotype containing the allele under selection experiences a sort of genetic hitchhiking and spreads rapidly through the population under selection; changes in the local site frequency spectrum, due to the reduced genetic diversity caused by the extended homozygosity, and increased number of rare variants due to the mutations occurred after the selection event, on a reduced diversity region. Three methods that have been applied to the data generated within this thesis are described below.

1.6.2.1 Fixation Index (F_{ST}) and Population Branch Statistic (PBS)

The fixation index or F_{ST} (Weir and Cockerham 1984) is probably the most widely used measure of genetic distance between populations. The classic formulation of this statistic can be described as $1-(H_S/H_T)$, where H_S is the average heterozygosity of each population and H_T is the heterozygosity of the total sample. A F_{ST} equal or near 0 is therefore obtained when the genetic distance between two populations is null and the diversity is all shared within both populations, while the opposite is true when the F_{ST} values are close to 1. When this measure of differentiation is applied to each genomic locus of a pair of populations, the loci falling in the top percentiles of the genomic distribution can be explored further for their biological meaning. Indeed the two processes that might bring a genomic locus to F_{ST} values that fall within the top percentiles of the genomic distribution can be either random fluctuations in the allelic frequency (genetic drift) or natural selection acting on a specific allele. To summarize the biological information contained in the F_{ST} outlier loci, a valuable approach involves the analysis of gene enrichment (Huang da et al. 2009) of specific gene categories. The overrepresentation of genes involved in a biological pathway, after correction for multiple testing, can suggest a selection process affecting that specific pathway. However, since the high F_{ST} values can be equally caused by changes in allele frequency in either of the two tested populations, it is problematic, with no a priori hypothesis, to identify which population could have putatively undergone natural selection.

To pinpoint the population branch where the selection event might have occurred, a three population test has been developed by Yi and colleagues (Yi et al. 2010). This population branch statistic (PBS) (Figure 1.10) takes advantage of the unambiguous tree that can be built from the pairwise genetic distances of three populations *a, b, c* and be used, thus, to compute the branch length specific to *a*. The genetic distances used to build the tree are defined, for each (*a,b*), (*a,c*) and (*b,c*) pair as $T(x,y) = -\log(1-F_{ST}(x,y))$, and the length of the branch specific to *a* is computed as $PBS(a) = (T(a,b)+T(a,c)-T(b,c))/2$. Since the genetic drift that has occurred on the *a* branch is not symmetrical when observed from the *b* or the *c* populations, it is important to include populations that are as closely related as possible to the test population to avoid statistical artefacts. Furthermore, as with many rank tests, when the top PBS signals are examined, there is no previous knowledge of the actual amount of natural selection experienced by the population taken in exam.

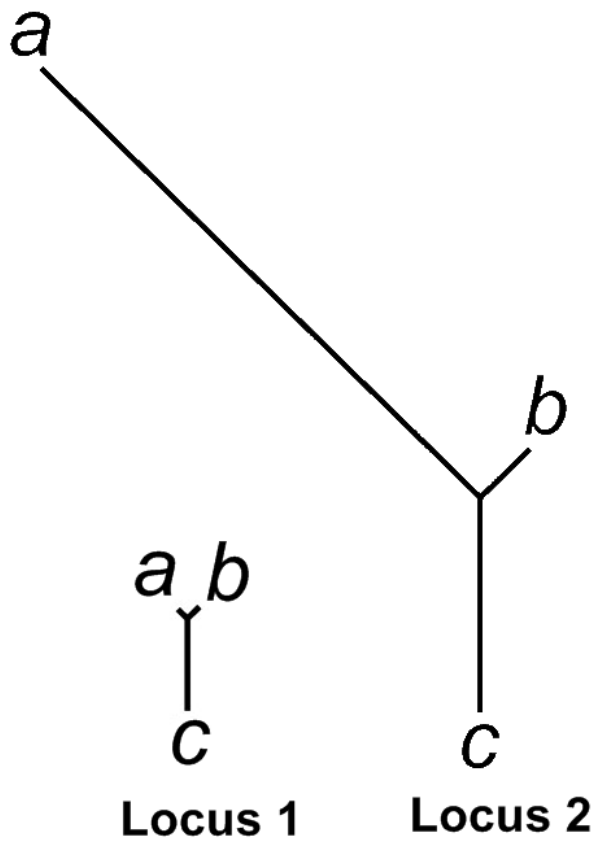

Figure 1.10 Population branch statistic (PBS). The difference in length of the *a* branch when calculated for different loci can imply population-specific processes such as genetic drift or natural selection. The three-population test allows detecting which positions show unusually high differentiation on any given branch of the population tree. This figure was adapted from (Yi et al. 2010).

1.6.2.2 iHS

As opposed to the single locus statistics described above, another set of methods instead takes into account the haplotype landscape surrounding the allele under selection. These methods search for genomic regions where the haplotype diversity is reduced or the length of the haplotypes is longer than the genomic average. Both these effects can be a consequence of a selective process which, by favouring one allele, increases the presence of the surrounding genomic region in the population hence creating a longer, more frequent, haplotype in that region. One of these methods, the integrated haplotype score (iHS) (Voight et al. 2006; Pickrell et al. 2009), takes into account the linkage disequilibrium blocks around each marker, to compute a score based on the length of each haplotype containing the allele under examination. The derived or ancestral allele showing outlier values of iHS at a given locus will be then taken

into consideration for gene enrichment or other follow-up analysis to look for their biological relevance.

1.6.3 Methods to infer demographic processes from individual diploid genomes.

1.6.3.1 PSMC

Following the generation of high-depth whole-genome sequences of individuals belonging to several human populations, new statistics to take advantage of this high resolution information from a limited number of samples have been explored. Among these, the Pairwise Sequentially Markovian Coalescent model (PSMC) developed by Li and Durbin (Li and Durbin 2011) considers each heterozygous site in a single genome as the contribution of two haploid genomes. The density of heterozygous sites in each genomic region can therefore be used to infer the time from the most recent common ancestor (TMRCA) between the two observed haploid genomes. The frequency of coalescent events within a particular time window is then used to infer the effective population size (N_e) over time and to inform about the population dynamics experienced by the ancestors of the sample considered. The effective population size can be seen as the minimum number of individuals needed to explain the observed (or inferred) genetic diversity, and is a good indicator of demographic processes. Since the effective population size changes more slowly than the census size, a high N_e can inform about the-long term genetic diversity of a population. In contrast, a small N_e can reveal past population bottlenecks which, like in the case of the non-African human groups, would otherwise be unsuspected by simply looking at the modern census size.

2. Genome-wide genotyping results

In order to provide a broad basis for characterizing the Ethiopian genetic diversity, 24 samples from each of the 15 Ethiopian and neighbouring populations described in section 1.5 were submitted for genotyping on an Illumina Omni 1M SNP array, and the results are described in this chapter. My personal contribution to the results described in this chapter was to design and execute all the experiments described in section 2.1 and appendix 2 and 4, and to design and perform all the analyses reported in the published paper enclosed to this chapter as well as writing its first draft. The Ethiopian samples employed here were selected by me from a broader set collected before the beginning of this project by Dr. Neil Bradman and collaborators at UCL.The DNA quantification, SNPchip genotyping and SNP calling quality checks were performed by the staff of the genotyping team of the Wellcome Trust Sanger Institute.

2.1 DNA quality assessment

Before starting any analysis on the more than 6000 samples from 60 Ethiopian populations available from Dr. Neil Bradman's collection at UCL, the feasibility of applying genotyping or sequencing approaches to these had to be tested. In order to assess the total amount, concentration and fragmentation of the DNA available from buccal swab extracts, and the potential presence of exogenous DNA from the oral flora, a preliminary study was carried out. The results of this study are reported in detail in Appendix 2. Furthermore this study aimed at assessing the consequences of whole genome amplification (WGA) on the relative proportion of human and exogenous DNA in the buccal swab extracts. The main outcome of the study was that the DNA extracted from the six tested buccal swabs was generally fragmented and in low quantity (less than 1µg per extract). Furthermore the results indicated that the proportion of exogenous DNA ranged between 12% and 76% after whole genome amplification (WGA). The WGA was indeed needed to ensure that at least 700ng of DNA were available, as normally required as input amount for the Illumina beadchip arrays. These two factors taken together suggest that whole genome sequencing of DNA from buccal swabs with the currently available technologies is not practical. However, the results encouraged us to test the feasibility of SNP array genotyping, using as input three times the amount required for pure human DNA, to correct for the high proportion of total exogenous DNA after WGA.

2.2 Pilot genotyping project

To estimate the potential genotyping success rate of the available buccal swab samples on an Illumina Omni 1M SNP array, 15 Amhara and 15 Oromo were processed by carrying out whole genome amplification using a three-fold increase in the input DNA amount, as suggested by the study described in Section 2.1.

The 30 samples submitted for genotyping following these steps yielded encouraging results, with an overall good call rate and only two samples falling just below the 92% acceptance threshold (Figure 2.1).

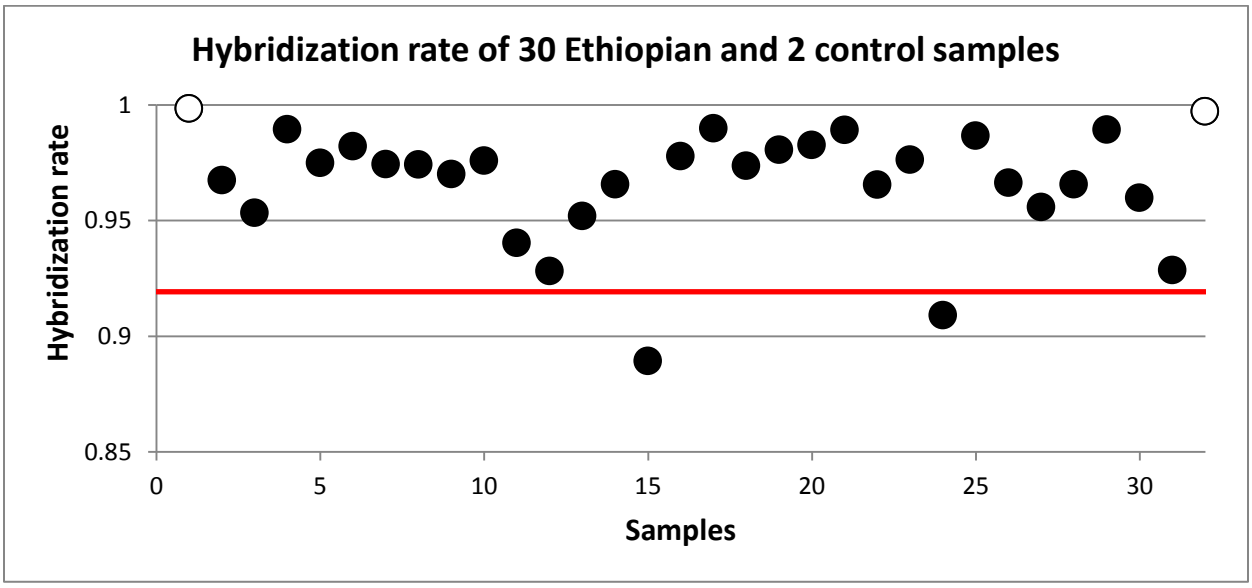

Figure 2.1 SNP array hybridization rates of 30 Ethiopian and 2 control samples. For each of the 30 Ethiopian buccal swab extracts (black dots) and two control cell line extracts (white dots), the proportion of successfully hybridized probes was assessed, as a measure of successful typing on the Illumina Omni 1M SNP array. The red line shows the default acceptance threshold (92%) recommended by the manufacturer. Overall, 28 out of 30 Ethiopian buccal swabs were successfully processed on the SNP array.

The presence of exogenous DNA in the buccal swabs and the DNA mutations potentially inserted by the WGA process suggested the need for further quality checks to be performed on the data obtained. To assess the biological relevance of the data, the 28 Amhara and Oromo samples that were successfully genotyped were analysed together with a set of African (including previously available genotype data for Amhara and Oromo (Behar et al. 2010)) and Middle Eastern control populations (Li et al. 2008; Behar et al. 2010) by a Multidimensional

Scaling analysis (MDS). The MDS was run using the packages available in PLINK (Purcell et al. 2007) and the axes of variation re-estimated from the available data, after pruning out SNPs with LD (r2) greater than 0.1 (Alexander et al. 2009).The new samples clustered with the other Ethiopians available from the literature and lay in an intermediate position between African and non-African populations, consistent with the historically-documented contacts of the Ethiopians with nearby Middle Eastern populations (Figure 2.2).

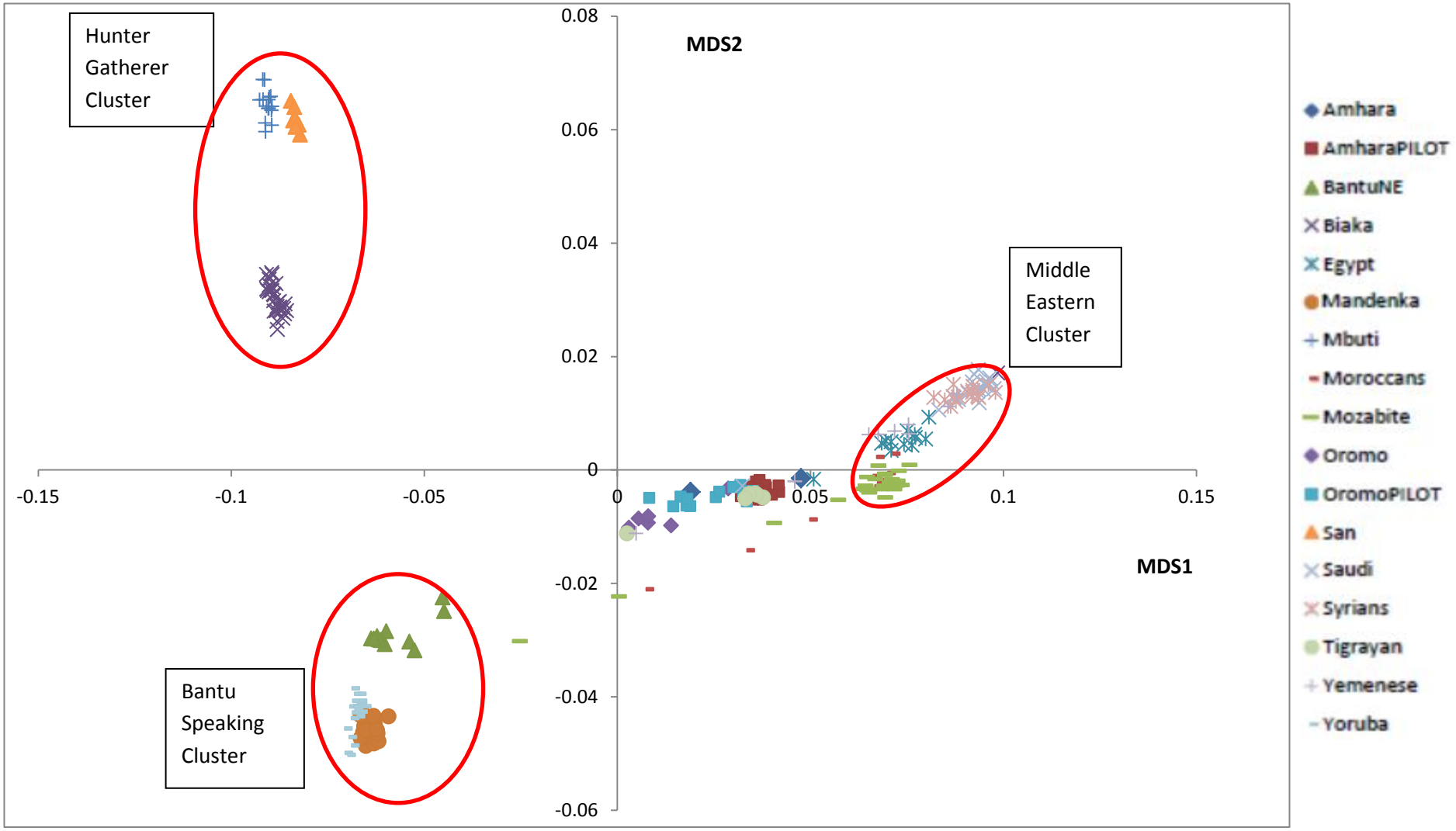

Figure 2.2. MDS plot based on pruned data from whole genome genotypes. As expected, all the pilot Ethiopian samples clustered with the other Ethiopian samples (Amhara, Oromo, Tygray) available from the literature and between the Hunter-Gatherer, the Bantu speaking and the Middle Eastern clusters. This provides a further quality check for the samples used in the pilot.

2.3 Results of the Illumina Omni 1M data analyses

Following the outcome of the analyses of the 30 pilot samples, the remaining nine Oromo and 11 Amhara samples, and 24 samples from each of the remaining 12 Ethiopian and two control populations, were submitted for genotyping. In addition, one Yoruba sample, NA19239, previously typed in Hapmap {Frazer, 2007 #2297} and sequenced by the 1000 Genomes Project was processed to assess the goodness of genotyping. Comparisons of the genotypes obtained for this sample after WGA and typing on the Illumina 1M Omni chip, and the data available for the same sample from Hapmap3 yielded a 99.75% match, 0,25% het mismatch and 0% hom-hom mismatch. This reassuring comparison shows that both the WGA and genotyping steps did not introduce any bias in the SNP calling process.

The genotyping was successful for 306 out of the 386 submitted samples (Table 2.1), although some populations were affected more than others by the QC failures, probably due to the DNA quality or extraction yields. Due to the small number of samples successfully genotyped, Dorze were removed from the analysis at this stage

The Y chromosome haplogroups inferred from the SNP array, and the haplogroups independently determined at UCL for the same samples, were compared to provide an additional quality control. Thanks to this check it was possible to see that the Gurage and Ari Blacksmiths batches of samples had been swopped at some stage of the sample transfer (Table 2.1 reports the correct population labels). Furthermore, potential cross-mixing between the re-labelled Gurage and the Agew was observed in these analyses and led to the removal of both the Agew and the Gurage from subsequent analyses. To keep the sample size consistent across the populations studied, only the AmharaHA were kept for further analyses and renamed just Amhara. The successfully genotyped samples were then pooled with a set of African and non-African samples available from the literature (Frazer et al. 2007; Li et al. 2008; Behar et al. 2010; Henn et al. 2011) and analysed for their demographic history and signatures of selection.

Table 2.1 Number of submitted and successfully genotyped samples for each population.

Population	Linguistic group	N submitted	N successfully genotyped	Success rate
Afar	Cushitic	24	12	50%
Agew	Cushitic	24	24	100%
Amhara	Semitic	24	21	88%
AmharaHA	Semitic	26	26	100%
Anuak	Nilotic	24	23	96%
Ari Blacksmiths	Omotic	24	17	71%
Ari Cultivators	Omotic	24	24	100%
Dorze	Omotic	24	2	8%
Ethiopian Somali	Cushitic	24	17	71%
Gumuz	Nilotic	24	19	79%
Gurage	Semitic	24	24	100%
Oromo	Cushitic	24	21	88%
Somali	Cushitic	24	23	96%
South Sudanese	Nilotic	24	24	100%
Tygray	Semitic	24	21	88%
Wolayta	Omotic	24	8	33%
TOTAL	-	386	306	79%

The details of the results of this study are reported at the end of this chapter in the form of a published paper and can be summarised here in three main points. First, the Semitic- and Cushitic-speaking Ethiopian populations have experienced gene flow from non-African populations, which accounts for as much as 50% of their genomes. This gene flow can be dated by methods that explore LD decay in admixed populations to around 3 kya, consistent with the date of a relevant split in the Semitic languages according to linguistic evidence (Kitchen et al. 2009) and with events reported in the Bible, the Quran and the Kebra Nagast (a traditional, epical account of the early Ethiopian history). Second, while the diversity and composition of mtDNA haplogroups in Ethiopian populations are compatible with Ethiopia being a source of the out-of-Africa mtDNA migration (Soares et al. 2011), the African component of Ethiopian

autosomes, when extracted with chromosome painting methods, is not closer to non-Africans than the African component of the Egyptians. However, this perhaps surprising point needs further investigation, due to the scarcity of data available from Egyptian and other North African populations. Thirdly, the analyses aimed at assessing which African populations show the lowest genome-wide LD placed the Ethiopians in an intermediate position in the South West-North East diversity cline described by Henn and colleagues, suggesting that they were not the best candidate for the source of modern human diversity.

2.4 High altitude adaptation in Ethiopia

The populations inhabiting the Ethiopian plateau, together with the ones from the Himalayan and Andean mountain ranges, are the three best characterized cases of high altitude adaptation from a physiological point of view (Beall 2002). Furthermore, recent genetic studies (Alkorta-Aranburu et al. 2012; Scheinfeldt et al. 2012) have started unveiling a potential genetic component to the high altitude adaptation of some Ethiopian highlanders. The Semitic Tygray and Amhara samples submitted for genotyping were selected, from among those available, to ensure that their birthplaces, as well as the ones of their parents, were at an altitude of 2000 meters or more above sea level. This choice allowed the subsequent setting-up of a study where these two Ethiopian highlander populations where compared with the lowlander Cushitic Afar and Nilotic Anuak, looking for traces of adaptation to high altitude. The other Amhara samples available from the project could not be used as lowlander controls, because the reported place of birth for those individuals was still above 1500 meters. To take into account the possible effect of the non-African gene flow in Tygray and Amhara, the observed allelic frequencies were corrected to adjust for the proportion of non-African admixture. Furthermore, the use of control populations closely related to the Ethiopian highlanders enabled a more confident interpretation of the results than previously achieved when using West African controls (Scheinfeldt et al. 2012). The main contribution stemming from the work reported in Appendix 3 is the detection of new candidate genes for high altitude adaptation in a genomic region surrounding *BHLHE41* (*DEC2*). The novelty of this finding stems from the fact that *DEC2* was not previously associated with high altitude adaptation in human populations, despite being involved in the regulatory pathway of the hypoxia response. Furthermore, DEC2 plays a role in the regulation of the circadian rhythm, perhaps linking the hypoxia response and circadian clock mechanisms in a broader adaptive response.

2.5 Identification of 22 markers to capture the Ethiopian genetic diversity

In order to help labs with limited resources to investigate the genetic history of a broader set of Ethiopian populations, a project was designed to summarize the observed Ethiopian genetic diversity through a small subset of the markers. The number of markers to be used was set to 22, one per autosome, in order to ensure full independence between them and to fit the capacity of PCR- or Sequenom-based genotyping facilities. The target of the project was to find a set of 22 markers that, when interrogated in a set of Ethiopian populations, could reproduce the F_{ST} values and PCA plots generated from the full 1M marker set on the same populations. At least two sets of 22 markers identified in this way performed better than 97.5% of a set of 1000 randomly generated lists in retrieving the observed Ethiopian diversity. Furthermore, these sets performed better than the randomly generated ones even when applied to a group of worldwide populations. However, they were outperformed in the worldwide analyses by sets of markers separately optimized for the worldwide populations. The main outcome of this exercise, described in detail in Appendix 4, was a set of markers that can be used in future studies to position new individuals and populations within the overall diversity of Ethiopian populations. Due to the design strategy, these markers will only be useful to describe Ethiopian samples, although to describe other samples an extension of this approach could be used. Nevertheless, since there are 6000 Ethiopian samples in the collections at UCL, the markers have substantial potential applications.

Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool

Luca Pagani,[1,2,*] Toomas Kivisild,[1] Ayele Tarekegn,[3] Rosemary Ekong,[4] Chris Plaster,[4] Irene Gallego Romero,[2] Qasim Ayub,[2] S. Qasim Mehdi,[5] Mark G. Thomas,[6] Donata Luiselli,[7] Endashaw Bekele,[3] Neil Bradman,[4] David J. Balding,[8] and Chris Tyler-Smith[2]

Humans and their ancestors have traversed the Ethiopian landscape for millions of years, and present-day Ethiopians show great cultural, linguistic, and historical diversity, which makes them essential for understanding African variability and human origins. We genotyped 235 individuals from ten Ethiopian and two neighboring (South Sudanese and Somali) populations on an Illumina Omni 1M chip. Genotypes were compared with published data from several African and non-African populations. Principal-component and STRUCTURE-like analyses confirmed substantial genetic diversity both within and between populations, and revealed a match between genetic data and linguistic affiliation. Using comparisons with African and non-African reference samples in 40-SNP genomic windows, we identified "African" and "non-African" haplotypic components for each Ethiopian individual. The non-African component, which includes the *SLC24A5* allele associated with light skin pigmentation in Europeans, may represent gene flow into Africa, which we estimate to have occurred ~3 thousand years ago (kya). The non-African component was found to be more similar to populations inhabiting the Levant rather than the Arabian Peninsula, but the principal route for the expansion out of Africa ~60 kya remains unresolved. Linkage-disequilibrium decay with genomic distance was less rapid in both the whole genome and the African component than in southern African samples, suggesting a less ancient history for Ethiopian populations.

Introduction

Much of the key fossil evidence for human origins and evolution is found in modern-day Ethiopia. Early putative hominin fossils such as *Ardipithicus kadabba* (5.2–5.8 million years ago [mya])[1] and *Ardipithecus ramidus* (4.4 mya; e.g., "Ardi"),[2] as well as the earliest indisputable hominin species, *Australopithecus anamensis* (3.9–4.2 mya) and the better-known *Australopithecus afarensis* (3.0–3.9 mya; e.g., "Lucy"),[3] have all been found there. It is also the homeland of the earliest known anatomically modern human remains: Omo 1 (195 thousand years ago [kya])[4] and *Homo sapiens idaltu* (154–160 kya).[5] Perhaps for these reasons and because of Ethiopia's geographical position between Africa and Eurasia, its capital, Addis Ababa, is often used in genetic studies as a proxy embarkation point for modern human range expansions.[6,7] However, such studies have seldom included Ethiopians; they are absent from widely used collections, such as the Human Genome Diversity Project (HGDP),[8] HapMap,[9] and 1000 Genomes[10] sets. In practice, our understanding of genome-wide patterns of diversity in Africa has been limited to populations from central and western Africa. Indeed, with a few exceptions,[11,12] studies of African genetic diversity that have included Ethiopians have been restricted to mtDNA[13–16] and the Y chromosome.[14,17] This deficiency has led to an incomplete picture of African genetic diversity that has implications for the study of our origins as a species, including the route followed during the dispersal(s) out of Africa and more recent demographic events involving East Africa.

In linking present-day genetic diversity to the Middle and Late Stone Age populations of Africa, it is important to consider the possibility of long-term population discontinuity in the region and the sparseness of information relating to Ethiopia over the past 200 thousand years (ky). Although archaeological studies focusing on the past few millennia document indigenous Ethiopian developments, including the early cultivation of local species such as teff (*Eragrostis tef*, a cereal), enset (*Musa ensete*), and coffee (*Coffea arabica*),[18] they also reveal some cultural influences from outside, such as the cultivation of wheat and barley, which originated in the Fertile Crescent and reached Ethiopia presumably through Egypt during the first documented trade links, around 5 kya.[19,20] External contacts with the Ethiopian region are also evident in the historical record from the first millennium BCE onward, wherein Sudanese, Egyptian, South Arabic, and Mediterranean influences are documented.[19,21] Another line of evidence for the variegated history of the Ethiopian people comes from linguistic studies. The spread of the two major language families spoken in Ethiopia today—Afro-Asiatic and Nilotic—is considered to be the outcome of cultural and demographic events over the past 10 ky.[22]

[1]Division of Biological Anthropology, University of Cambridge, Cambridge CB2 1QH, UK; [2]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK; [3]Addis Ababa University and Center of Human Genetic Diversity, P.O. Box 1176, Addis Ababa, Ethiopia; [4]The Centre for Genetic Anthropology, University College London, London WC1E 6BT, UK; [5]Centre for Human Genetics, Sindh Institute of Urology and Transplantation, Karachi 74200, Pakistan; [6]Molecular and Cultural Evolution Lab, University College London, London WC1E 6BT, UK; [7]Anthropology Unit, Department of Experimental Evolutionary Biology, University of Bologna, Bologna 40126, Italy; [8]UCL Genetics Institute, University College London, London WC1E 6BT, UK
*Correspondence: lp8@sanger.ac.uk
DOI 10.1016/j.ajhg.2012.05.015.

The presence of three diverse Afro-Asiatic branches (Omotic, Semitic, and Cushitic) makes the Horn of Africa one potential source of this family, although the Ethio-Semitic branch is likely to have originated at a later stage in the Middle East.[23] The Nilotic languages, represented in Ethiopia by the East Sudanic, Kunama, and Koman branches, are more widespread in Sudan, and their presence in Ethiopia is probably the result of recent demographic processes.[24] Similarly, genetic studies indicate that a major component of recent Ethiopian ancestry originates outside Africa: for example, half of the mtDNA haplotypes[16] and more than one-fifth of Y haplotypes[17] found in Ethiopia belong to lineages that, on the basis of phylogeographic criteria, have been attributed to a non-African rather than a sub-Saharan African origin. These historical admixture events are themselves of interest to historians, anthropologists, and linguists, as well as to geneticists.

Our current study is motivated by four questions. First, where do the Ethiopians stand in the African genetic landscape? Second, what is the extent of recent gene flow from outside Africa into Ethiopia, when did it occur, and is there evidence of selection effects? Third, do genomic data support a route for out-of-Africa migration of modern humans across the mouth of the Red Sea? Fourth, assuming temporal stability of current populations, what are the estimated ages of Ethiopian populations relative to other African groups? In order to address these questions, we generated genome-wide SNP genotypes from Ethiopian individuals.

Given that little genetic information on Ethiopian populations was available in advance, we sought to analyze a broad sample of 188 Ethiopians from ten diverse populations, chosen from a collection of > 5,000 samples assembled by N.B.[25,26] The samples genotyped included representatives of a range of geographical regions and all four linguistic groups (Semitic, Cushitic, Omotic, and Nilotic). For comparative studies, we combined our Ethiopian data with published data from the HGDP[27] and HapMap3[9] projects, as well as more focused studies.[28,29] Furthermore, to compensate for the lack of published data of populations immediately surrounding Ethiopia, we additionally genotyped 24 South Sudanese and 23 Somali samples.

Material and Methods

Samples and Genotyping

The Ethiopian and Sudanese DNA samples used in this study were extracted from buccal swabs collected in various Ethiopian and Sudanese locations from apparently healthy, anonymous male donors who provided their informed consent. The collection was performed by members of The Centre for Genetic Anthropology at University College London (UCL) and of Addis Ababa University in Ethiopia, and samples were enrolled into the current study when self-reported ethnicity matched that reported for the donor's parents, paternal grandfather, and maternal grandmother. The populations sampled (numbers) were the Semitic-speaking Amhara (26) and Tigray (21); the Cushitic-speaking Oromo (21), Ethiopian Somali (17), and Afar (12); the Omotic-speaking Ari Cultivators (24), Ari Blacksmiths (17), and Wolayta (8); and the Nilotic-speaking Gumuz (19) and Anuak (23). In addition to these groups, we also generated South Sudanese data from mixed populations (24) and Somali data from Somali populations (23). Additional information, together with the sampling locations of these populations, is available in Table S1 available online. The use of the samples for the present study was approved by the UK research ethics committee (approval numbers 99/0196 and 0489/001). The Somali DNA samples (previously obtained from Somali expatriates in Islamabad, Pakistan) were extracted from lymphoblastoid cell lines in the collection created by S.Q.M.

All the samples were whole-genome amplified with the GE GenomiPhi HY DNA Amplification Kit (catalog no. 25-6600-25, General Electric) and genotyped on the Illumina Omni 1M SNP array at the Wellcome Trust Sanger Institute. SNP calls and quality checks were performed by the Sanger genotyping facility with the use of GenoSNP.[30] Y-chromosomal haplogroups were also determined at both UCL and Sanger labs. The above 235 genotypes were pooled with data from published sources,[9,27–29] providing ~280,000 overlapping markers in 4,442 individuals.

For the fixation index (F_{ST}), mtDNA, and genomic minimum pairwise distance, we chose to reference non-African populations along the two putative routes: Bedouin, Druze, Palestinian, Syrian, Lebanese, Jordanian, Iranian, Greek, French, Pathan, Han, and Surui populations representing the northern route; Yemeni, Saudi Arabian, Dravidian, and Papuan populations representing the southern route.

Summary Statistics

SNP frequencies, heterozygosity, and linkage disequilibrium (LD, r and r^2) were calculated for each group with PLINK,[31] and pairwise F_{ST} values were calculated with an in-house script implementing the Weir and Cockerham formula.[32] The F_{ST} and heterozygosity values were interpolated and plotted on a geographic map with Surfer (Golden Software). The merged data set was pruned to remove SNPs in high LD ($r^2 > 0.1$), and ADMIXTURE analyses were run as described[33] after removal of samples showing high relatedness (PLINK identity-by-descent score ≥ 0.125) with any other sample in the same population (1 Amhara, 2 Ari Cultivators, 6 Ari Blacksmiths, 3 South Sudanese, and 1 Gumuz).[34] Cross validation was used to estimate the optimum number of clusters (K). Principal-component analysis (PCA) was implemented with EIGENSTRAT[35] on the same pruned data set.

We phased the one million Ethiopian SNPs with BEAGLE,[36] incorporating information from the HapMap3 YRI (Yoruba in Ibadan, Nigeria from the CEPH collection) trios.[9] Candidate population-specific signals of positive selection were identified with the integrated haplotype score (iHS) statistic.[37]

Genome Partitioning

We implemented the following approach, modified from published chromosome-painting methodology,[28] to partition each individual genome into windows that were more similar to the African and non-African populations, respectively. To obtain a list of SNPs that were independent in each of the reference populations, we LD pruned[34] the data in three steps, using 20 French, 20 Han Chinese, and 20 Yoruba samples, sequentially. The pruned markers were then divided into 40-SNP, nonoverlapping windows covering the whole genome. Every window was then phased independently within each population with the

PHASE program,[38] and the phased haplotypes were used in the following steps.

Each test haplotype was compared with haplotypes from the corresponding genomic window taken from 20 individuals from each of the three reference populations (Han Chinese, French, and Yoruba). The comparison was performed by running a PCA with the use of the "princomp" function of the R package. Three reference clouds (Han Chinese, French, and Yoruba) were defined by the median and 50% confidence radius, calculated from the relevant haplotypes. The Euclidean distance between the principal component (PC) coordinates of the test haplotype and the confidence perimeter of each cloud were then calculated. Due to the similarity between the European and Asian haplotypes relative to the African haplotypes and the consequent difficulty in drawing a clear separation between the two non-African clouds, we then labeled each test 40-SNP haplotype as either "African" or "non-African" according to its position in the PCA plot, or "NA" if there was no separation between the reference clouds. The "NA" haplotypes (less than 1% of the total) were removed from the downstream analyses.

Analyses of Partitioned African and Non-African Genomic Components

The resulting genome partitions were used in a series of analyses whereby either the African or the non-African component of a set of populations was taken into consideration. In order to compare various populations with different levels of African and non African components, we pooled together either the African or non-African haplotypes to create ten mosaic haploid genomes per population. Each mosaic haploid genome would then include either African or non-African haplotypes from different individuals of the same population.

To analyze the LD of the African component of each genome, we included all available SNPs and calculated LD decay over a range of distances as described.[28]

The minimum pairwise distance between African and non-African populations was calculated using ten mosaic non-African haploid genomes (made of either African or non-African haplotypes only) from each Ethiopian, Somali, and Sudanese population (together, "Ethiopian+"). For each Ethiopian+ 40-SNP window, we calculated the shortest distance to the same window in the non-African population, and averaged the distance over all windows in each population.

A Z-score based on the number of chromosomes in the non-African state was assigned to each 40-SNP window in each of the five Semitic-Cushitic populations. The Z-score was calculated for each 40-SNP window in each population on the basis of the average and SD of the full set of regions for that population. We then binned the Z-scores and counted the number of regions occurring for a given bin in a given number of the examined populations. Any region showing a Z-score > 2 or < -2 in more than two of the five populations examined was flagged as an outlier, and its gene content was examined for functional interest.

Assuming that the African and non-African components of the Ethiopian genomes result from a single admixture event, we used ROLLOFF[39] to estimate the midpoint of the period of admixture. However, if there were multiple or continuous admixture events, as with the North African populations, this method detected[39] the most recent event or the admixture midpoint, respectively. ROLLOFF computes the correlation between (1) a (signed) statistic for LD between a pair of markers and (2) a weight that reflects their allele-frequency differentiation in the ancestral populations. We used as putative ancestral populations either CEU (Utah residents with ancestry from northern and western Europe) and YRI (as previously described[39]) or CEU and Ari, chosen because of their extremal positions in a PC plot (Figure S4). Because of the lack of publicly available code at the time of the analyses, the ROLLOFF algorithm was recoded in-house (details available upon request) from the description provided,[39] following advice kindly provided by its authors, and was shown to give similar age estimates ($r^2 > 0.9$, data not shown) for a set of test populations previously analyzed with the use of this approach[39] (African Americans, Palestinians, Sardinians, Bedouins, and Druze; all treated as a mixture of CEU and YRI). Before running the analyses, we performed a PCA on the Ethiopian, North African, and Middle Eastern individuals, together with YRI and CEU, to identify and remove outlier individuals (1 Amhara, 1 South Sudanese, 1 Bedouin, 2 Egyptian, 3 Moroccan, 4 Mozabite, 1 Saudi, and 1 Yemeni) and to split those populations forming more than one cluster (e.g., Oromo was divided into Oromo1 and Oromo2), as recommended by the authors.

Results

In the following sections, we consider sequentially the four questions identified in the Introduction, and thus move from more recent to more ancient events.

Modern Ethiopians in the African Genetic Landscape

The first PC of the African samples separates sub-Saharan Africans from North Africans, with Ethiopians positioned between them (Figure 1A), whereas the second and third components separate the hunter-gatherers (click speakers and Pygmies) and the East Africans, respectively (Figures 1A and 1B). Both plots separate the Ethiopian samples according to their linguistic origin. This linguistic clustering appears to be more important than geographical structure, especially for the Semitic and Cushitic populations (Figure 1D), and is also supported by the neighbor-joining tree of Figure S2. Remarkably, the Ethiopian clusters, taken together, span half of the space delimited by all the African populations and surround the Maasai from Kenya. To investigate this high diversity further, we performed an African-only PCA (Figure S1A) using five randomly chosen samples from each Ethiopian population, in order to eliminate bias that might arise from including a large number of Ethiopian samples, and a worldwide PCA using the full data set (Figure S1B). Both plots confirmed the high diversity in Ethiopia; Ethiopians spanned most of the African branch in the worldwide PCA (Figure S1A) and showed similar internal structure in both PCA plots (Figures 1B and S1B).

ADMIXTURE[34] was applied to the same African data set, with the addition of the HGDP French[27] as a reference group for the non-African component (Figure S1C). The best-supported[34] clustering ($K = 7$, Figure 1C) divided the Ethiopians into two main groups: the Semitic-Cushitic Ethiopians stand out as a relatively uniform set of

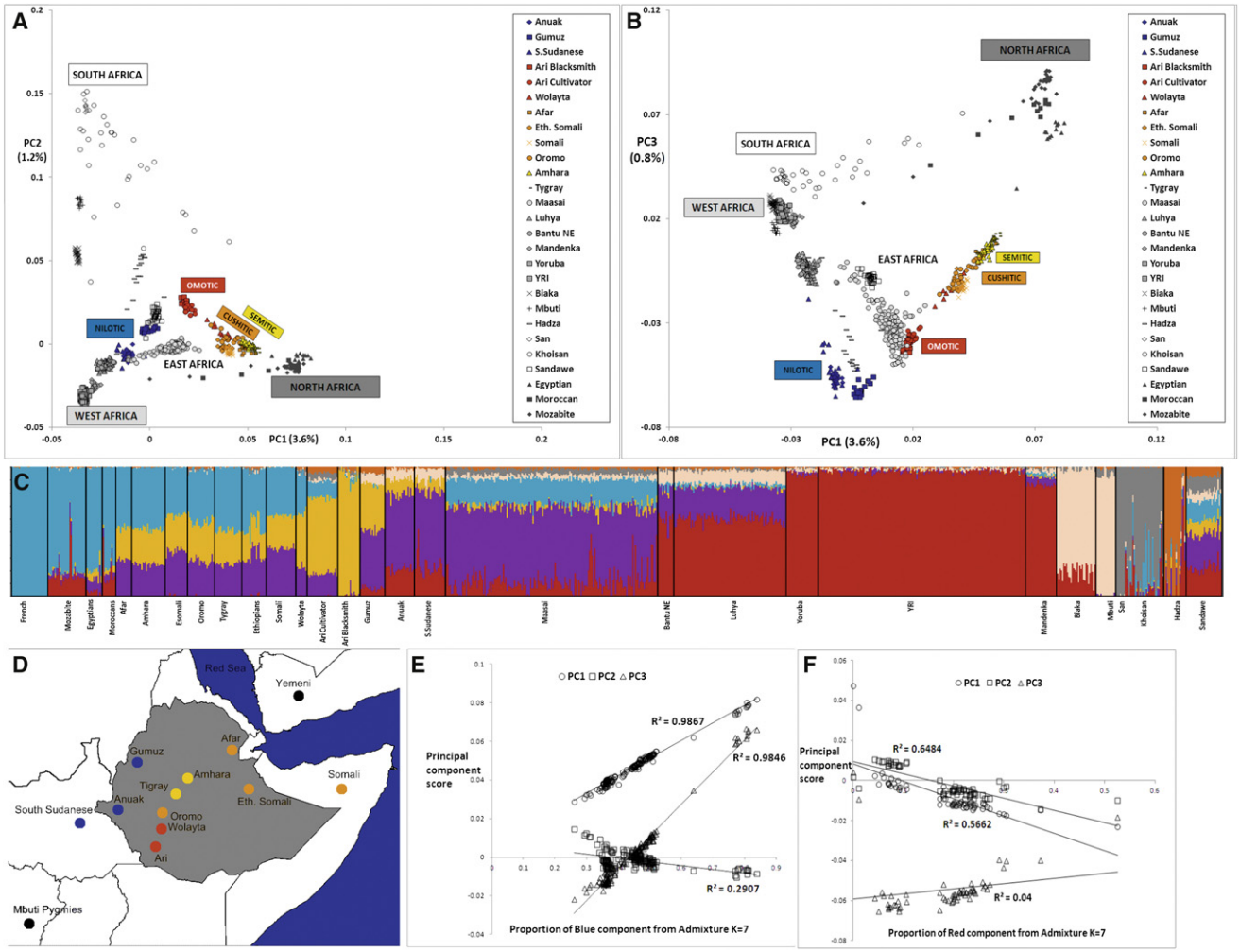

Figure 1. Principal Components and STRUCTURE-like Analyses of the Full African Data Set
The first three PCs are represented in bidimensional plots (first versus second in A and first versus third in B). The samples genotyped in this study are represented in yellow (Semitic), orange (Cushitic), red (Omotic), or blue (Nilotic); the rest of the African samples are shown with the use of a gray scale. The proportion of explained variance is reported next to each axis.
(C) displays the best fit (K = 7) ADMIXTURE result, including all the African samples and with the addition of French as a non-African population. The colors in (C) do not match those in (A) and (B).
(D) shows the sampling locations in Ethiopia. Each population is colored according to the linguistic family to which it belongs.
(E) Correlation between the proportion of "non-African" admixture (x axis, blue component from C) and the first three PCs for the Semitic, Cushitic, Omotic, and Egyptian samples.
(F) Correlation between the proportion of Nigerian-Congolese admixture (x axis, red component from C) and the first three PCs for the Anuak, Gumuz, and South Sudanese samples.

individuals characterized by a strong (40%–50%) non-African component (light blue in Figure 1C) and an African component split between a broad East African (purple in Figure 1C) and an apparently Ethiopia-specific component (yellow); the Nilotic and Omotic Ethiopians show little or no non-African component and are instead characterized by eastern (purple and yellow) or western (dark red) African components, with some traces of additional components. The yellow and purple components represent the major proportion of the African component in the Egyptian Afro-Asiatic population, but are less predominant than the red West African component among northwestern African populations who also speak Afro-Asiatic languages. However, it is striking that North Africans share substantially more variation with non-African populations (80%) than do Ethiopians (40%–50%).

To investigate the role played by the non-African component in the PCA clustering of the Semitic and Cushitic samples, we looked for correlations between the former (obtained from ADMIXTURE, K = 7) and the first three PCs. As shown in Figure 1E, both PC1 and PC3 strongly correlate (both r^2 values are above 0.98) with the blue component of Figure 1C, whereas PC2 shows a weaker correlation (r^2 = 0.29). The strong PC1 and PC3 correlations therefore seem to indicate that the proportion of non-African admixture is the main driver of the Ari-Egyptian cline formed by the Semitic-Cushitic samples in the PCA plot, regardless of their population of origin.

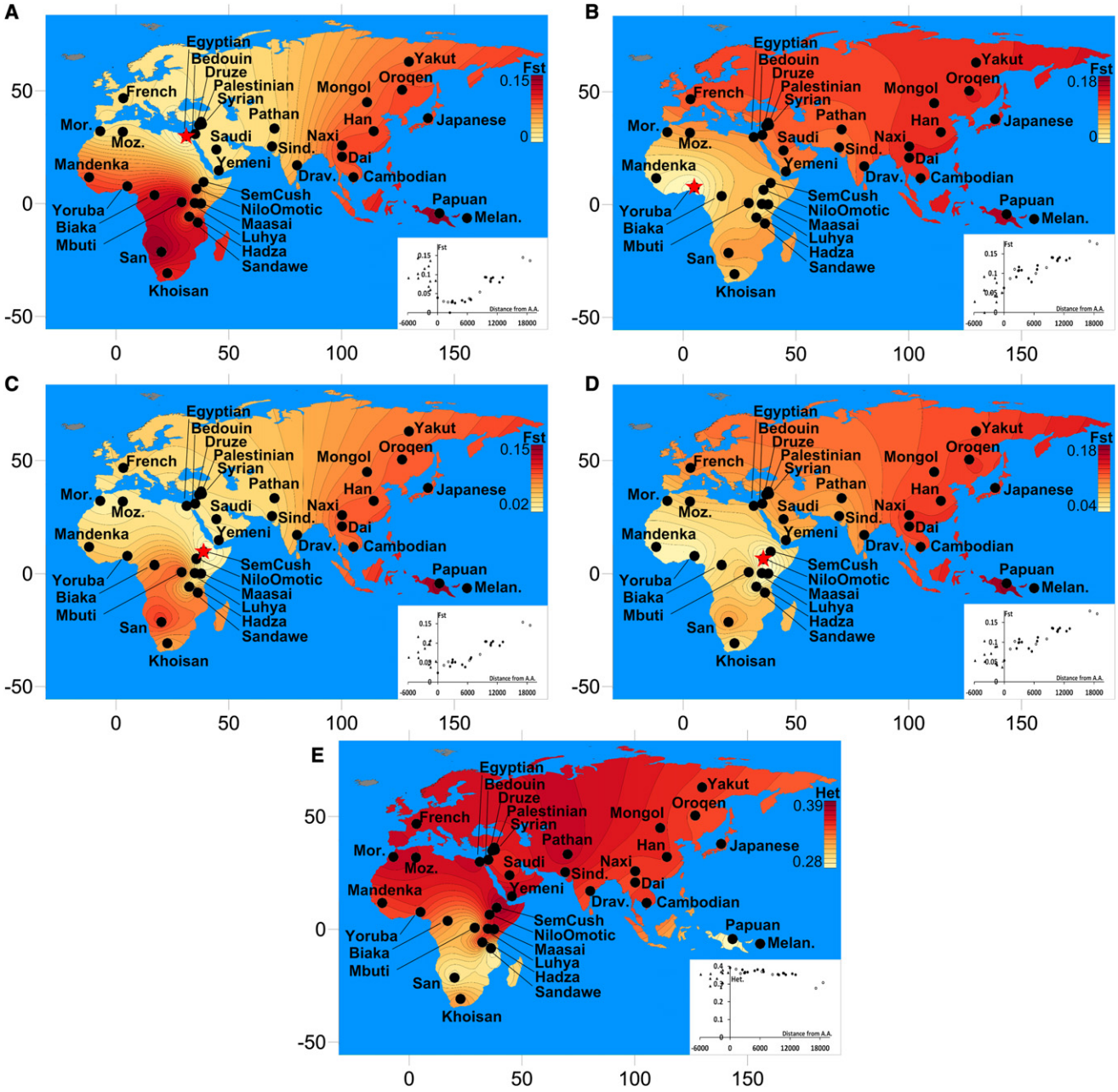

Figure 2. Pairwise F_{ST} and SNP Heterozygosity in a Set of Worldwide Populations

F_{ST} was calculated with the use of ten individuals from each worldwide population and Egyptians (A), Yoruba (B), and Semitic-Cushitic (C) and Nilotic-Omotic Ethiopians (D), and is displayed as a heat surface, produced with the Surfer software. Values in (C) and (D) are the averages for all the Semitic-Cushitic or Nilotic-Omotic populations. (E) shows the average genomic heterozygosity calculated for the same samples with the use of the available SNPs. The bottom-right section of each panel includes a scatter plot displaying the actual values of either *F*st or heterozigosity over the geographic distance (in km) from Addis Ababa (negative for sub-Saharan populations). Filled and empty circles represent non-African populations along the putative northern or southern routes, respectively. Triangles represent sub-Saharan populations.

However, when looking for correlations between the Nigerian-Congolese component (blue in Figure 1C) and the first three PCs in the Nilotic populations, we found a much weaker correlation (Figure 1F) than observed for the Semitic-Cushitic component. The Ari-Yoruba cline observed for the Nilotic samples cannot therefore be explained as a simple admixture event between Ethiopians and Nigerian-Congolese populations.

To compare the level of genetic variation in the populations investigated, we estimated average SNP heterozygosity in the pruned genomes of ten individuals from each population and the pairwise F_{ST} between African and worldwide populations (Figure 2 and Table S2). The Semitic-Cushitic and North African populations showed the highest values of heterozygosity worldwide, which may reflect a combination of SNP ascertainment bias and

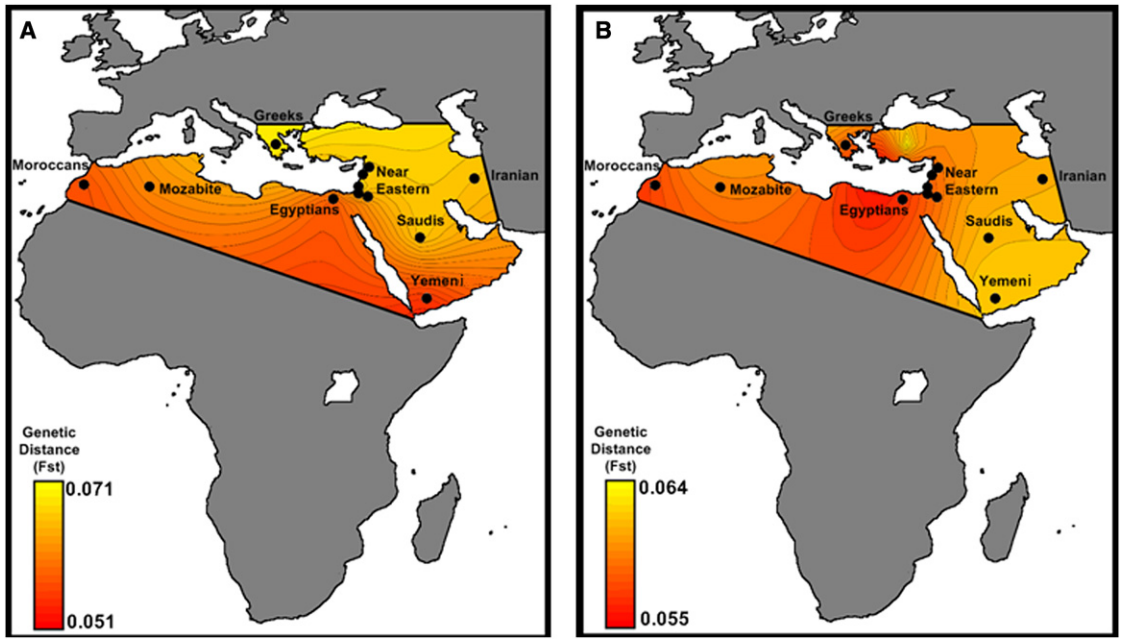

Figure 3. Pairwise F_{ST} between Semitic-Cushitic Ethiopians and Surrounding Populations
Contour plots derived from F_{ST} were calculated with (A) ten haploid genomes from the Semitic-Cushitic Ethiopians, showing that modern Yemeni, Egyptians, and Moroccans are closest to the Ethiopians, and (B) ten haploid non-African genomes from the same groups, showing instead a prevalence of Egyptian and Middle Eastern contributions to the non-African Ethiopian gene pool.

the mixture of African and non-African components in these populations. The observed pattern of uniform decline of F_{ST} values away from North, West, or East Africa is consistent with previous interpretations of a single exit, followed by "isolation by distance."[6,40,41]

Back to Africa

Before considering questions related to ancient demographic events, we needed to separate the probable ancient African components from that which might have originated from more recent (<60 kya) gene flow back to Africa (light blue in Figure 1C).

In order to perform this partitioning, we modified a PCA-based method,[28] dividing the genome into haploid windows of 40 SNPs and labeling each as either African or non-African (see Material and Methods). The effectiveness of this method was assessed through comparison of the proportion of each individual genome assigned to an African or non-African origin by PCA with the ADMIXTURE K = 2 clustering. The patterns are very similar (Figures S3A and S3B), and the correlation between the proportions is high ($r^2 > 0.99$; Figure S3C). The added value of the PCA approach is that it locates the African and non-African haplotype windows within each genome, and thus allows their subsequent analysis.

We calculated the genetic distance (F_{ST}) between Semitic and Cushitic Ethiopians and populations of the Levant, North Africa, and the Arabian Peninsula using two approaches: (1) the whole genome and (2) only the non-African component. In the whole-genome analysis, Ethiopian Semitic and Cushitic populations appear to be closest to the Yemeni (Figure 3A); when only the non-African component is used, they are closer to the Egyptians and populations inhabiting the Levant (Figure 3B). We explored this finding further by calculating the minimum pairwise difference (see Material and Methods) between Africans and non-Africans for their whole genome, and for the non-African component only. The results are concordant with the results of the F_{ST} analyses in showing that the Egyptians are closer than Yemeni to Ethiopians in their non-African component (Table S3). A possible explanation for this result is that there has been gene flow into Ethiopia from the Levant and Egypt, although we cannot say whether the gene flow was episodic or continuous. The Ethiopian similarity with the Yemeni detected throughout the genome could be explained as an Ethiopian contribution to the Yemeni gene pool, consistent with that observed with mtDNA.[16]

We considered two sources (western and eastern) for the African component of the Ethiopian genomes. The distinction between the East and West African components is supported by the PCA, wherein our samples formed a triangle (Figure S4) with the three corners represented by West Africans (YRI), non-Africans (CEU), and East Africans (Ari Cultivators and Blacksmiths). The other populations were distributed along the three sides of the triangle in a way that could imply different patterns of admixture. We applied ROLLOFF to estimate admixture dates for the Ethiopian populations, considered as a combination of West Africans with non-Africans or East Africans with

Table 1. Admixture Date Estimates in East and North African Populations

Region	Population	YRI-CEU Admixture Date	Ari-CEU Admixture Date
East Africa	Ari Blacksmith	−1228	NA
East Africa	Ethiopian Somali	−1094	−1201
East Africa	Ari Cultivator	−1017	NA
East Africa	Somali	−953	−1996
East Africa	Amhara	−637	−1502
East Africa	Tygray	−425	−1319
East Africa	Wolayta	−209	−1418
East Africa	Afar	−170	−1039
East Africa	Oromo1	−168	−1062
East Africa	Anuak	71	NA
East Africa	Oromo2	96	−906
West Asia	Druze	767	958
East Africa	Maasai	883	NA
West Asia	Saudi2	1109	1232
North Africa	Egyptian	1117	1283
West Asia	Bedouin2	1130	1122
West Asia	Palestinian	1159	1137
West Asia	Saudi	1164	1466
North Africa	Moroccan	1176	1407
West Asia	Bedouin1	1256	1365
North Africa	Mozabite	1267	1388
West Asia	Yemeni	1548	1548
East Africa	Gumuz	1588	NA
East Africa	South Sudanese	1839	NA
North America	African American	1855	NA

The date of admixture for each populations reported in the table was calculated with an in-house version of the ROLLOFF algorithm.[39] To facilitate the interpretation of results, we converted the number of generations into years using 30 years per generation, and then into a CE or BCE date by subtracting 2011. Column 3 reports this date, and models the populations as a mixture of CEU and YRI (Utah residents with ancestry from northern and western Europe and Yoruba in Ibadan, Nigeria, respectively, from the CEPH collection).[39] Column 4 reports corresponding estimates, modeled assuming admixture between CEU and the Ari Ethiopians. The rationale for these two analyses is provided in Figure S4. NA, not available.

non-Africans, depending on their position in the PC plot (Figure S4). The dates of admixture (assuming 30 years per generation)[42] are reported in Table 1. Notably, in most of the Semitic, Cushitic, and Omotic populations, the admixture of African and non-African ancestry components dates to 2.5–3 kya, whereas in North Africa, the admixture dates are ~2 ky more recent, clustering around 1 kya, consistent with previous reports.[43] The consistency between the Ethiopian estimates and the appearance in the area of a linguistic family (Ethio-Semitic) with a West Asian origin[23] support the hypothesis of a recent gene flow from the Levant. Although ROLLOFF estimated a date for an admixture event involving the Nilotic populations, examination of the relationship between the correlation coefficient and genetic distance (Figure 4) revealed no exponential decay for these populations, implying less support for an admixed origin of the Nilotic populations than of the Semitic, Cushitic, and Omotic populations.

Selection Following Admixture

An intriguing consequence of admixture between populations is the opportunity for packages of genes to be "tested" in different environments. As a result, the genomic regions containing functionally divergent genes might experience either positive or negative selection, depending on whether their adaptive contribution was beneficial or damaging in the new environment, or whether it affected social factors such as sexual selection. To look for such outlier regions of admixture in Ethiopian populations (Semitic and Cushitic) where the estimated proportions of African and non-African ancestries were roughly equal, we listed those regions showing an excess or a deficit (see Material and Methods) of non-African haplotypes (Table S4). Of the fourteen 40-SNP windows observed with a Z-score > 2, we noted one that contained *SLC24A5* (MIM 113750). This gene is a major contributor to the pigmentation differences between Africans and Europeans and a strong candidate for positive selection in Europe.[44,45] Given that *SLC24A5* is one of the most highly differentiated genes between African and European populations,[10,46] we then looked for other highly differentiated genes[10] among the outlier windows, but found none. We also checked whether the 24 large Z-score windows reported in Table S4 showed enrichment for regions with extreme distances between the African and non-African clouds. After ranking all the 40-SNP windows by the distance between the African and European cloud centers divided by the SD of the European cloud around its center, none of the large Z-score windows were present within the top 1%. We therefore speculate that the excess of non-African SLC24A5 haplotypes must be linked to the biological function of that gene.

The iHS scan performed on the Semitic-Cushitic populations (considered as a whole) confirmed that *SLC24A5* was within the top 5% of selection signals, whereas the gene was not detected as an outlier in the other groups of Ethiopians. The unusual history of this gene was further supported by the presence of the derived A allele of the SNP rs1834640, associated with the light skin pigmentation of Europeans and western Asians,[47] at higher frequencies in Semitic-Cushitic groups compared with Omotic, Nilotic, or Nigerian-Congolese groups (0.55 versus 0.23, 0.07, and 0.04, respectively). To further investigate the effect of admixture on the genetic landscape of skin pigmentation in Ethiopia, we also looked at other genes associated with pigmentation in Europe;[46] however, none were found in our outlier regions.

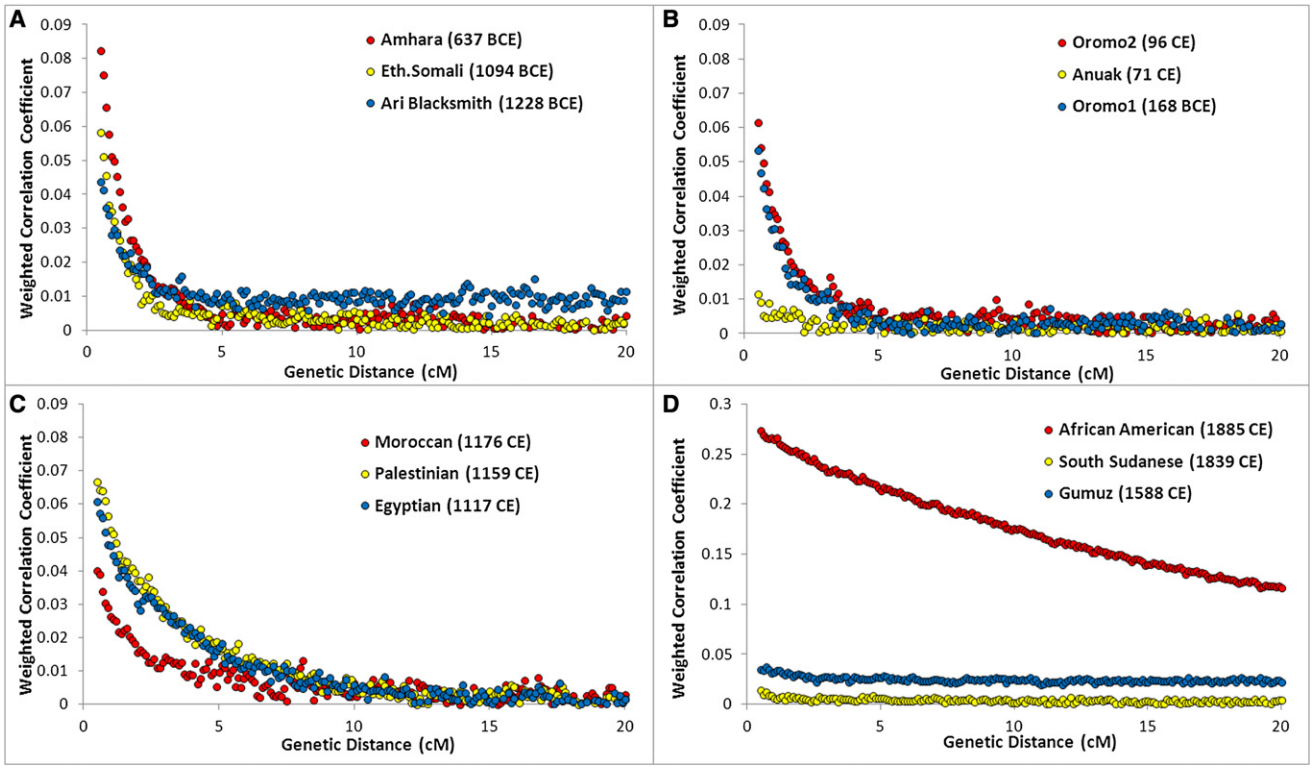

Figure 4. ROLLOFF Plots
Three populations from each of the four historical periods of admixture (A: <500 BCE, B: ~0 CE, C: ~1000 CE, and D: >1500 CE) are plotted to show their LD decay (represented by a weighted correlation coefficient as previously described[39]) with genetic distance. The legend reports the name of each population, with the estimated date of admixture in brackets. Notably, all three Nilotic populations (Gumuz, South Sudanese, and Anuak) have very flat decay curves compared to those of the other populations in the same plot.

Source of the Major Out-of-Africa Migration

Consistent with previous studies' reports of a steady decline in genetic similarity among non-African populations as a function of geographical traveling distance from East Africa, we found that the F_{ST} values estimated between either Ethiopian or North African populations and non-African populations followed the same pattern (Figure 2, Table S2). This steady decline has been argued[27] to be compatible with a single exit followed by isolation-by-distance, rather than with two distinct African sources contributing to the non-African diversity. Neither including nor excluding the Ethiopian data altered the pattern. To follow the thread left by this dispersal in more detail, we used the genome partitioning performed earlier to calculate the minimum pairwise difference between the African component of the Egyptian and Ethiopian populations and the equivalent genomic segment in non-Africans. The partitioning would remove noise, caused by recent backflows into Africa, which might otherwise mask the original out-of-Africa signal. If the mouth of the Red Sea had been a major migration route out of Africa, we might observe a closer affinity of Ethiopians, rather than Egyptians, with non-Africans.

As a proof of principle, we first applied the approach to a genetic system with a well-understood phylogeographic structure: mtDNA. Virtually all indigenous sub-Saharan African mtDNA lineages belong to L haplogroups, whereas the presence of haplogroups M and N in North and East Africa has been interpreted as a signal of gene flow back to Africa.[48,49] With the full set of 18 mtDNA SNPs used in our genome-wide data set, Egyptians and Moroccans proved to be the closest African population to any non-African population examined (Table 2A). However, when we first partitioned the mtDNA lineages into African and non-African (i.e., L and non-L) and considered only the L component, a different pattern emerged: Ethiopians were the closest population to the non-Africans (Table 2B), consistent with inferences drawn from more detailed mtDNA analyses.[50]

Applying the same principle, we then calculated the shortest distance between the African and non-African populations on the basis of either full genome data or the African component of this data set. In contrast to the mtDNA results, the Egyptians proved to be the closest to the non-Africans in both cases (Tables 2A and 2B).

Relative Ages of the Ethiopian and Other African Populations

The decay of LD with time provides a robust proxy for the "age" of a population of a constant size: that is, the length of time that the ancestors of the sampled individuals have been evolving as part of the same breeding unit. To assess

how relatively "old" the patterns of LD are in Ethiopian populations, we compared the LD at different distances between the Ethiopian populations and a range of other African populations (Figure 5).[28] We also performed the same analyses on the African components of each population to reduce the bias introduced by the recent genetic back-flow (Figure 5B). In both cases, the Ethiopians displayed less LD decay than did the click speakers, Pygmies, or Nigerian-Congolese groups, suggesting a younger age, a smaller long-term effective population size, or a combination of these.

Discussion

We present an extensive genome-wide data set representing Ethiopian geographical, linguistic, and ethnic diversity. Its study has allowed us to cast light on a number of questions, some long-standing, about both ancient and recent demographic events in human evolution. In the Discussion, we again follow a roughly chronological path from the more recent to the older events.

The Ethiopian populations show high genetic diversity, with stratification matching the linguistic families (Figure 1B), except for the overlap in both PCA and F_{ST} analyses of populations belonging to two mutually unintelligible linguistic groups (Semitic and Cushitic). This overlap reflects both the similar amount of non-African genome present in these individuals and the similar African component (Figures 1C and 1E). It may also reflect factors such as the recent expansion of some Cushitic and Semitic groups and landscape such as highland and lowland environments. Of particular interest is the distinctiveness of the Omotic groups, whose position in Figures 1A and S3 is intriguingly compatible with being a putative ancestral Ethiopian population. One insight provided by the ADMIXTURE plot (Figure 1C) concerns the origin of the Ari Blacksmiths. This population is one of the occupational caste-like groups present in many Ethiopian societies that have traditionally been explained as either remnants of hunter-gatherer groups assimilated by the expansion of farmers in the Neolithic period or as groups marginalized in agriculturalist communities due to their craft skills.[51] The prevalence of an Ethiopian-specific cluster (yellow in Figure 1C) in the Ari Blacksmith sample could favor the former scenario; the ancestors of this occupational group could have been part of a population that inhabited the area before the spread of agriculturalists. Further study of multiple groups comparing agriculturists and caste-like groups would reveal whether there is a pattern of a greater Ethiopia-specific genomic profile associated with caste-like occupations, an observation which would support the absorption rather than the exclusion hypothesis.

ADMIXTURE analyses revealed a major (40%–50%) contribution to the Ethiopian Semitic-Cushitic genomes that is similar to that of non-African populations. Our estimates of genetic similarity between this component and extant non-African populations suggest that the source was more likely the Levant than the Arabian Peninsula. We estimate that this admixture event took place approximately 3 kya. The more recent admixture dates for the Oromo and Afar can be explained by the effect of a subsequent Islamic expansion that particularly impacted these groups, as well as the North Africans.[52] Levant people may have arrived in Ethiopia via land or sea subsequently, leaving a similar signature also in modern Egyptians, or the similarity between Ethiopians and Egyptians may be a consequence of independent genetic relationships. This putative migration from the Levant to Ethiopia, which is also supported by linguistic evidence, may have carried the derived western Eurasian allele of *SLC24A5*, which is associated with light skin pigmentation. Although potentially disadvantageous due to the high intensity of UV radiation in the area, the *SLC24A5* allele has maintained a substantial frequency in the Semitic-Cushitic populations, perhaps driven by social factors including sexual selection. The "African" component of the Ethiopian genomes may also result in part from recent migrations into Ethiopia from other parts of Africa, a possibility that we have not examined here.

The estimated time (3 kya) and the geographic origin (the Levant) of the gene flow into Ethiopia are consistent with both the model of Early Bronze Age origins of Semitic languages and the reported age estimate (2.8 kya) of the Ethio-Semitic language group.[23] They are also consistent with the legend of Makeda, the Queen of Sheba. According to the version recorded in the Ethiopian *Kebra Nagast* (a traditional Ethiopian book on the origins of the kings), this influential Ethiopian queen (who, according to Hansberry,[53] reigned between 1005 and 955 BCE) visited King Solomon—ruler, in biblical tradition, of the United Kingdom of Israel and Judah—bringing back, in addition to important trading links, a son. The ancient kingdom of Axum adopted Christianity as early as the fourth century. Historical contacts established between Ethiopia and the Middle East were maintained across the centuries, with the Ethiopian church in regular contact with Alexandria, Egypt. These long-lasting links between the two regions are reflected in influences still apparent in the modern Ethiopian cultural and, as we show here, genetic landscapes.

An abundance of evidence suggests that all modern non-Africans descend predominantly from a single African source via a dispersal event some 50 to 70 kya.[6,7,27,49] However, debate continues about whether the principal migratory route out of Africa was north of the Red Sea to the Levant, or across its mouth to the Arabian Peninsula. The actual source of the migrations within Africa is a different question, but we assume that the migrators would have left genetic signatures in Egypt if they took the northern route or in Ethiopia if they took the southern route. We chose reference non-African populations along

Table 2. Minimum Pairwise Difference between Africans and Non-Africans Calculated for the Whole Genome or mtDNA and for their African Component Only

Population	Cushitic-Semitic	Omotic	Nilotic	Egyptian	Moroccan	Mozabite
Whole Genome						
Han	0.0407	0.0418	0.0422	0.0402	0.0407	0.0406
Bedouin	0.0385	0.0402	0.0409	0.0365	0.0375	0.0379
Druze	0.0386	0.0403	0.0412	0.0365	0.0376	0.0379
French	0.0391	0.0409	0.0419	0.0372	0.0381	0.0378
Greek	0.0389	0.0408	0.0416	0.0369	0.0378	0.0378
Iranian	0.0389	0.0406	0.0413	0.0372	0.0382	0.0382
Jordanian	0.0386	0.0402	0.0410	0.0365	0.0379	0.0379
Lebanese	0.0385	0.0403	0.0411	0.0368	0.0376	0.0377
Moroccan Jews	0.0386	0.0403	0.0412	0.0364	0.0375	0.0376
Palestinian	0.0387	0.0403	0.0411	0.0370	0.0377	0.0379
Saudi	0.0386	0.0404	0.0412	0.0367	0.0377	0.0378
Syrians	0.0387	0.0404	0.0414	0.0364	0.0377	0.0379
Yemeni	0.0384	0.0396	0.0399	0.0372	0.0373	0.0384
Yemeni Jews	0.0385	0.0404	0.0412	0.0367	0.0375	0.0380
AVERAGE	0.0388	0.0404	0.0412	0.0370	0.0379	0.0381
Whole mtDNA Pool						
Bedouin	0.0024	0.0033	0.0041	0.0024	0.0012	0.0024
Palestinian	0.0020	0.0023	0.0028	0.0017	0.0006	0.0011
Saudi	0.0008	0.0015	0.0012	0.0025	0.0019	0.0025
Yemeni	0.0031	0.0046	0.0062	0.0044	0.0040	0.0040
Yemeni Jews	0.0018	0.0022	0.0022	0.0017	0.0022	0.0022
French	0.0019	0.0014	0.0023	0.0000	0.0000	0.0006
Pathan	0.0020	0.0017	0.0028	0.0011	0.0006	0.0017
Dravidian	0.0008	0.0008	0.0011	0.0000	0.0000	0.0006
Papuan	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
AVERAGE	0.0017	0.0020	0.0026	0.0016	0.0012	0.0017
African Component						
Han	0.0420	0.0414	0.0412	0.0415	0.0418	0.0434
Bedouin	0.0397	0.0395	0.0394	0.0364	0.0382	0.0406
Druze	0.0399	0.0398	0.0396	0.0365	0.0388	0.0410
French	0.0406	0.0404	0.0402	0.0379	0.0392	0.0416
Greek	0.0403	0.0401	0.0400	0.0375	0.0389	0.0412
Iranian	0.0402	0.0400	0.0397	0.0375	0.0389	0.0412
Jordanian	0.0399	0.0397	0.0395	0.0371	0.0385	0.0408
Lebanese	0.0399	0.0396	0.0394	0.0371	0.0381	0.0407
Moroccan Jews	0.0400	0.0398	0.0395	0.0367	0.0385	0.0409
Palestinian	0.0400	0.0399	0.0395	0.0366	0.0382	0.0408
Saudi	0.0399	0.0398	0.0394	0.0366	0.0385	0.0408
Syria	0.0401	0.0398	0.0397	0.0366	0.0387	0.0411

Table 2. *Continued*

Population	Cushitic-Semitic	Omotic	Nilotic	Egyptian	Moroccan	Mozabite
Yemeni	0.0394	0.0391	0.0387	0.0367	0.0378	0.0403
Yemeni Jews	0.0399	0.0397	0.0395	0.0364	0.0384	0.0407
AVERAGE	0.0401	0.0399	0.0397	0.0372	0.0388	0.0411
L-mtDNA Only						
Bedouin	0.0036	0.0034	0.0051	0.0077	0.0050	0.0058
Palestinian	0.0026	0.0024	0.0035	0.0056	0.0032	0.0032
Saudi	0.0023	0.0016	0.0015	0.0041	0.0027	0.0035
Yemeni	0.0054	0.0047	0.0077	0.0102	0.0072	0.0072
Yemeni Jews	0.0029	0.0023	0.0028	0.0037	0.0032	0.0032
French	0.0025	0.0015	0.0029	0.0038	0.0033	0.0033
Dravidian	0.0013	0.0009	0.0014	0.0019	0.0016	0.0016
Pathan	0.0032	0.0017	0.0035	0.0056	0.0040	0.0048
Papuan	0.0008	0.0006	0.0007	0.0010	0.0008	0.0008
AVERAGE	0.0027	0.0021	0.0032	0.0048	0.0034	0.0037

the two putative routes. However, both the northern and eastern Africans have genetic distances (F_{ST}) that gradually increase with geographic distance along both routes. This also holds true when Ethiopian populations that show little evidence of recent non-African gene flow (Omotic and Nilotic) are used as a source. A minimum-pairwise-distance measure based on the African component of the genome found that the Ethiopian mtDNA component was closer to non-African populations than was the Egyptian mtDNA component, as previously reported,[50] but that the autosomal genome of non-Africans was closer to the African component of the Egyptian rather than Ethiopian populations. This could be interpreted as supporting a northern exit route. However, the 80% non-African proportion of the Egyptian genome (Figure 1C) reduces the power of our comparisons and, taken together with the requirement for the African state in at least ten chromosomes, means that this conclusion is based on just ~1,800 SNPs (compared to 18,960 for the Ethiopians, 30,798 for the Mozabite, and 5,920 for the Moroccans). Therefore, the question requires further investigation beyond the scope of the present study.

On a broader time scale, the LD analyses pointed to click speakers, Pygmies, and a Nigerian-Congolese group as all having a deeper population history than both the whole genome and the African component of the East Africans sampled. Although this result might seem inconsistent with the outstanding fossil record available from Ethiopia, it may illustrate that genetic diversity assessed from modern populations does not necessarily represent their long-term demographic histories at the site. Alternatively, the rich record of human fossil ancestors in Ethiopia, and indeed along the Rift Valley, may reflect biases of preservation and discovery, with more fossils being exposed in regions of geological activity. Fluctuations in effective population size in the past and dispersals within Africa may have further confounded our analyses and their correlation with the fossil record. The fact that the observed genetic diversity in Ethiopia is lower than in some other African populations does not negate the possibility that Ethiopia was the cradle of anatomically modern humans. However, interpretations of the LD-based analyses may be challenged by future work in two key respects. First, whole-genome sequences can provide an independent measure of the demographic history of the groups studied,[54] but they have not yet been applied to Ethiopian samples. Second, there is a need for a better understanding of the implication for the genomic recombination landscape of the observed allelic differences in *PRDM9* (MIM 609760).[55] The higher frequencies of the active allele reported for the West African Yoruba compared with the Eastern African Maasai might therefore imply the need for rethinking the direct correlation between LD patterns and population age.

In conclusion, Ethiopian SNP genotypes give insights into evolutionary questions on several timescales. Whether or not modern Ethiopians can be identified as the best living representatives of an ancestral human population, or even of the out-of-Africa movement, the data presented here reveal imprints of historical events that accompanied the formation of the rich cultural and genetic diversity observed in the area. Furthermore, we observe strong genetic structuring in East Africa, including a strong match between the linguistic and genetic structures. This is exemplified by the three distinct PC clusters (Omotic, Nilotic, and Semitic-Cushitic), confirming Ethiopia as one of the most diverse African regions.

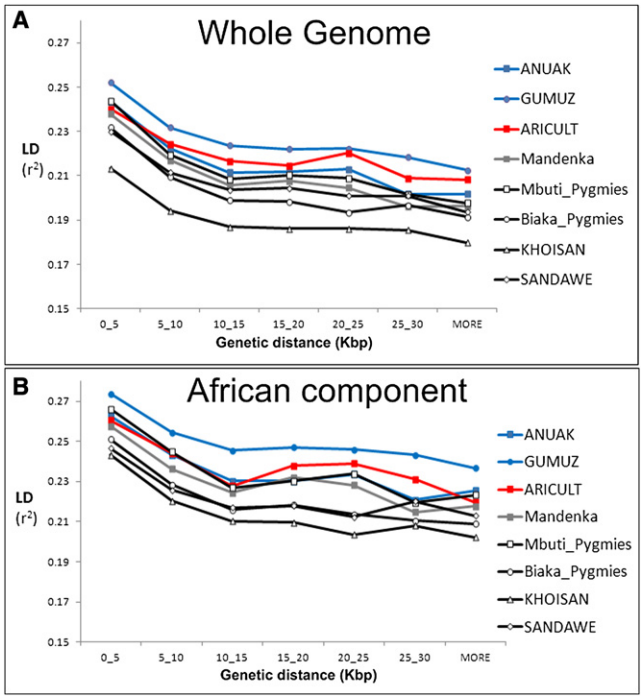

Figure 5. LD Decay over Distance
Analyses were performed with the use of 12 individuals from a set of African populations (A), including Ethiopians (red-yellow scale), west-central Africans (gray scale), and click speakers (blue scale). A modified version of the same analyses (B) was performed with the use of only ten haploid African-genome equivalents. In both cases, the Ethiopian samples show less-rapid LD decay than the other African populations in the figure.

Supplemental Data

Supplemental Data include four figures and four tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

The authors would like to acknowledge all the Ethiopian donors and collaborators, as well as Sarah Edkins, Emma Gray, Sarah Hunt, Avazeh Tashakkori Ghanbarian, and the staff at the Sanger Institute who performed the genotyping. Great help has also been provided by Priya Moorjani, David Reich, and Nick Patterson for a better understanding of the mathematics underlying the ROLLOFF approach. This work was supported by grant number 098051 from the Wellcome Trust. L.P. would like to thank the providers of a Domestic Research Scholarship, the Cambridge European Trust, and Emmanuel College, Cambridge, UK for sponsoring his research. N.B. is the settlor and senior trustee of Melford Charitable Trust and owner of Cordell Homes Ltd., which have in part funded this research. Neither N.B., the charitable trust, nor the company have any intellectual property or other rights with respect to the results of the study.

Received: March 9, 2012
Revised: May 3, 2012
Accepted: May 21, 2012
Published online: June 21, 2012

Web Resources

The URLs for data presented herein are as follows:

ADMIXTURE, http://www.genetics.ucla.edu/software/admixture/
BEAGLE, http://faculty.washington.edu/browning/beagle/beagle.html
EIGENSTRAT, http://genepath.med.harvard.edu/~reich/Software.htm
Illumina 1M Newly Generated Genotypes, http://mega.bioanth.cam.ac.uk/data/Ethiopia
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.omim.org
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/
R: Principal Component Analysis, http://stat.ethz.ch/R-manual/R-patched/library/stats/html/princomp.html

References

1. Haile-Selassie, Y. (2001). Late Miocene hominids from the Middle Awash, Ethiopia. Nature *412*, 178–181.
2. White, T.D., Asfaw, B., Beyene, Y., Haile-Selassie, Y., Lovejoy, C.O., Suwa, G., and WoldeGabriel, G. (2009). *Ardipithecus ramidus* and the paleobiology of early hominids. Science *326*, 75–86.
3. Johanson, D.C., and White, T.D. (1979). A systematic assessment of early African hominids. Science *203*, 321–330.
4. McDougall, I., Brown, F.H., and Fleagle, J.G. (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature *433*, 733–736.
5. White, T.D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G.D., Suwa, G., and Howell, F.C. (2003). Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. Nature *423*, 742–747.
6. Prugnolle, F., Manica, A., and Balloux, F. (2005). Geography predicts neutral genetic diversity of human populations. Curr. Biol. *15*, R159–R160.
7. Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W., and Cavalli-Sforza, L.L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. USA *102*, 15942–15947.
8. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.
9. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al; International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
10. 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
11. Campbell, M.C., and Tishkoff, S.A. (2010). The evolution of human genetic and phenotypic variation in Africa. Curr. Biol. *20*, R166–R173.
12. Lovell, A., Moreau, C., Yotova, V., Xiao, F., Bourgeois, S., Gehl, D., Bertranpetit, J., Schurr, E., and Labuda, D. (2005). Ethiopia: between Sub-Saharan Africa and western Eurasia. Ann. Hum. Genet. *69*, 275–287.
13. Quintana-Murci, L., Semino, O., Bandelt, H.J., Passarino, G., McElreavey, K., and Santachiara-Benerecetti, A.S. (1999).

Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat. Genet. *23*, 437–441.
14. Passarino, G., Semino, O., Quintana-Murci, L., Excoffier, L., Hammer, M., and Santachiara-Benerecetti, A.S. (1998). Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. Am. J. Hum. Genet. *62*, 420–434.
15. Poloni, E.S., Naciri, Y., Bucho, R., Niba, R., Kervaire, B., Excoffier, L., Langaney, A., and Sanchez-Mazas, A. (2009). Genetic evidence for complexity in ethnic differentiation and history in East Africa. Ann. Hum. Genet. *73*, 582–600.
16. Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J., Geberhiwot, T., Usanga, E., and Villems, R. (2004). Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am. J. Hum. Genet. *75*, 752–770.
17. Semino, O., Santachiara-Benerecetti, A.S., Falaschi, F., Cavalli-Sforza, L.L., and Underhill, P.A. (2002). Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. Am. J. Hum. Genet. *70*, 265–268.
18. Phillipson, D.W. (1998). Ancient Ethiopia. Aksum: its antecedents and successors (London: British Museum Press).
19. Pankhurst, R. (1998). The Ethiopians (Oxford: Blackwell Publishers Ltd).
20. Phillipson, D.W. (1993). The antiquity of cultivation and herding in Ethiopia. In The Archaeology of Africa: Food, Metals and Towns, T. Shaw, P. Sinclair, B. Andah, and A. Okpoko, eds. (London: Routledge), pp. 344–357.
21. Levine, D.N. (1974). Greater Ethiopia (Chicago: The University of Chicago).
22. Ehret, C. (1995). Reconstructing Proto-Afroasiatic (Proto-Afrasian): Vowels, Tone, Consonants, and Vocabulary (Berkeley, CA: University of California Press).
23. Kitchen, A., Ehret, C., Assefa, S., and Mulligan, C.J. (2009). Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proc. Biol. Sci. *276*, 2703–2710.
24. Blench, R. (2006). Archaeology, Language, and the African Past (Lanham, MD: AltaMira Press).
25. Horsfall, L.J., Zeitlyn, D., Tarekegn, A., Bekele, E., Thomas, M.G., Bradman, N., and Swallow, D.M. (2011). Prevalence of clinically relevant UGT1A alleles and haplotypes in African populations. Ann. Hum. Genet. *75*, 236–246.
26. Plaster, C.A. (2011). Variation in Y chromosome, mitochondrial DNA and labels of identity in Ethiopia. PhD thesis, University College London, London.
27. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.
28. Henn, B.M., Gignoux, C.R., Jobin, M., Granka, J.M., Macpherson, J.M., Kidd, J.M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A., et al. (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc. Natl. Acad. Sci. USA *108*, 5154–5162.
29. Behar, D.M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., et al. (2010). The genome-wide structure of the Jewish people. Nature *466*, 238–242.
30. Giannoulatou, E., Yau, C., Colella, S., Ragoussis, J., and Holmes, C.C. (2008). GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. Bioinformatics *24*, 2209–2214.
31. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
32. Cockerham, C.C., and Weir, B.S. (1986). Estimation of inbreeding parameters in stratified populations. Ann. Hum. Genet. *50*, 271–281.
33. Alexander, D.H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics *12*, 246.
34. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
35. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
36. Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84*, 210–223.
37. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*, e72.
38. Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. *68*, 978–989.
39. Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., and Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet. *7*, e1001373.
40. Liu, H., Prugnolle, F., Manica, A., and Balloux, F. (2006). A geographically explicit genetic model of worldwide human-settlement history. Am. J. Hum. Genet. *79*, 230–237.
41. Barbujani, G., and Colonna, V. (2010). Human genome diversity: frequently asked questions. Trends Genet. *26*, 285–295.
42. Fenner, J.N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. *128*, 415–423.
43. Henn, B.M., Botigué, L.R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J.K., Fadhlaoui-Zid, K., Zalloua, P.A., Moreno-Estrada, A., Bertranpetit, J., et al. (2012). Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet. *8*, e1002397.
44. Lamason, R.L., Mohideen, M.A., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynec, M.J., Mao, X., Humphreville, V.R., Humbert, J.E., et al. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science *310*, 1782–1786.
45. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al; International HapMap Consortium. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.
46. Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., and Pritchard, J.K. (2009). Signals of recent positive

selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.
47. Stokowski, R.P., Pant, P.V., Dadd, T., Fereday, A., Hinds, D.A., Jarman, C., Filsell, W., Ginger, R.S., Green, M.R., van der Ouderaa, F.J., and Cox, D.R. (2007). A genomewide association study of skin pigmentation in a South Asian population. Am. J. Hum. Genet. *81*, 1119–1132.
48. Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., Scozzari, R., Cruciani, F., Behar, D.M., Dugoujon, J.M., et al. (2006). The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science *314*, 1767–1770.
49. Underhill, P.A., and Kivisild, T. (2007). Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu. Rev. Genet. *41*, 539–564.
50. Soares, P., Alshamali, F., Pereira, J.B., Fernandes, V., Silva, N.M., Afonso, C., Costa, M.D., Musilová, E., Macaulay, V., Richards, M.B., et al. (2012). The Expansion of mtDNA Haplogroup L3 within and out of Africa. Mol. Biol. Evol. *29*, 915–927.
51. Freeman, D., and Pankhurst, A. (2003). Peripheral People: The Excluded Minorities of Ethiopia (London: Hurst and Company).
52. Kaplan, I. (1971). Area Handbook for Ethiopia (Washington, D.C.: US Government Printing Office).
53. Hansberry, W.L. (1974). Pillars in Ethiopian History (Washington, D.C.: Howard University Press).
54. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.
55. Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African Americans. Nature *476*, 170–175.

The American Journal of Human Genetics, Volume 91

Supplemental Data

Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool

Luca Pagani, Toomas Kivisild, Ayele Tarekegn, Rosemary Ekong, Chris Plaster, Irene Gallego Romero, Qasim Ayub, S. Qasim Mehdi, Mark G. Thomas, Donata Luiselli, Endashaw Bekele, Neil Bradman, David J. Balding, and Chris Tyler-Smith

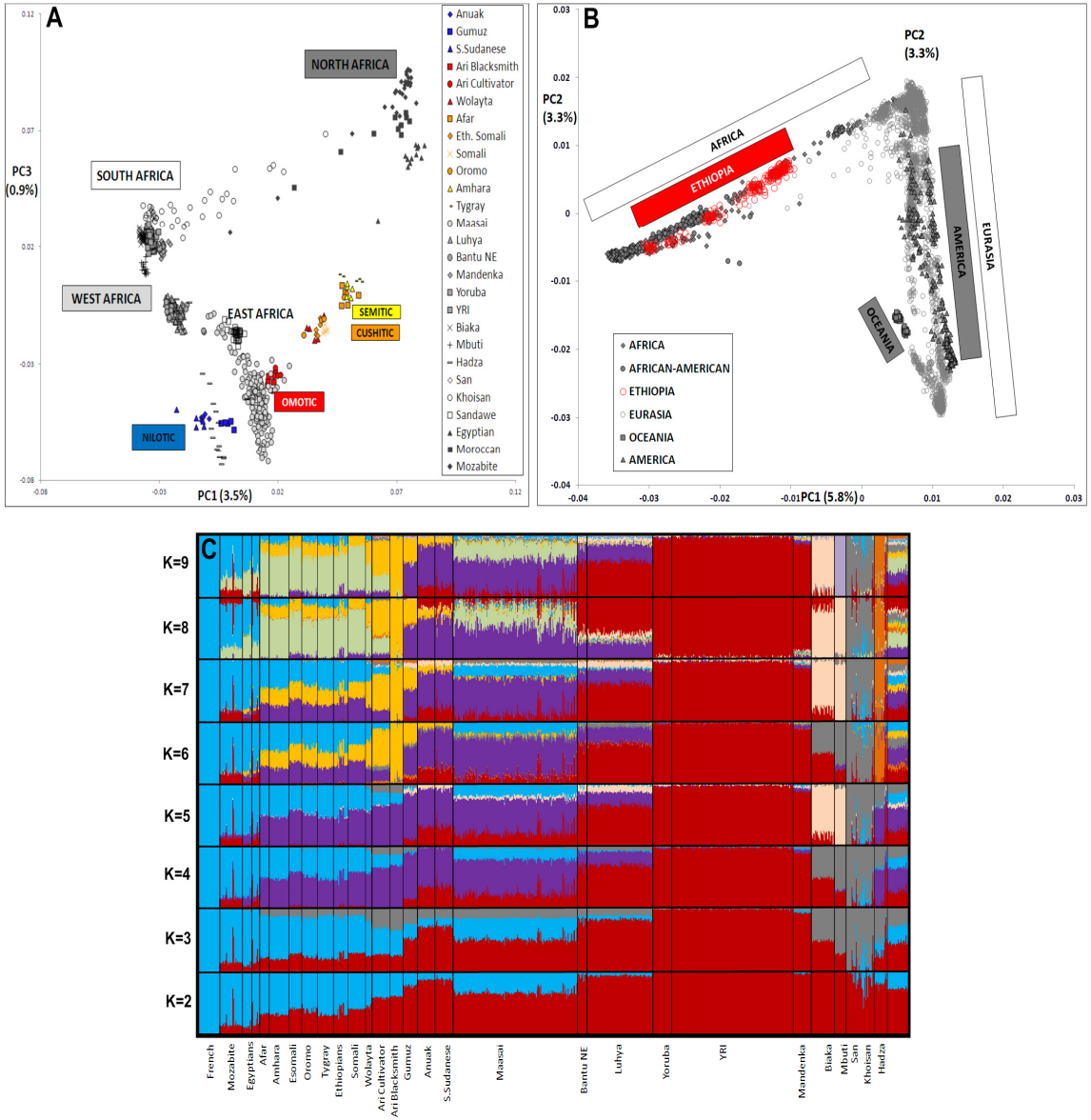

Figure S1. Alternative PCA and STRUCTURE-like Plots

The 1st and 3rd component of the PCA based on the full African panel using only 5 individuals per Ethiopian population is shown in **A.** The African samples are viewed in a worldwide context in **B**, where the first 2 components show the Ethiopian samples (in red) spanning most of the African diversity. **C** shows a series of ADMIXTURE plots using values of K ranging between 2 and 9.

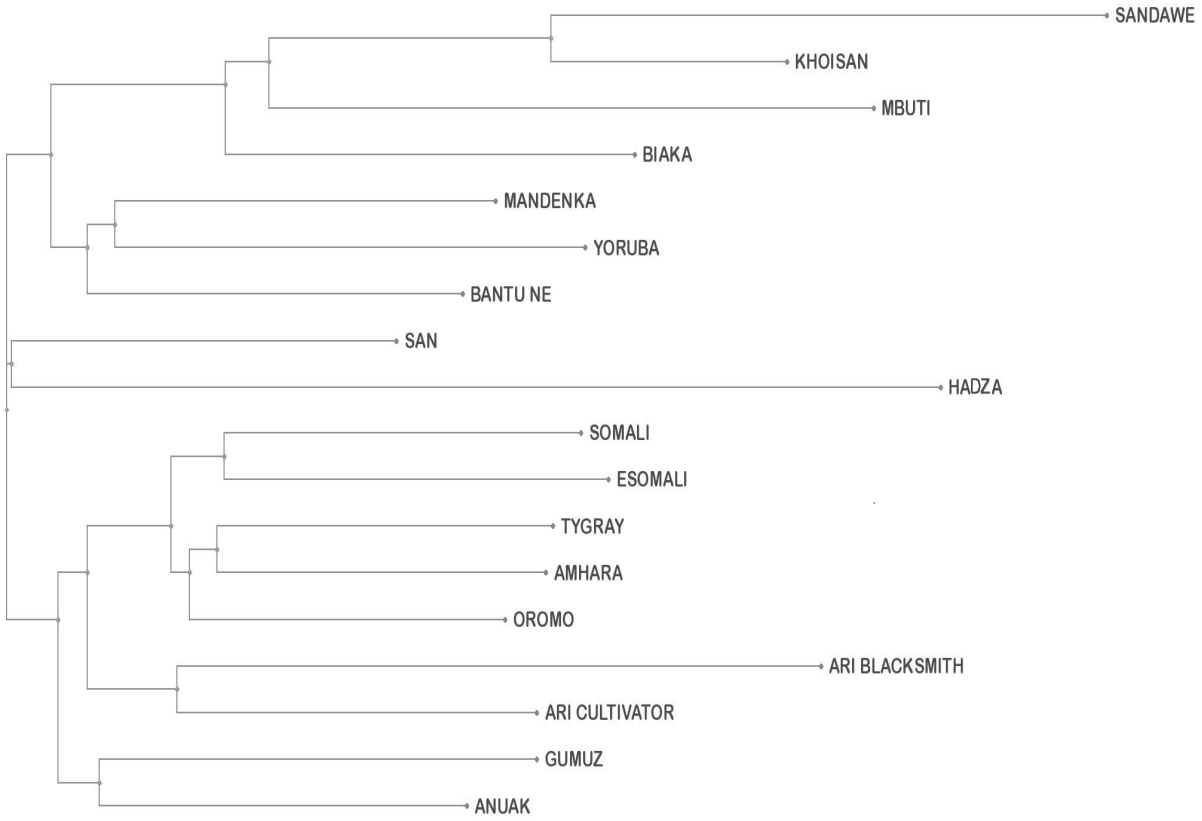

Figure S2. Neighbor-Joining Tree

The genetic distance (Fst) calculated on the African component only was used to draw a neighbor-joining tree of the studied Sub-Saharan and Ethiopian populations.

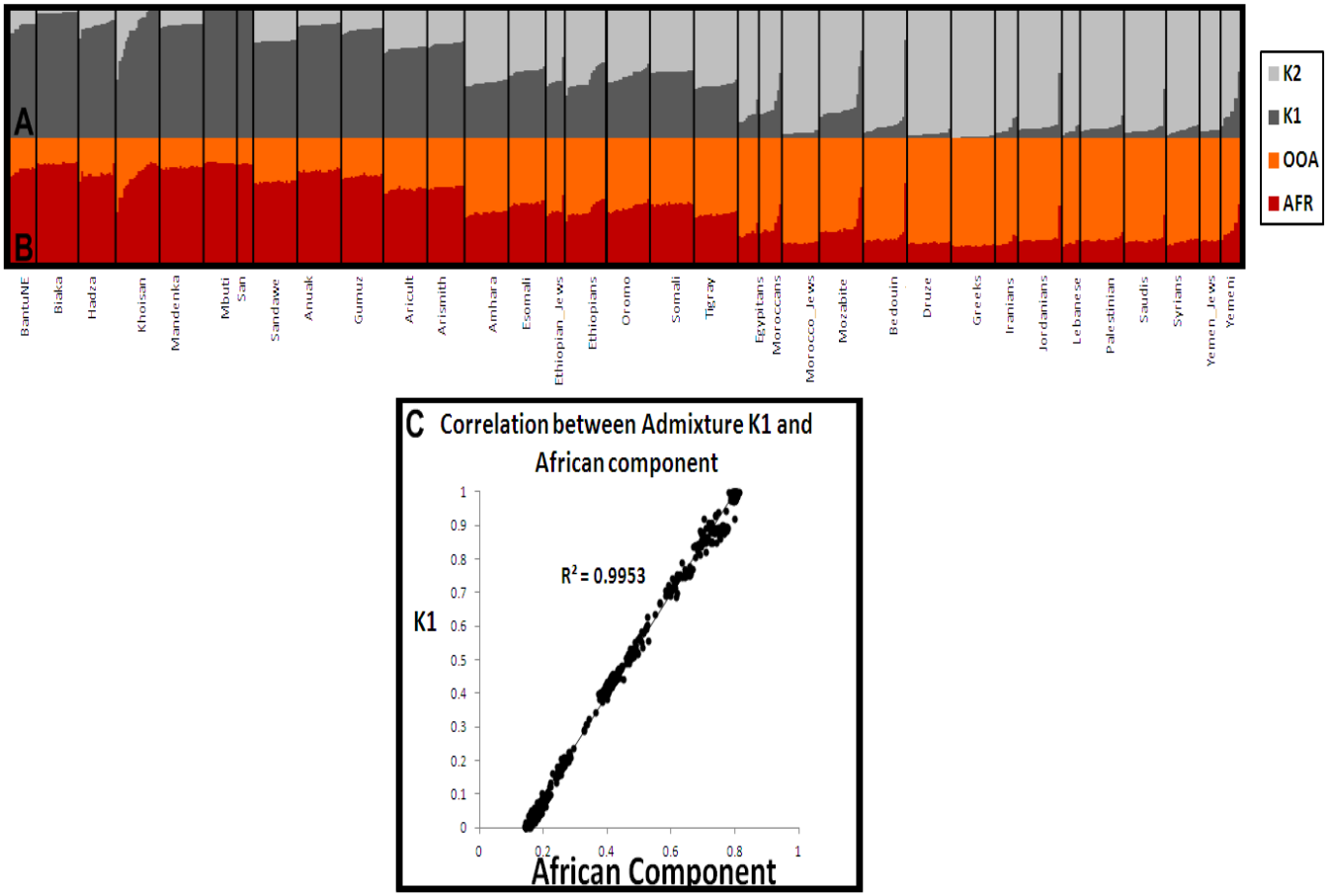

Figure S3. Comparison between the PCA-Based Genome Partitioning and the ADMIXTURE Results

A maximum of 20 pruned samples were processed in ADMIXTURE with K=2 (**A**) and PCA-based genome partitioning (**B**). The correlation (r^2) between the K1 and the African components from these plots is displayed in **C.**

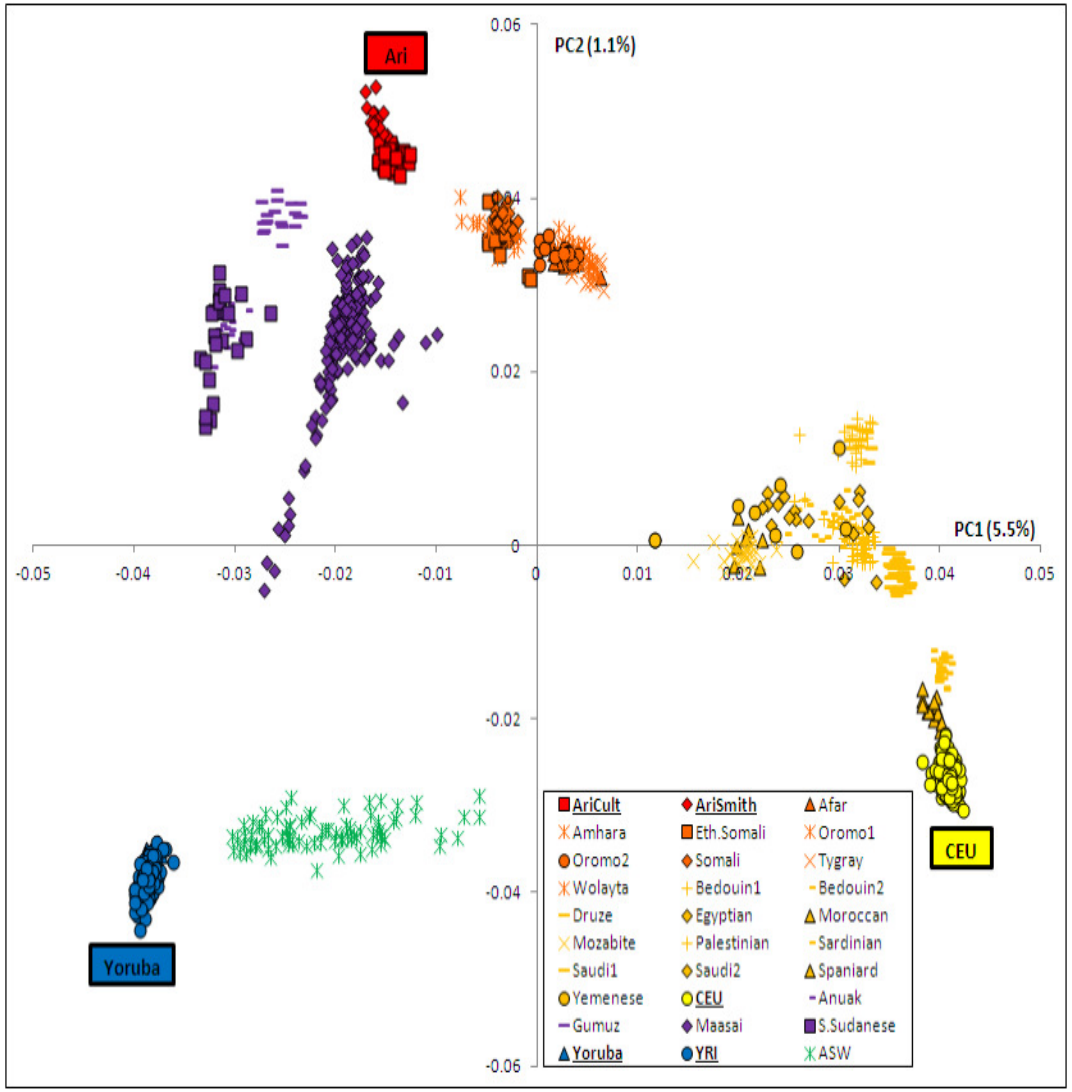

Figure S4. PCA Plot of African, West Asian, and European Samples

The Yoruba (blue), CEU (Yellow) and Ari (Red) define a triangular shape with the other samples analyzed (colored according to their position between the three primary colors) distributed on its three sides.

Table S1. Sample Size, Location, and Sociological Features of the Genotyped Populations

Pop	N	Lat	Long	Elev	Geo. Location	Ling. Group	Language	1998 Census	Endo/ Exogamous	Patri/ Matrilocal	Mono/ Polygamous	Patri/ Matrilineal	High/ Lowland	Food Production
Afar	12	12	41	379	Wag Hemra Zone	Cushitic	Xamtan	143369	Endo	Patrilocal	Poly	Patrilineal	Highland	Agriculturalist
Amhara	26	10	39	2088	Amhara Region	Semitic	Amharic	17372913	Endo	Patrilocal	Mono	Patrilineal	Highland	Agriculturalist
Anuak	23	8	34	500	Gambella	Nilotic	Anuak	45646	Endo	Patrilocal	Poly	Patrilineal	Lowland	Mixed Farming
Ari Blacksmith	17	6	37	1348	South Omo	Omotic	Ari	158857	Endo	Patrilocal	Poly	Patrilineal	Highland	Agriculturalist
Ari Cultivator	24	6	37	1348	South Omo	Omotic	Ari	158857	Endo	Patrilocal	Poly	Patrilineal	Highland	Agriculturalist
Ethiopian Somali	17	9	42	1543	Somali Region	Cushitic	Somali	3334113	Endo	Patrilocal	Poly	Patrilineal	Lowland	Pastoralist
Gumuz	19	NA	NA	NA	Beni-Shangul Gumuz	Nilo-Saharan	Gumuz	120424	Endo	Patrilocal	Poly	Patrilineal	Lowland	Pastoralist
Oromo	21	8	37	1758	Oromia Region	Cushitic	Oromo	Ca. 17000000	Endo	Patrilocal	Mono/Poly	Patrilineal	Highland	Agriculturalist/Mixed Farming/Pastoralist
Somali	23	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
South Sudanese	24	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Tigray	21	9	38	1696	Tigray Region	Semitic	Tigrayan	3224875	Endo	Patrilocal	Mono	Patrilineal	Highland	Agriculturalist
Wolayta	8	6	37	1737	Wolayta Zone	Omotic	Wolayta	1231673	Endo	Patrilocal	Mono	Patrilineal	Highland	Agriculturalist

Table S2. Genomic Average Heterozygosity and F_{ST}

Calculated using the same pruned SNPs as for the genome partitioning. The values here formed the basis of the heat maps in Figure 2.

(This table is available as a supplemental Excel file)

Table S3. Minimum Pairwise Genetic Distance in Whole Genome (A) and Non-African Component Only (B) between Ethiopians and Surrounding Populations to Compare with the F_{ST} Results Reported in Figure 3

A	Whole Genome	Bedouin	Druze	Egyptian	Greek	Iranian	Jordanian	Lebanese	Moroccan	Mozabite	Palestinian	Saudi	Syrian	Yemeni
	AMHARA	0.0382	0.0383	0.0381	0.0387	0.0387	0.0384	0.0383	0.0380	0.0387	0.0385	0.0381	0.0384	0.0383
	ESOMALI	0.0380	0.0383	0.0380	0.0385	0.0385	0.0382	0.0381	0.0377	0.0387	0.0383	0.0381	0.0384	0.0381
	OROMO	0.0386	0.0388	0.0383	0.0391	0.0390	0.0388	0.0387	0.0380	0.0389	0.0388	0.0386	0.0389	0.0385
	SOMALI	0.0390	0.0391	0.0384	0.0394	0.0394	0.0391	0.0389	0.0380	0.0391	0.0391	0.0389	0.0393	0.0388
	TYGRAY	0.0377	0.0379	0.0380	0.0382	0.0382	0.0379	0.0377	0.0380	0.0386	0.0378	0.0377	0.0380	0.0380
	AVERAGE	0.0383	0.0385	**0.0381**	0.0388	0.0388	0.0385	0.0383	0.0380	0.0388	0.0385	0.0383	0.0386	**0.0383**

B	Non-African Component Only	Bedouin	Druze	Egyptian	Greek	Iranian	Jordanian	Lebanese	Moroccan	Mozabite	Palestinian	Saudi	Syrian	Yemeni
	AMHARA	0.0373	0.0374	0.0372	0.0375	0.0376	0.0373	0.0373	0.0377	0.0377	0.0374	0.0372	0.0373	0.0375
	ESOMALI	0.0377	0.0379	0.0376	0.0380	0.0380	0.0377	0.0378	0.0381	0.0381	0.0378	0.0377	0.0379	0.0378
	OROMO	0.0377	0.0378	0.0375	0.0379	0.0380	0.0378	0.0377	0.0380	0.0379	0.0377	0.0376	0.0378	0.0379
	SOMALI	0.0379	0.0381	0.0377	0.0383	0.0383	0.0380	0.0380	0.0381	0.0382	0.0379	0.0379	0.0381	0.0381
	TYGRAY	0.0370	0.0373	0.0368	0.0373	0.0374	0.0372	0.0371	0.0375	0.0375	0.0371	0.0370	0.0372	0.0373
	AVERAGE	0.0375	0.0377	**0.0373**	0.0378	0.0379	0.0376	0.0376	0.0379	0.0379	0.0376	0.0375	0.0376	**0.0377**

Table S4. 40-SNP Windows Showing an Outlier Number of Non-African Chromosomes in the Majority of the Semitic-Cushitic and Ari Blacksmith Populations

	Z-scores bins for the proportion of European haplotypes in each 40 SNP window							
Number of populations showing a given Z score per given region	**-3.5**	**-2.5**	**-1.5**	**-0.5**	**0.5**	**1.5**	**2.5**	**3.5**
1	9	87	362	469	487	384	82	0
2	0	29	217	476	454	231	25	0
3	**1**	**5**	103	348	332	116	**8**	**0**
4	**0**	**4**	51	132	139	42	**6**	**0**
5	**0**	**0**	6	37	24	16	**0**	**0**

Genomic regions showing deficiency (light grey) or excess (dark grey) of non-African component

Chr	**Start**	**End**	**Z-score**
19	61523852	62240101	**-3**
1	202611308	203863493	**-2.5**
1	233121260	234561036	**-2.5**
2	159910201	162832297	**-2.5**
12	114756940	115796582	**-2.5**
13	62291269	65337535	**-2.5**
15	27642703	29438608	**-2.5**
15	31797258	32775283	**-2.5**
17	18231718	21377174	**-2.5**
18	14930293	19200794	**-2.5**
1	238073408	238678237	**2.5**
5	149063400	149839428	**2.5**
5	149875685	150607133	**2.5**
6	154486907	155207161	**2.5**
Chr	**Start**	**End**	**Z-score**
7	95305557	96856821	**2.5**

7	155434633	155960870	**2.5**	
8	132605749	133525623	**2.5**	
11	102147471	103368144	**2.5**	
12	123717947	124441419	**2.5**	
14	21063338	21641738	**2.5**	
15	45473882	47025722	**2.5**	****includes SLC24A****
15	51368181	52099177	**2.5**	
15	68167541	69443763	**2.5**	
19	1290058	2416737	**2.5**	

The number of regions showing a given Z-score in the specified number of populations is reported. Regions showing Z-scores <-2 (light grey) or >2 (dark grey) in at least 3 populations were considered outlier. The genomic coordinates on the highlighted windows are reported below.

2.6 Notes to the published paper

This paragraph is provided to expand on some aspects of the above reported paper, in order to reconcile the succinct style of a scientific publication with the more descriptive one of a thesis.

2.6.1 Genome partitioning and mosaic genomes

All the Illumina 1M Omni SNPs overlapping a panel of reference populations were pruned three subsequent times with PLINK (Purcell et al. 2007) using standard pruning parameters (Alexander et al. 2009) on the LD patterns of CHB, CEU and YRI respectively. The pruned SNPs were then divided into 40 SNPs chunks for a total of ~2000 chunks across the whole genomes. Each chunk of both Ethiopians and reference populations was phased independently using PHASE (Stephens et al. 2001) and each phased Ethiopian haplotype compared with a set of reference YRI or CEU haplotypes using an in house version of PCAdmix (Brisbin et al. 2012), as described in the published paper. The available Ethiopian haplotypes of each 40 SNPs chunk were therefore divided into African and non-Africans, depending on their similarity with the YRI or CEU reference sets. In order to create the mosaic genomes to be used in the downstream analyses, ten of the African or non-African haplotypes were selected for each chunk. The ten African haplotypes chosen for every chunk formed, together, 10 haploid African genomes. Similarly, the ten non-African genomes selected for each chunk, when pasted together, formed 10 haploid on-African genomes.

2.6.2 Admixture plot

To facilitate the reading of the population labels, as well as providing the linguistic groups of the Ethiopian populations, the Admixture plot represented in Figure 1C of the published paper is here reproduced in its full size (Figure 2.3)

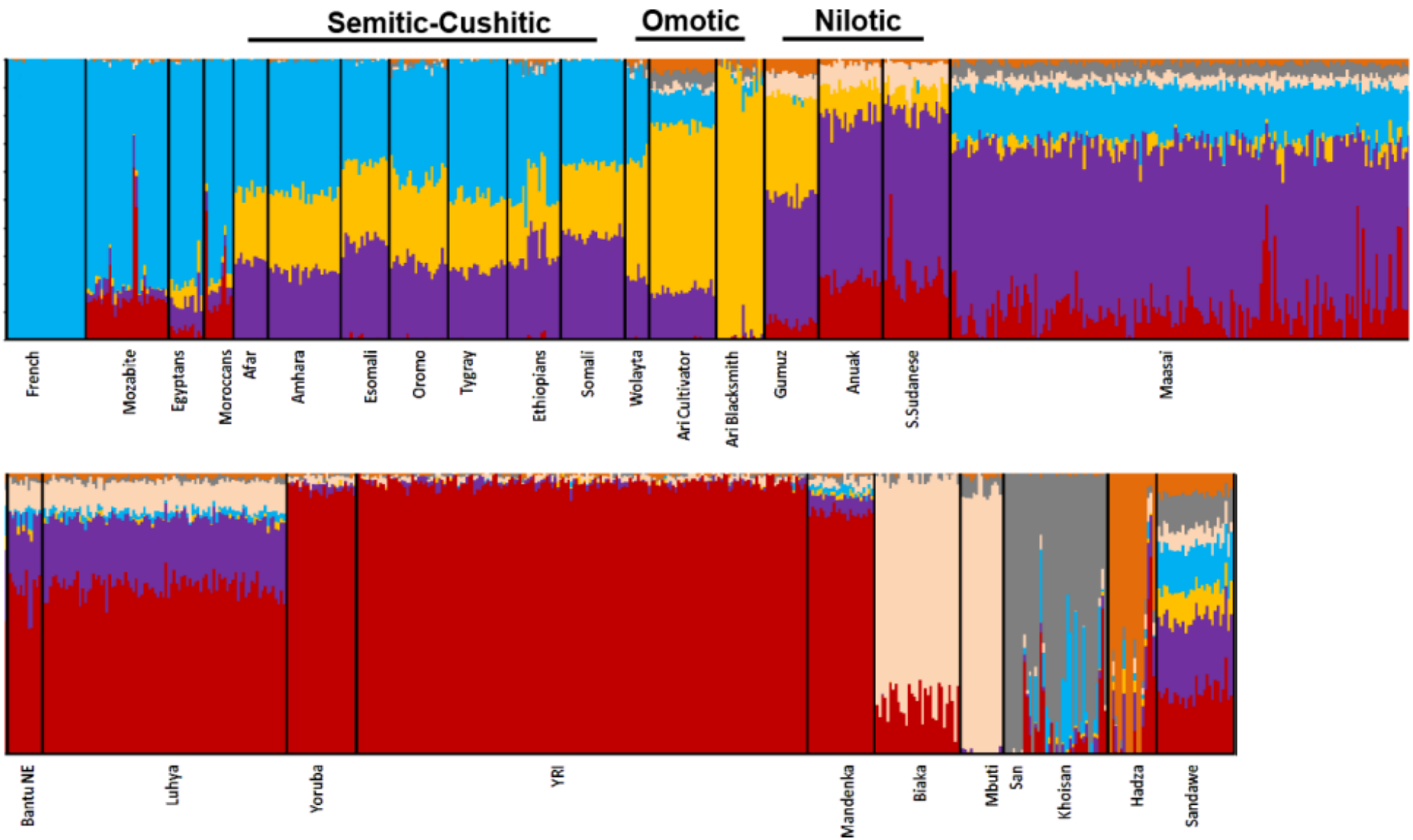

Figure 2.3 Detailed view of Admixture plot from SNPchip paper

2.6.3 Admixture date estimate using Rolloff on different ancestry sources

The Table 1 reproduced in the published paper reports the estimated admixture dates for the analysed Ethiopian populations, using as proxy for the ancestral sources either YRI and CEU or Ari and CEU. Unexpectedly, the ROLLOFF estimates for the date of admixture differ for up to 900 years, depending on which pair of source populations is used. In order to compare the ROLLOFF exponential declines in both cases, the results for Tygray, one of the populations for which the two dating estimates are most divergent, are reported in Figure 2.4. From the figure it appears how, when Ari is used as proxy for the African source, the exponential decay starts from lower values than when YRI is used. The difference in dating is therefore supported by the observe data and it is not an artefact of the fitting procedure. A possible explanation for this would be that the true African source of the Tygray populations was somewhat more closely related with the Yoruba than with the Ari or, more likely, that since the admixture events Ari underwent some sort of genetic drift that decreased their performance as source population.

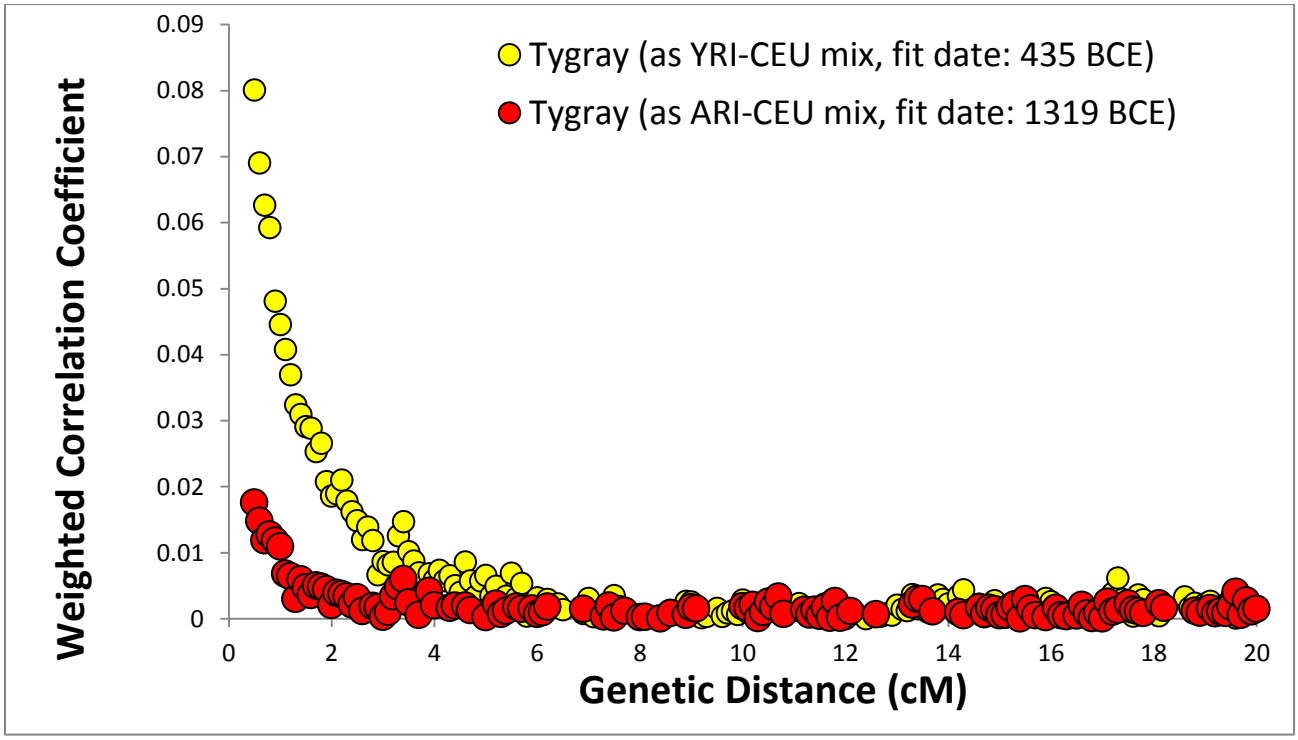

Figure 2.4 ROLLOFF exponential decay for Tygray using either Ari or YRI as African source population

2.6.4 Ancestry related traits and admixture

While the finding of SLC24A5 among the set of windows showing a dramatic excess of non-African haplotypes within the Ethiopian samples was thoroughly discussed in the paper, this ancestry-biased selection might have affected other class of genes. However no other skin pigmentation associated gene, nor any ancestry markers such as HLA, FY or any of the most highly differentiated genes between Africans and non-Africans (The 1000 Genomes Project Consortium et al. 2012)was found within these regions.

2.6.5 LD decay and age of populations

In order to integrate the results reported by Henn and colleagues (Henn et al. 2011) on the putative origin of all human populations, the same LD based analyses was performed on the newly generated Ethiopian samples. However, as already discussed in the Introduction of this thesis, the decay of LD over genetic distance can only provide information on the long term effective population size of a population. Any conclusion drawn from these results on the age of a population, as well as the concept itself of “age” applied to a population, must be considered as speculations. Furthermore, as stated in the paper, differences in allelic variants of PRDM9 or RNF212 across the population considered, might affect the conclusion of any LD based analysis. Specific variants of the above mentioned genes have indeed been reported to alter the overall recombination structure of the genome (Hinch et al. 2011), with subsequent effect on the LD structure on the populations where these are found.

3. Whole-genome sequencing of 125 Ethiopian samples

To generate unbiased genomic information for a representative set of the populations described in chapter 2, the whole genome sequences of 125 Ethiopian samples were generated and the preliminary results are described in this chapter. The rationale for generating whole genome sequences follows the need of better describing the selected populations through the discovering of new variants of all classes, including structural rearrangements. The availability of low frequency and less biased genetic markers would indeed allow a deeper exploration of the Ethiopian diversity. Particularly, overcoming the SNPchip ascertainment bias can provide reliable estimation of the Ethiopian heterozigosity, in the context of other worldwide populations, as well as refining their pairwise distances. The availability of a comprehensive set of low frequency markers (namely doubletons), could instead inform us on finer scale patterns of allele sharing between different populations. My contribution to the results described in this chapter involved the choice of the populations to be re-sampled, the participation to the collection and extraction campaign, and the design and execution of all the described analyses, with the exception of the PSMC which was run by Dr. Stephan Shiffels of the WTSI. The purified DNA samples were quantified and sequenced by the staff of the sequencing team of the WTSI, while the SNP calling and genotype refinement were performed by Dr. John Maslen and Dr. Petr Danecek of the WTSI.

3.1 Identification of five Ethiopian populations to be sampled for re-sequencing

A campaign to collect further samples in Ethiopia was organized in parallel with the processing and the analyses of the DNA extracted from the buccal swabs that led to the results described in the past chapter. The rationale for collecting new samples was that, in order to generate whole genome sequences using the Illumina HiSeq platform, the starting DNA quantity had to be of the order of 5 µg. Furthermore, as next generation sequencing technologies do not discriminate between human and exogenous DNA the presence of non-human DNA (i.e. from bacteria or other symbiotic organisms likely to be present in the buccal swabs) has to be kept to a minimum. As a consequence the collection of new blood samples from Ethiopian donors was necessary for the whole-genome sequencing step.

Following the assessment of genetic diversity among Ethiopian populations on the basis of the SNP array data (Chapter 2), five populations: Amhara, Oromo, Ethiopian Somali, Gumuz and Wolayta, were chosen to represent most efficiently the known genetic diversity in Ethiopians. Figure 1 of the Pagani et al. 2012 paper, reproduced in Chapter 2, summarizes the observed

patterns of diversity among Ethiopian populations classified by their linguistic affiliation: Amhara and Oromo were included as representative of the Semitic and Cushitic groups, respectively, and chosen before other populations from the same groups because, together, they account for up to 60% of the total Ethiopian census population size. Ethiopian Somali, who belong to the Cushitic group like Oromo, were included because they formed a separate cluster in the PCA (Figure 1A), and showed a lower contribution of the non-African component than other Cushitic populations in the ADMIXTURE analyses (Figure 1 C). Of the two Nilotic-speaking populations, Anuak and Gumuz, that were analysed using genome-wide genotype data, Gumuz were chosen to represent this distinct linguistic cluster (Figure 1A) because they showed a lower proportion of the "red" ancestry component which is associated with the Bantu-speaking populations and a higher proportion of the "yellow" component specific to East African populations (Figure 1C). The red Omotic cluster of Figure 1A is formed by Ari and Wolayta. Wolayta were chosen to represent this cluster because of uncertainty about the proper labelling of the Blacksmith samples (as described in Paragraph 2.3) at the time when the populations were chosen. Furthermore, the scattered nature of the Ari Blacksmiths (only few blacksmiths live in each Ari village) would have affected the sampling strategy and made sampling considerably longer and more laborious.

During the fieldwork in the summer of 2011, 10ml blood from peripheral veins was collected from 343 anonymous volunteers (participating with written informed consent) from the five aforementioned ethnic groups. In addition to the collection of blood for DNA extraction, a number of phenotypic measurements as well as sociological information were taken from the participants. The sociological interview conducted with each participant, and available in Appendix 5 as a blank form, was aimed at comparing the self-reported ethnicity of the donors with the ones declared for their relatives up the grandparents level. Furthermore, this interview provided useful information about the birthplace and lifestyle of the donor. The phenotypes collected (a list of which is provided in Appendix 5 as a blank form) included quantitative skin pigmentation measurement using a DSM II ColourMeter, blood oxygenation, lung capacity, heart rate, blood pressure, cranio-facial measurements, body shape, height, weight and electric conductance.

3.2 Fieldwork in Ethiopia

Before sample collection could start, applications to the UCL Research Ethical Committee (REC) and to the Addis Ababa equivalent were submitted to seek for approval to sample up to

500 Ethiopians from ethnic groups that would be chosen later. While the consents were being granted (UCL REC approval number 0489/002 and Addis Ababa University REC approval number 310/538/04), the preliminary results available from the SNP arrays informed the choice of the five populations to be collected: Amhara, Oromo, Ethiopian Somali, Wolayta and Gumuz (Section 3.1). While Dr. Bradman and Dr. Rosemary Ekong from UCL took care of ordering and shipping all the consumables to the University of Addis Ababa, Prof. Endashaw Bekele was involved in recruiting a team of three Ethiopian researchers (in addition to himself and Dr. Ekong) and a trained nurse to actually perform the collection in the field. My role was to adapt to the facilities available in Prof. Bekele's lab a set of existing protocols for the collection of blood and phenotypes and for the extraction of the DNA from blood during my visit at the University of Addis Ababa between 16th and the 28th of June 2011. In addition I was also involved in the training and supervision of the four people who would perform the collection and extraction of the samples in order to make them independent throughout the collection campaign.

In order to collect blood and sociological and phenotype information, keeping the time devoted to each donor to a minimum and making the best use of the training of the staff, three stations were set up. In the first station, after having read and signed the informed consent form reproduced in Appendix 5, the donors had their skin colour, heart pulse, blood pressure, blood oxygenation and lung capacity measured. To control for intra-individual variation of skin pigmentation, the skin colour was measured from three body parts (once on the sun exposed cheek and twice on the non exposed left and right inner arms) using DSM II ColourMeter. To monitor the response to a small physical exercise, systolic and diastolic pressure, pulse, blood oxygenation and pulmonary capacity were measure before and after six minutes of walk. Particularly, blood pressure and pulse were measured using an electronic sphygmomanometer, while blood oxygenation and pulmonary capacity were measured using MIR Spirobank. This instrument is composed of a finger clip that infers the blood oxygenation (SpO_2) from the reflectance of the capillary bed, and a disposable turbine and mouthpiece that measures the exhaled air flux for the instrument to compute the subject's pulmonary capacity. The second station was devoted to the measurement of body shape and physical characteristics. The height and weight of the donors were measured using a scale and a rigid meter commonly used in the medical ambulatories. In addition, the waist and hips circumferences were measured with a tape meter, while the craniofacial dimensions, right and left subscapular, suprailiac, biceps and triceps skinfolds were measured using anthropometric callipers. The body electric conductance was assessed using Bodystat 1500MDD. This instrument infers the proportion of lean and fat

mass as well as total water composition by estimating the total body impedance. The instrument can also provide an estimate of the nutritional state of the subject upon input of the age, weight and height. Furthermore, to accommodate for the longer times required by the first station, the donors were asked to fill in the sociological questionnaire at this stage, to improve the overall flow across the three stations. To assess the error that could have stemmed from the reading of the most problematic devices such as meters and callipers, the person in charge of such measurements and whom carried out all the measurements throughout the collection campaign, assessed the same person three times. These readings, reported in Table 3.1, show the error that might affect the phenotypes collected on analogical devices. The same error assessment was not performed for those phenotypes that relied on digital devices, since the reading error is virtually absent there. The third and final station involved the blood drawing alone, which was carried out by a qualified nurse.

The first donors that were gathered both for training and collection purposes were mostly Prof. Bekele's students who showed a particular interest in our project. The donors were first asked to read the information sheet and to sign the informed consent and briefed on the various instruments and collection procedures. The DNA from the samples thus collected was then extracted to test the efficiency of the machinery and protocols. With the successful collection and extraction of the first blood samples and phenotypes, the training period reached an end. The team set off for a two month campaign of blood and phenotype collection in the homeland regions of the Amhara, Oromo, Ethiopian Somali, Wolayta and Gumuz people who would then be studied for the remainder of the project

Table 3.1 Reading errors on the analogical devices. For each of the measurements involving tape measures or callipers, the operator took three measurements on the same subject. The maximum discrepancy between the three measurements is reported in the last column.

Phenotype	1st Measure (cm)	2nd Measure (cm)	3rd Measure (cm)	Max Error (cm)
Cranial length	19.8	20.3	20	0.5
Cranial width	15.5	15.8	15.7	0.3
Facial length	12.2	11.1	11.9	1.1
Facial width	13.4	13.8	13.8	0.4
Left bicep skinfold	0.7	0.7	0.6	0.1
Left tricep skinfold	1.7	1.6	1.6	0.1
Left scapular skinfold	0.9	0.9	0.9	0.0
Left iliac skinfold	2.0	2.1	1.8	0.3
Left arm circumference	29.5	29	29	0.5
Waist circumference	87.3	88	89.5	2.2
Hips circumference	103	102	103	1.0

3.3 Whole genome sequencing

3.3.1 Sample processing

With the successful conclusion of the sampling campaign in Ethiopia, 343 purified DNA samples were shipped to the Wellcome Trust Sanger Institute (WTSI). There, an initial screening on the sociological record was performed, to shortlist only those samples whose reported ethnicity matched the reported ethnicity of both parents, maternal grandmother and paternal grandfather. The samples that met these criteria and belonged to the five target populations were submitted for quality checks where the quantity, concentration and fragmentation of the available DNA were assessed. A total of 125 samples, 25 from each population, that passed these filters were submitted for whole genome sequencing on an Illumina HiSeq platform with a 100bp paired-end reads on a median library insert size of 350 bp. Of these, 120 were processed at low depth (8x) while five, one from each group, were sequenced at high depth (30x). The average number of reads per each locus achieved for the high depth samples was 36, 27, 29, 29 and 32 for the Amhara, Gumuz, Eth. Somali, Oromo and Wolayta samples respectively. The 120 low depth samples had instead an average number of reads that ranged between 4.0 and 10.9, with a median across samples of 7.5. Additionally, 200 samples that passed the quality checks, including the 125 that were chosen for sequencing, were processed on an Illumina Omni 2.5 M SNP array to provide a set of reference calls for the sequenced samples and additional projects.

3.3.2 Variant calling

The single nucleotide and insertion/deletion variants of the high-depth samples were called applying standard pipelines available at the WTSI to the five samples alone, exploiting the high confidence achieved through the high number of reads available at each genomic site. A different calling approach was applied to the low-depth samples to compensate for the lower confidence due to the smaller number of reads available per each sample. The newly generated 120 low depth samples were pooled together with 850 samples from the Phase1 of the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2012) as reported in Table 3.2 and the variants were called in the global set of samples using pipelines developed for the 1000 Genomes Project (See Appendix 6). Furthermore, to provide a low-depth comparison with the high-depth samples, two thirds of the reads of the high-depth samples (for Amhara, Oromo, Ethiopian Somali and Gumuz) were randomly removed to create an *in silico* set of samples of approximately 10x sequencing depth. The Wolayta sample was instead sequenced at high and low depth independently, so there was no need to use the *in silico* reduction for this sample.

This set of reduced-depth reads was pooled together with the rest of the low-depth samples to benefit from the same calling process. The total number of SNPs and Insertion/Deletions (Indels) detected from the 5 high depth samples and from the 120 Ethiopian and 850 1000 Genomes Project low depth samples is reported in Table 3.2.

Table 3.2 Samples used to call variants in the low-depth set. Ethiopian populations, in bold, were each composed of 24 low-depth and one high-depth sample that was sub-sampled *in silico* to low coverage.

Sample	**N**
African ancestry in Sout West USA (ASW)	49
Utah residents (CEPH) with North and West European ancestry (CEU)	91
Han Chinese in Beijing, China (CHB)	83
Han Chinese South (CHS)	92
Colombian in Medellin, Colombia (CLM)	55
Gujarati Indian in Houston, Texas (GIH)	13
Iberian populations in Spain (IBS)	49
Luhya in Webuye, Kenya (LWK)	83
Mexican ancestry in Los Angeles, California (MXL)	58
Peruvian in Lima, Peru (PEL)	29
Puerto Rican in Puerto Rico (PUR)	64
Tuscan from Italy (TSI)	100
Yoruba in Ibadan, Nigeria(YRI)	84
Amhara	**24+1**
Ethiopian Somali	**24+1**
Gumuz	**24+1**
Oromo	**24+1**
Wolayta	**24+1**

Despite the higher resolution provided by the high depth data, the number of samples processed at low depth was almost 200 times higher. The total number of variants found in the low depth samples was therefore expected to be higher than the ones found in the high depth ones. Interestingly the ratios between multiallelic variants detected in low and high depth samples (Table 3.3) are much lower than the ratios obtained for the biallelic variants. This relative enrichment of multiallelic variants in the high depth samples might be consequence of the increased resolution provided by the higher number of reads per each sample. However,

due to the documented difficulty in calling multiallelic variants from re-sequencing data (The 1000 Genomes Project Consortium et al. 2012), a more plausible explanation could be that the higher number of samples available for the low depth data decrease the otherwise high false discovery rate affecting these type of variants.

Table 3.3 Number of variants detected in the high and low depth sequencing datasets and low to high enrichment ratio.

Variant type	N in 120+850 low depth samples	N in 5 High depth samples	Low/High ratio
Multiallelic SNP	14,476	13,019	1.11
Multiallelic Indel	144,952	205,065	0.71
Biallelic Indel	3,097,988	1,039,808	2.98
Biallelic SNP	38,183,772	8,629,166	4.42

To provide a quality control for the biallelic indels detected, these were divided by length classes in both the high and low depth datasets, and their relative proportions plotted in Figure 3.1. With the exception of a small excess of 4bp long indels, the decay of number of indels with the increase in length fits the expected trend for human populations (The 1000 Genomes Project Consortium et al. 2012). The raw file processing and variable positions calling were performed by Dr. John Maslen and Dr. Petr Danecek respectively, both from the WTSI.

To assess the genotype calling accuracy in the low depth samples, a subset of the genotypes obtained for the latter were compared with the ones obtained from the same samples typed on the Illumina 2.5 M SNP array. Of the 620K genotypes examined in each of the 120 samples, 93.21% matched between the low depth sequencing and SNP array genotyping results. This matching rate, obtained before phasing and imputation improvements, is slightly better than the 91% reported for the 1000 genomes samples at a comparable stage, perhaps due to the higher depth (~8x as opposed to ~6x) used for the Ethiopian samples. A representation of the concordance rate for each minor allele frequency SNP category is reported in Figure 3.2. Further orthogonal validation procedures will be needed to assess the false discovery rate (FDR) of the sequenced samples.

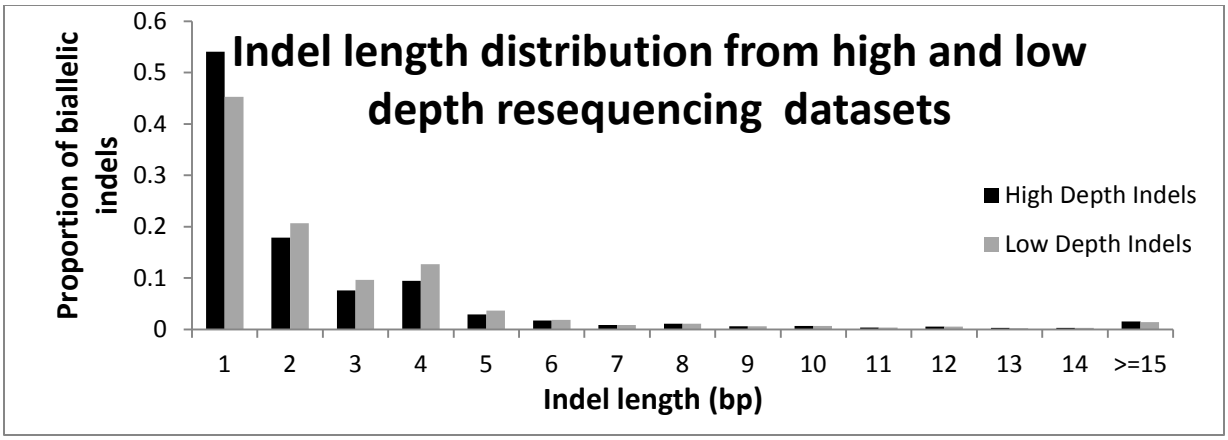

Figure 3.1 Length distribution of the biallelic Indels in high and low depth datasets. The biallelic indels detected for the high (black bars) and low (grey bars) datasets were divided into length classes and their relative proportions plotted.

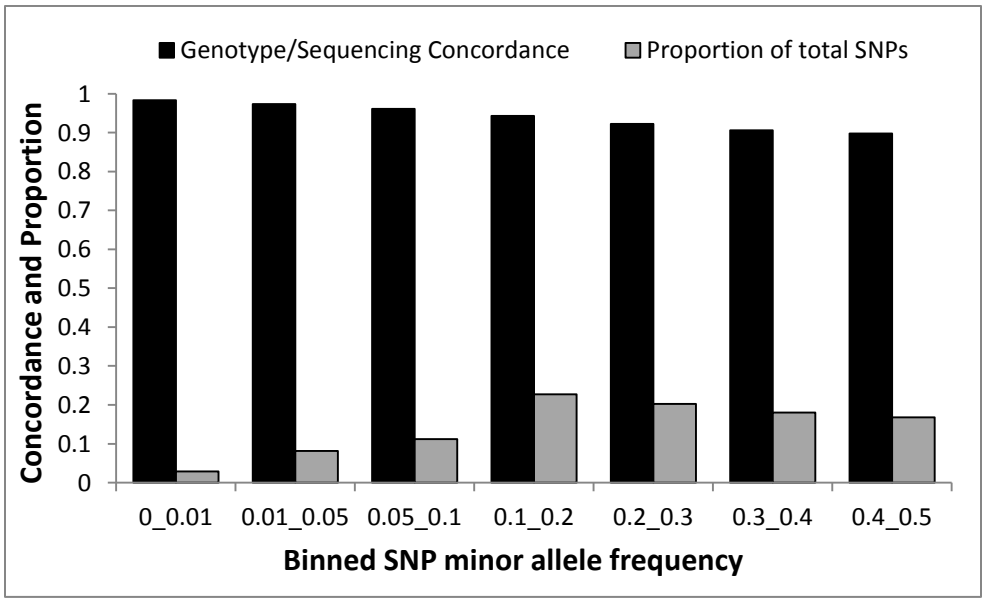

Figure 3.2 Concordance rate of SNP calling between low coverage samples and Illumina 2.5M genotypes. Black bars show, per each frequency class, the proportion of matching genotype calls for the 120 samples sequences at low coverage (8x) and the same samples typed on Illumina 2.5M Omni. Grey bars show the proportion of total SNPs in each frequency class.

3.3.3 Summary statistics from whole genome sequences

To provide an initial view of the genetic diversity in the low-depth Ethiopian samples, some summary statistics were calculated and the results compared with the 1000 Genomes Project samples. The ancestral alignments released by the 1000 Genomes Project were used to determine the ancestral state of each called SNP. The total heterozygosity, median proportion of heterozygous, ancestral homozygous and derived homozygous and total number of variable sites per population are reported in Table 3.4. The total number of derived and proportion of homo and heterozygous sites for each Ethiopian population are similar to the ones observed for Yoruba (YRI), the other non-admixed African population and greater that any of the non-African populations. The higher number of derived sites observed in Luhya (LWK) and African Americans (ASW) could be explained by the mixture of East and West African genomes for the former (Henn et al. 2012) and putative of both diverse West African and European components in the ancestors of the latter. Furthermore the higher proportion of total homozygous sites (both ancestral and derived) in non African populations fits with the documented bottleneck events that characterized their genetic history (Gravel et al. 2011).

Table 3.4 Summary statistics for the low depth samples. For each of the Ethiopian and 1000 Genomes Project samples the following summary statistics were calculated: number of variable sites (N. Var Sites), genomic positions where in 24 samples at least one variant was observed (with the exception of GIH, where only 13 samples were available); average genome-wide expected heterozygosity (Exp. H) calculated on the total set of variable sites; median proportion of ancestral homozygous sites (Anc_Homoz) calculated over all the samples available from each population; median proportion of derived sites (Het. Sites); median proportion of derived homozygous sites (Der_homoz). Each column is independently formatted to highlight high values in red and low values in blue.

	N. Var Sites	Exp H.	Anc_Homoz	Het. Sites	Der_homoz
LWK	12,366,127	0.067	0.884	0.067	0.049
YRI	11,241,830	0.064	0.888	0.062	0.049
ASW	12,815,279	0.067	0.885	0.065	0.05
Gumuz	9,674,165	0.063	0.893	0.057	0.05
Eth. Somali	10,355,332	0.063	0.89	0.058	0.052
Wolayta	10,817,393	0.063	0.89	0.059	0.051
Oromo	10,795,823	0.063	0.889	0.058	0.052
Amhara	10,278,207	0.061	0.892	0.055	0.053
GIH	7,212,792	0.052	0.893	0.049	0.058
CEU	7,768,428	0.048	0.896	0.046	0.058
IBS	8,303,527	0.052	0.892	0.05	0.058
TSI	7,795,947	0.051	0.895	0.047	0.058
CHB	7,215,199	0.045	0.897	0.044	0.059
CHS	7,182,687	0.046	0.896	0.045	0.059
PUR	9,603,748	0.056	0.89	0.054	0.056
MXL	8,715,477	0.051	0.892	0.05	0.058
CLM	9,300,613	0.053	0.892	0.051	0.057
PEL	8,172,480	0.05	0.892	0.048	0.06

The average genomic pairwise F_{ST} values for the same populations, reported in Table 3.5, show a strong separation between the African and non-African populations, and further subdivisions within each of these two groups. The non-African populations belonging to the three European, Asian and American groups indeed show smaller genetic distances with the populations from their own group than with other non-African populations, as expected from the isolation by distance model (Prugnolle et al. 2005). The Afro-Asiatic-speaking Ethiopian populations (Amhara, Eth. Somali, Oromo and Wolayta) have lower F_{ST} values among each other than those observed among other African populations. Compared to distances with West Africans, these Ethiopian populations also showed lower F_{ST} values with the ASW and LWK. Surprisingly,

the F_{ST} values between Afro-Asiatic Ethiopians and Gumuz (Nilotic) are higher than between the former and any other African population. In contrast, Gumuz are closer to YRI, LWK and ASW than to any Ethiopian population, hence suggesting complex relationships between Nilotic and Afro-Asiatic Ethiopians and the other African groups analysed. It is also important to note that, when compared with the F_{ST} values calculated from SNPchip data (see Chapter 2), the values reported for whole genome data appear smaller, due to the increased presence of low frequency variants. The F_{ST} between two populations, one carrying a low frequency variant and the other not showing such mutation at all, is very close to zero. Therefore the presence of such variants tend to lower down the average genomic F_{ST} between any given pair of populations.

Table 3.5 Average genomic F_{ST} values between each pair of analysed populations. The table is formatted to show high values in red and low values in blue.

	LWK																	
YRI	0.006	YRI																
ASW	0.007	0.007	ASW															
Gumuz	0.014	0.017	0.017	Gumuz														
Eth.					Eth.													
Somali	0.015	0.017	0.014	0.021	Somali													
Wolayta	0.014	0.017	0.013	0.020	0.014	Wolayta												
Oromo	0.015	0.017	0.013	0.020	0.013	0.012	Oromo											
Amhara	0.015	0.018	0.014	0.021	0.013	0.012	0.011	Amhara										
GIH	0.031	0.035	0.026	0.045	0.035	0.032	0.029	0.029	GIH									
CEU	0.025	0.028	0.020	0.035	0.025	0.023	0.020	0.019	0.020	CEU								
IBS	0.027	0.030	0.021	0.037	0.025	0.023	0.020	0.020	0.022	0.006	IBS							
TSI	0.025	0.027	0.019	0.034	0.023	0.021	0.019	0.018	0.020	0.004	0.005	TSI						
CHB	0.030	0.033	0.027	0.043	0.035	0.033	0.031	0.030	0.029	0.023	0.026	0.023	CHB					
CHS	0.030	0.033	0.027	0.042	0.035	0.033	0.031	0.030	0.029	0.023	0.026	0.023	0.004	CHS				
PUR	0.020	0.022	0.015	0.028	0.019	0.017	0.015	0.014	0.017	0.007	0.007	0.006	0.019	0.019	PUR			
MXL	0.026	0.029	0.021	0.036	0.027	0.024	0.023	0.022	0.021	0.012	0.013	0.012	0.019	0.018	0.008	MXL		
CLM	0.023	0.026	0.018	0.032	0.022	0.020	0.018	0.017	0.019	0.008	0.009	0.008	0.020	0.019	0.005	0.007	CLM	
PEL	0.035	0.039	0.030	0.049	0.040	0.037	0.035	0.034	0.036	0.026	0.027	0.025	0.027	0.027	0.018	0.011	0.015	

To represent the sharing of genetic diversity between individuals of the same Ethiopian population in relation to other available data (The 1000 Genomes Project Consortium et al. 2012), the derived site frequency spectrum of the five sequenced populations together with three African and two non-African control populations are reported in Figure 3.3. The figure shows, across 24 samples (48 chromosomes) chosen from each population, the proportion of derived sites observed in 1, 2, ..., 48 chromosomes. The higher proportion of singletons and doubletons (derived sites observed only once or twice in the sample, respectively) in the African populations shows an overall higher genetic diversity in these populations, while a higher proportion of "fixed" derived variants at the right-hand end of the spectrum in non-African populations can be interpreted as the effect of population bottlenecks that occurred during the migration out of Africa (Gravel et al. 2011). Furthermore, ASW, LWK and some Ethiopian populations show an even higher presence of low frequency variants, perhaps due to the admixed nature of these populations. Further analyses are needed to explain these differences and to explore the impact of small differences in coverage between populations. However, the Ethiopian Gumuz appear as the most distinct of the African populations studied, with an increase in the intermediate frequency variants which cannot simply be explained by the decrease in frequency of singletons. The cumulative proportion of variants observed in 2-6 chromosomes in Gumuz is indeed higher than in any other population studied, which can perhaps be explained by a novel demographic history of this population that needs further study.

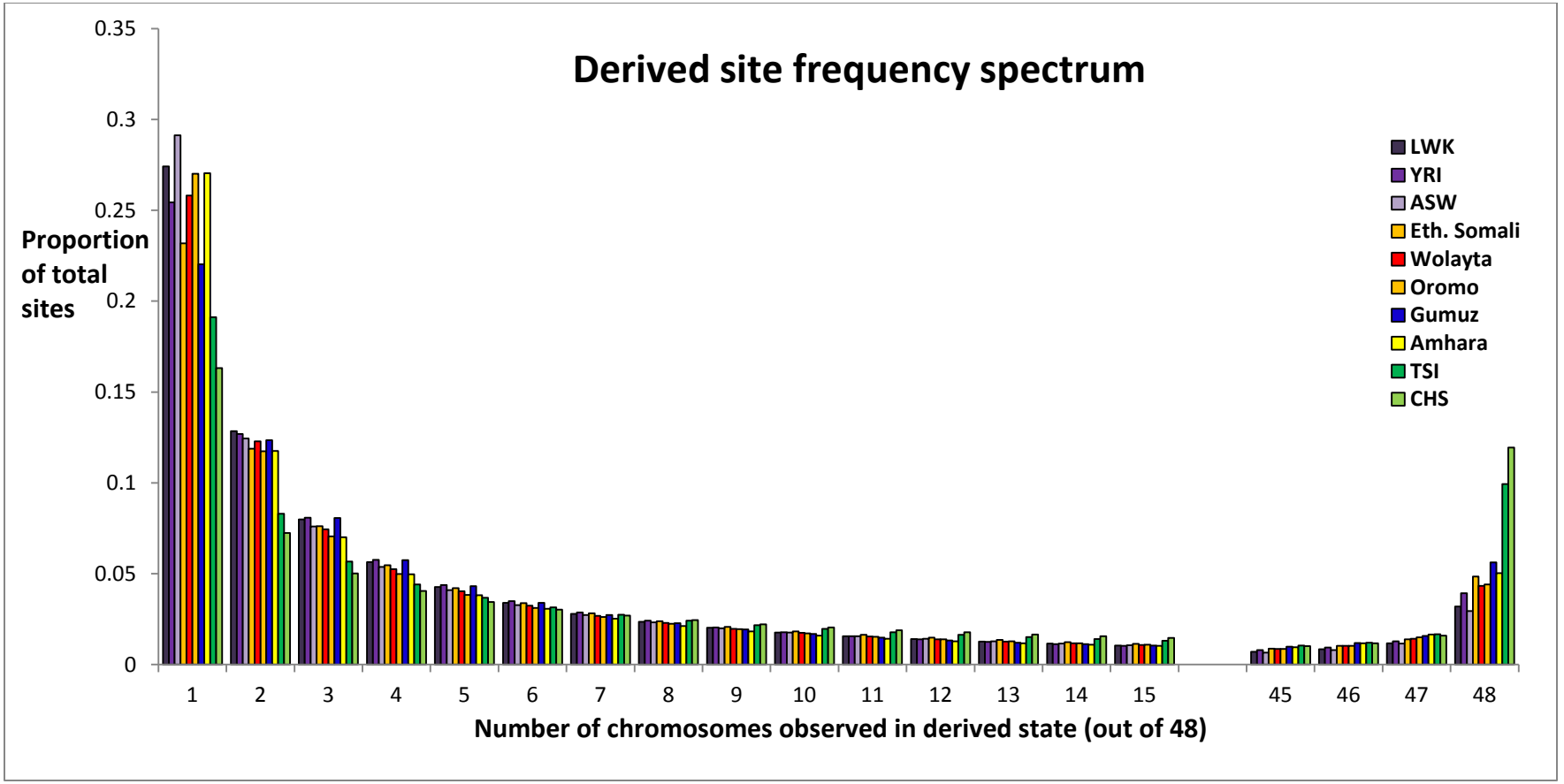

Figure 3.3 Derived site frequency spectra of five Ethiopian and five control populations. The proportion of sites derived in different numbers of chromosomes is plotted for a total of 48 chromosomes from each population. The Ethiopian populations are coloured according to their linguistic group of origin and sorted as shown in the legend (Cushitic in orange, Semitic in yellow, Omotic in red and Nilotic in blue). The spectra of five control populations (three African in shades of purple and two non African in shades of green) are shown for reference. The difference in sequencing depths of Ethiopian (~8x) and 1000 Genomes Project (~6x) populations has not been taken into account to produce this plot.

3.3.4 A comparison with the SNP array results

The results of the analyses performed on the newly-generated sequence and 2.5M SNP array data were compared with the data previously obtained from the Illumina 1M Omni SNP array (Pagani et al. 2012). This comparison aimed both to assess the relationship between the newly collected individuals and the ones previously analysed from the same, self-reported, ethnic groups and to investigate the extent of ascertainment bias when comparing SNP array with sequencing results. A principal component analysis of the set of 143K LD pruned (Alexander et al. 2009) markers overlapping between the 2.5M Omni SNP array data available for the new samples and the 1M Omni SNP array data from the previous samples is reported in Figure 3.4. The newly-generated samples, identified by the suffix "seq", cluster with their counterparts, showing a good concordance between the inclusion criteria of the two sampling strategies. The highest divergence between previously- and newly-sampled groups is observed for the Ethiopian Somali. This is probably due to the presence of stratification (i.e. clans) within this broadly-defined group. Among the newly-sampled individuals, the ones that were sequenced at high depth are identified by the "HC" suffix. These samples fall within the main core of each typed population, therefore providing an indication that the chosen individuals are not outliers in representing their respective groups.

A different comparison between the two sets of data is provided by the heterozygosity estimates (Figure 3.5). The values estimated from genotypes obtained from the SNP array and sequencing results are different both in relative and absolute terms. The heterozygosity estimated from sequence data is almost one order of magnitude smaller than the one from the SNP array results. This is likely due to the higher proportion of low-frequency variants in the sequence data (Figure 3.3) which decreases the overall probability of finding a heterozygous locus in the overall sample. However, the most striking change affects the relative difference between African and non-African values of heterozygosity. Based on the SNP array data the non-African populations (in green in Figure 3.5) show values equal to or higher than the African control populations (in purple). In contrast, from sequencing, the African control populations show heterozygosity values that are higher than any other sampled population, even higher than the Ethiopians which, for the SNP array, showed the highest heterozygosity values among Africans. This discrepancy between heterozygosity values estimated from different sets of genotype calls can be explained by the ascertainment bias that affects the choice of markers included on the SNP array. As discussed in paragraph 1.1.1, the results from the SNP arrays have the tendency to magnify the genetic diversity present outside Africa because more of the

SNPs were discovered outside Africa. As a result, populations with a considerable proportion of both African and non-African genomes such as the Afro-Asiatic Ethiopians will show the highest values (having the highest probability of observing a non-African allele in combination with an African allele) while populations with a low or undetectable proportion of non-African ancestry such as the control African populations will show the smallest values of heterozygosity. The sequencing data provide an unbiased perspective, bringing the African populations back to their positions of “high genetic diversity”, consistent with a commonly accepted African origin of the human genetic diversity.

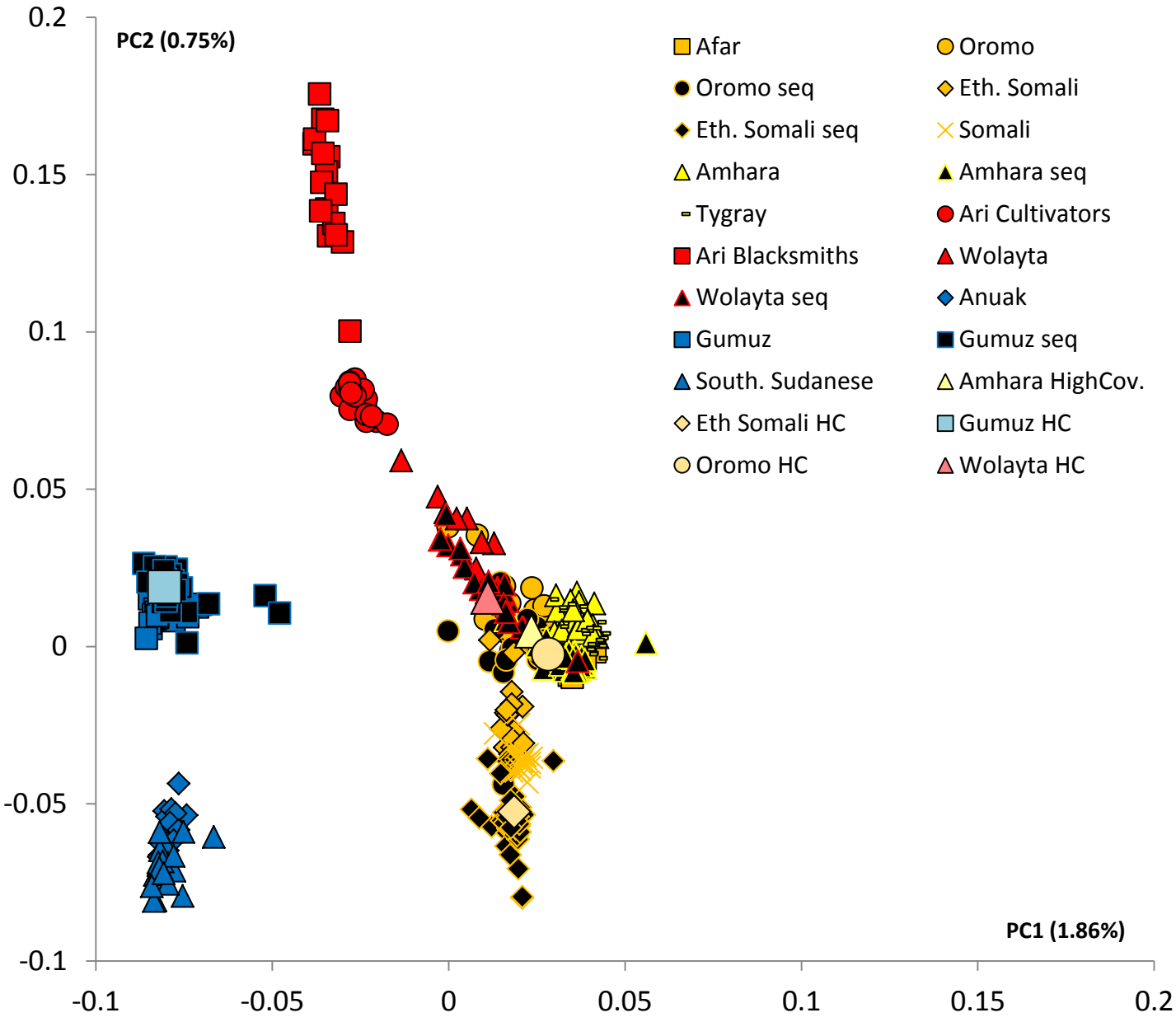

Figure 3.4 Principal component analysis of SNP- array genotypes from the previously- and newly-sampled Ethiopian populations. The newly-sampled groups are designated by the "seq" suffix, while the five individuals sequenced at high depth are labelled "HC". The PCA was carried on the set of 143K LD pruned markers overlapping both the 1M and 2.5M Omni SNP array.

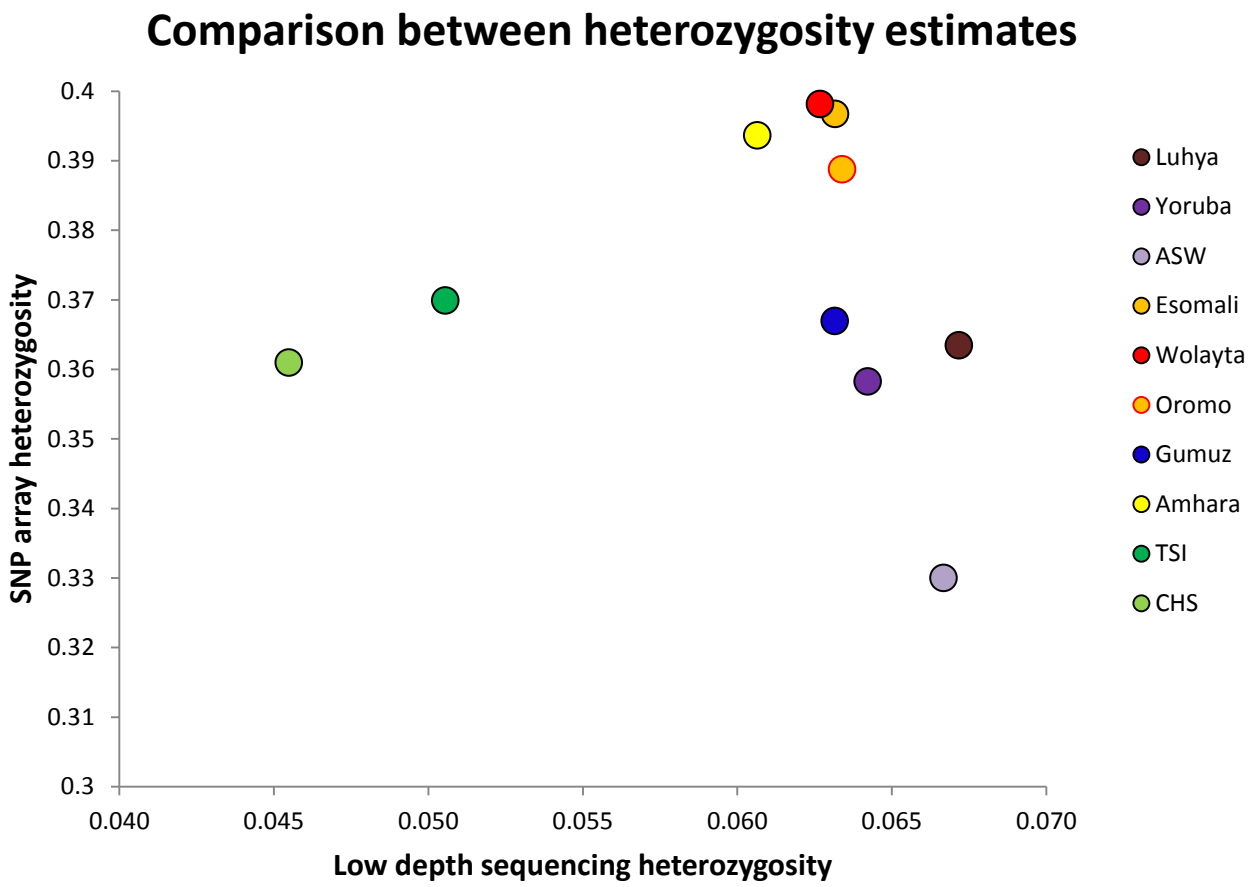

Figure 3.5 Comparison between heterozygosity estimates. The heterozygosity values estimated from the genotypes obtained from the SNP array and low-depth sequencing are plotted on the y and x axes respectively. From the sequencing results, the African populations, shown in shades of purple, have heterozygosity values that are higher than either the non-African (in shades of green) or the Ethiopian populations (coloured according to their linguistic group of origin).

3.3.5 Sharing of low frequency variants within and between populations

The availability of low frequency sites from whole genome sequencing, and particularly of sites only observed twice in the total sample (doubletons), allowed to refining patterns of admixture between populations. If the two copies of a doubleton have originated from the same mutation event and if the analysed populations are somewhat stratified from a genetic point of view, the expectation is to find both copies of any doubleton within the same population. However, if one population experienced genetic introgression from another population, the two populations might share one copy each of a certain number of doubletons. Figure 3.6 shows, for each population, the proportion of doubleton allele sharing with the other analysed populations.

Following the procedures described when this statistic was first introduced (The 1000 Genomes Project Consortium et al. 2012), each column in Figure 3.6 represents the proportion of doubleton alleles that a given population share with any other population. Remarkably Oromo and Amhara share with each other the same proportion of alleles they do with themselves. Wolayta, Somali and particularly Gumuz show, instead, a less panmictic pattern having each a greater fraction of private doubletons. The observed pattern is consistent with Amhara and Oromo being the main acceptors of genetic flow from the neighbouring Wolayta and Somali populations, while more distant Gumuz, Yoruba and Luhya do not contribute to this flow. Furthermore this gene exchange seems to be unbalanced by a greater flow toward the Oromo and Amhara genomes. Remarkably little or no allele sharing was detected between the Ethiopian and non-African populations. This might appear in contrast with the marked patterns of non-African admixture described in Chapter 2. However the lack of sharing of low frequency, and hence recently arisen variants between Ethiopians and non-Africans can be seen as a confirmation that such admixture event is ancient and, perhaps, no further admixture took place during the last few generations.

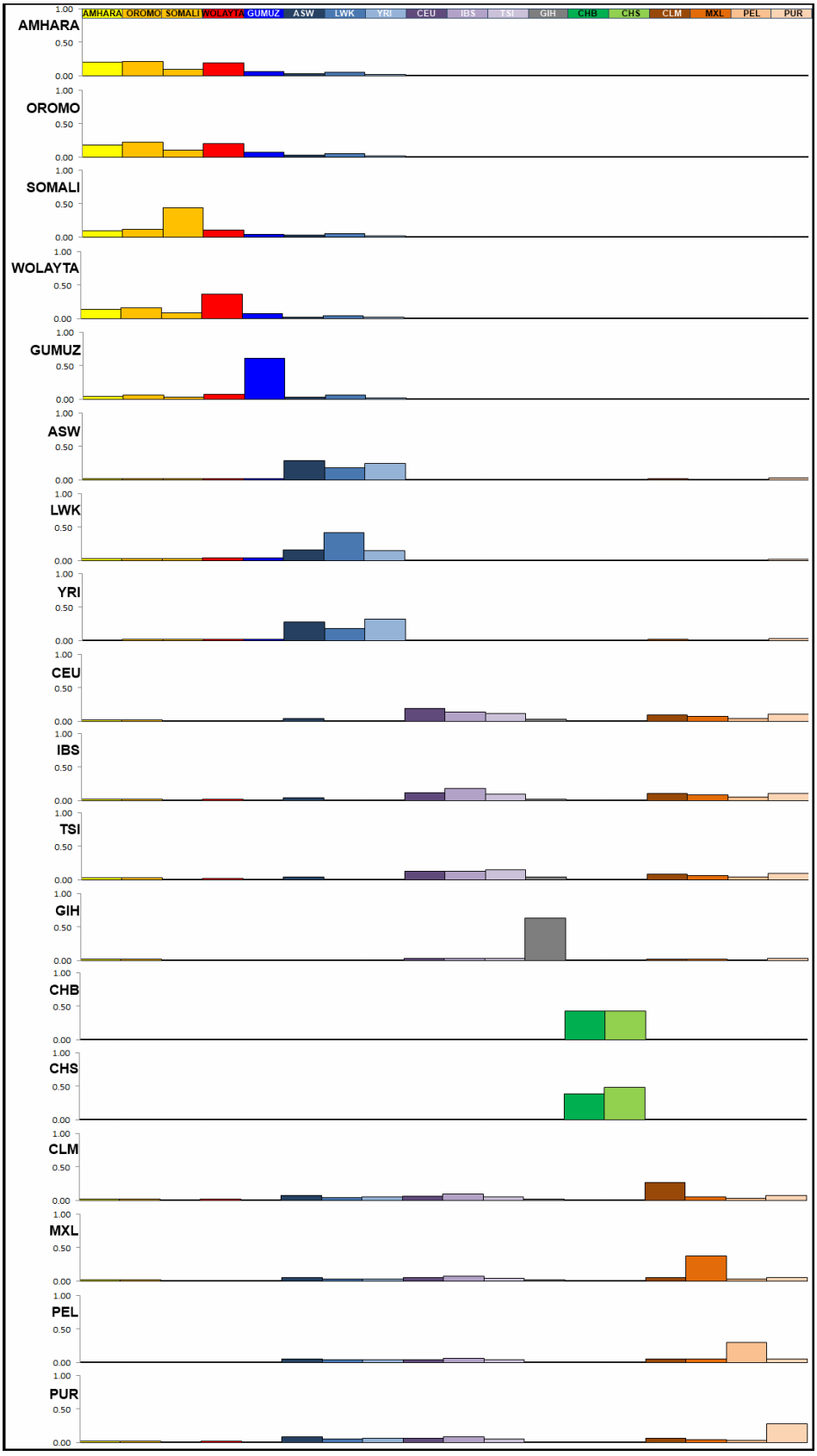

Figure 3.6 Doubleton allele sharing between worldwide populations. Each column represents the proportion of doubleton allele sharing between a given population (row) and every other population (columns).

3.3.6 A comparison between high and low coverage sequencing on the same samples

Another interesting comparison is between the high-depth samples and their low-depth or *in silico* reduced counterparts. As it appears from Figure 3.7 A, the proportion of called sites that matched in the low and high depth samples is as high as 90% in all the five samples (black bars). The non-matching sites (white bars) were further subdivided into different categories (Figure 3.7 B). From this subdivision it appears that a significant proportion of the non-matching sites are explained by sites that were not detected in the low coverage samples. In fact, approximately one half of the differences in each sample are caused by sites that were observed only once in the high depth sample (singletons) and one fifth of the differences due to sites that were observed as variable in the overall sample but not detected in the low depth sample. The remaining one third of the mismatching sites is made up by the ones that were undetected in the high depth data and the ones where the zygosity did not match between the two call sets. Overall, the 90% match between the two sets of calls reflects the high efficiency of the integrated calling system for the low-depth samples. Furthermore, the non-matching sites are predominantly explained by sites discovered by the high-depth calling only, which both shows the added value of sequencing samples at high depth and excludes major artefacts in the low depth calls. The proportion of sites that were called only in the low depth set and, therefore, which might be enriched for false discoveries, amount in each sample to less than 1% (brown and green bars in figure 3.7).

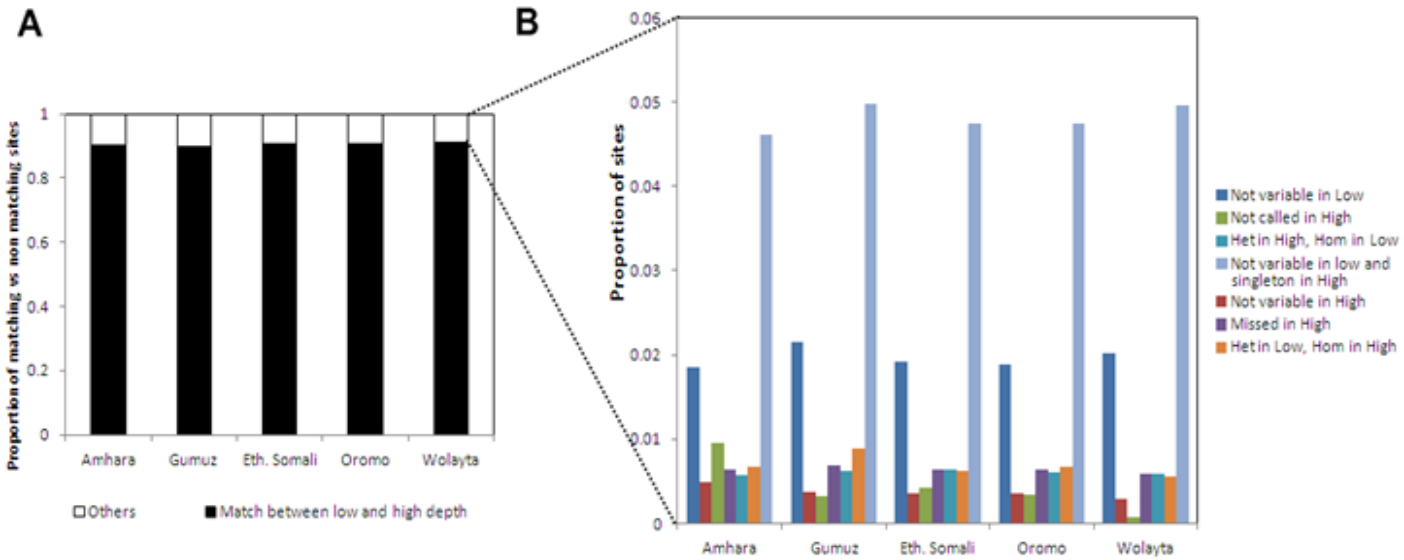

Figure 3.7 Match between high and low depth sequenced samples from five Ethiopian populations. Panel **A** shows the total proportion of sites that match (black bars) or do not match (white bars) between the high- and low-coverage call sets in five individuals. The proportion of non-matching sites is further subdivided (**B**) into different categories (coloured bars). These further categories are: not

variable in low (listing those sites called only in the high depth call set, dark blue); not variable in high (only variable in the low call sets, brown); not called in high (sites that are variable in the low call set but not called in the high call set, light green); missed in high (sites missing from the high call set and not variable in the low call set, purple); het in high hom in low (sites called as heterozygous in the high depth and as homozygous in the low depth call set, turquoise); het in low hom in high (sites called as heterozygous in the low depth and as homozygous in the high depth call set, orange); not variable in low and singleton in high (those sites that were not called in the low depth and called only on one chromosome on the high depth call set, light blue).

3.3.7 Single genome demography from five high-depth Ethiopian genomes

The main advantage of sequencing at high depth is the achievement of a virtually unbiased representation of the diploid genome of an individual. This representation can be exploited to trace the demographic events that characterized the genetic pool of origin of a given sample. Each genome can be seen as the temporary coexistence of genetic fragments pooled from individuals who, in turn, were mosaics themselves. Treating an individual genome as a composite of such events can therefore inform about the genetic history of the whole genetic pool. One method recently developed to estimate the change in effective population size (N_e) over time from high-depth re-sequenced samples, PSMC, was introduced by Li and Durbin in 2011 (Li and Durbin 2011) and described in paragraph 1.6.3.1. The five high-depth Ethiopian samples were run together with one CEU and one YRI control sample from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010) and the resulting curves showing the change of effective population size (N_e) over time are reported in Figure 3.8. The run was performed by Dr. Stephan Shiffels from the WTSI using an in-house modified algorithm with the supervision of Dr. Richard Durbin, one of the co-authors of the original publication. At a glance the Ethiopian genomes fit well within the scenario outlined by the CEU and YRI controls, with intermediate Ne between the two at the time of the out of Africa (~2000 ga, (Gronau et al. 2011)) and with Amhara and Oromo showing the highest Ne in recent times, consistently with their high census size, accounting for up to 60% of the total Ethiopian population. Remarkably, this improved algorithms yielded a previously undetected separation between the CEU and YRI samples between 10^4 and 10^5 generations ago (ga) with all the Ethiopian samples falling between the two. Overall the intermediate position of the Ethiopian samples is not surprising, given their geographic location and detectable proportion of non-African ancestry. The PSMC analysis assumes a single panmictic population with no inbreeding; the non-African

introgression into the Ethiopian genomes described in Chapter 2 might introduce a confounding effect to this assumption, causing the Ethiopians to fall between the Yoruba and CEU. However, the difference of estimated N_e values between the CEU and YRI samples at a time depth that predates the split of African and non-African populations (~2000 ga, (Gronau et al. 2011)) needs further investigation, as well as the effect caused by the above mentioned introgression.

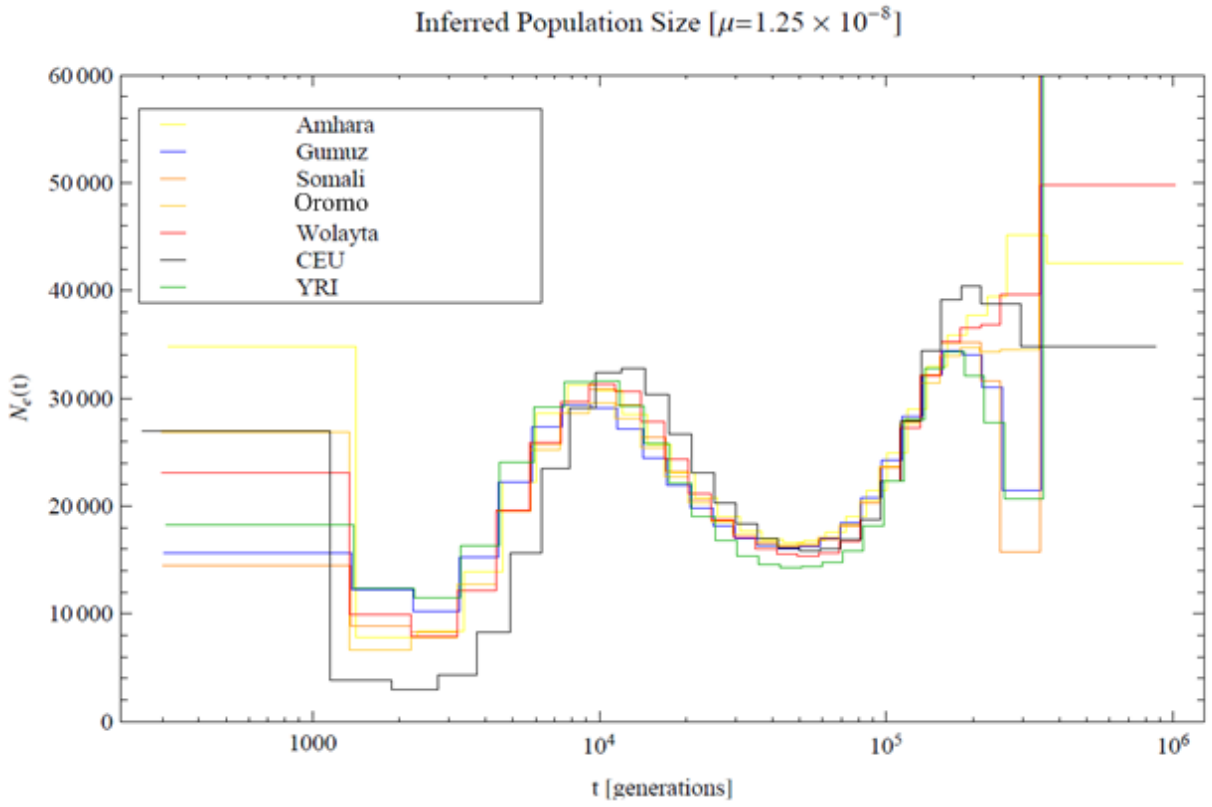

Figure 3.8 PSMC analysis. The high depth Ethiopian Amhara (yellow), Gumuz (blue), Eth. Somali (dark orange), Oromo (light orange) and Wolayta (red) samples were run together with one CEU (purple) and one YRI (green) to compare the changes in their effective population sizes (N_e) over time.

To confirm whether the distance between the Ethiopian samples and the controls, on one hand, and the distances between CEU and YRI, on the other, could be explained by differences in the read lengths of the sequenced samples, a further analysis was carried out. The same CEU (NA12878) and YRI (NA19240) samples were run together with an additional YRI (NA18606) and one Ethiopian (Gumuz) samples. Furthermore, in order to provide statistical support to the obtained PSMC curve, the same Gumuz sample was run 20 times through bootstrapping on the variable sites, by dividing the genomes into 5 cM units and shuffling them each time. The Gumuz sample was chosen among the available Ethiopians to minimize biases due to recent

admixture with non-African individuals: Gumuz showed no admixture in the ADMIXTURE and ROLLOFF analyses described in Chapter 2. This PSMC run, represented in Figure 3.8, showed that the ancient separation between the CEU (purple) and YRI (green) samples persists also when the samples are sequenced with different read lengths (NA12878 and NA18606 with 100bps while NA19240 with 35 bps). Furthermore, the separation between the Gumuz samples and both of the control populations was supported for most of the curve length by the bootstrap underlying a complex relationship between the ancestors of these three populations. The analysis shown in Figure 3.9 therefore supports the robustness of the findings reported in Figure 3.8.

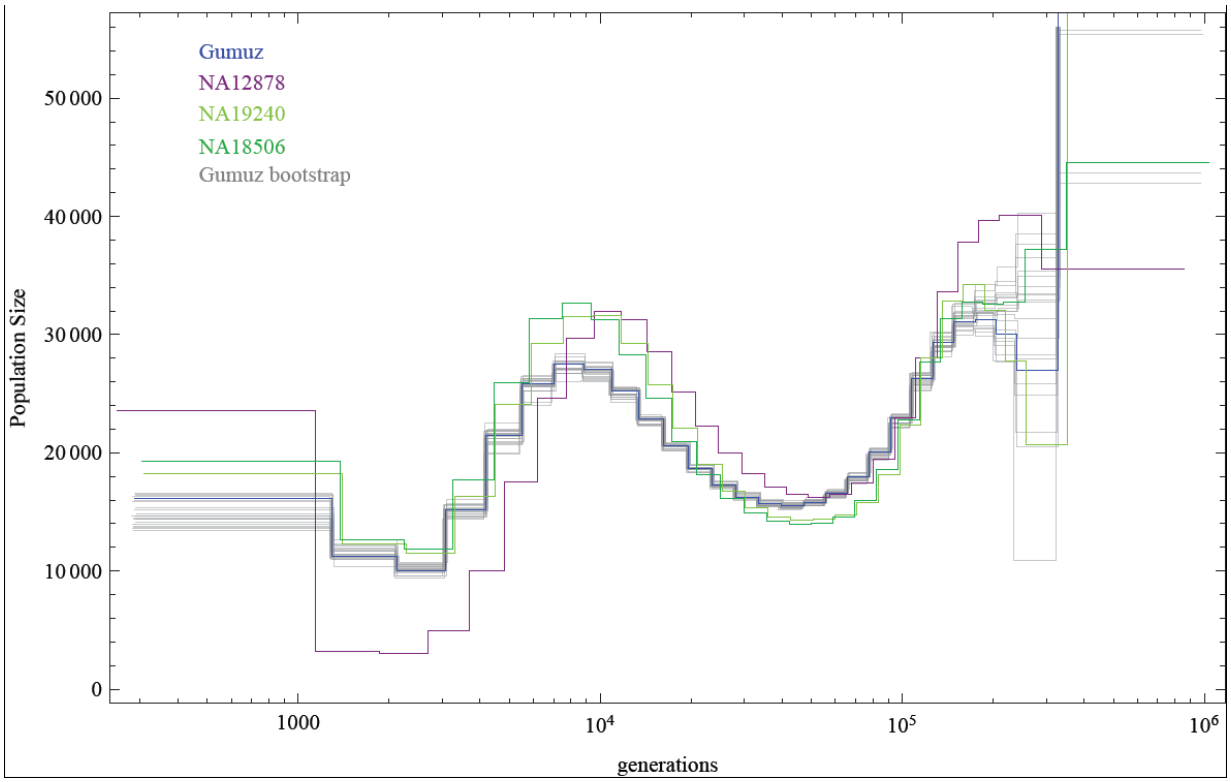

Figure 3.9 PSMC analysis and bootstrap. To confirm the differences between Ethiopian and control populations displayed in Figure 3.8, the high depth Gumuz sample (blue) was run together with one CEU (NA12878, purple, 100bp read length) and two YRI (NA19240, light green, 35 bp read length and NA18506, dark green, 100bp read length) to compare the change in their effective population sizes over time. The distance between the Gumuz and the other curves was tested through 20 bootstraps on the number of variable sites in Gumuz. Both the bootstrap and the use of samples with different read lengths confirmed the differences between the three populations. The mutation rate used was the same as for the previous analysis: $\mu=1.25\ 10^{-8}$.

Although potential bias could have been introduced by unforeseen differences in sequencing procedures of the analysed samples, and keeping in mind that the inferences made are just based on one or two individuals from each population, a first description and interpretation of the PSMC results of Figure 3.9 can be provided. It is also important to keep in mind that the mutation rate used for this analysis ($\mu=1.25\ 10^{-8}$) is one-half of that used in the original publication from Li and Durbin shown in Figure 1.6; therefore to compare the two plots the time scale has to be adjusted accordingly. The mutation rate adopted in the current analysis is based on the average of the directly calibrated rates, as reported in the publication of the Gorilla genome paper (Scally et al. 2012). The three CEU, YRI and Gumuz populations start their curves as a single population before 10^5 ga. After this point the West African YRI split from the main CEU and Gumuz cluster. The three populations follow a general trend of decrease in N_e up to $5x10^4$ ga (when the Gumuz seems to split from the CEU and form a third, separate curve) and N_e increases until around 10^4 ga. This trend, already observed by Li and Durbin, was tentatively interpreted as an ancient population stratification within Africa (Li and Durbin 2011). The above described split, coinciding with the aforementioned population stratification, and supported by the bootstrap performed on the Gumuz sample, could provide further evidence towards the proposed explanation. The ancestors of the three analysed populations could have indeed formed two or more sub-groups up to $8x10^3$ generations ago (ga) with the Gumuz remarkably sharing the early part of their evolutionary history with the ancestors of CEU and the more recent part with the YRI. After this point, which roughly fits with the emergence of the first anatomically modern humans in Africa, the ancestors of the modern CEU started experiencing a decrease in their N_e which eventually led to the out-of-Africa migration around $2\ 10^3$ ga as seen by the drop in N_e of CEU, while the Gumuz curve joins the ones of the two YRI samples. The presence of ancient population stratification in African hominins could be linked to the speciation processes that led to the emergence of the ancestors of Neanderthals around 350kya (Green et al. 2010), to some extent shedding light on the complex genetic relationships observed between modern humans and two recently sequenced hominins, Neanderthals and Denisovans (Green et al. 2010; Reich et al. 2010; Eriksson and Manica 2012). Despite the higher noise in the N_e estimate more recent than 1000 ga, the dramatic increase in population size detected for Amhara and Oromo in Figure 3.8, as well as the already documented increase of the CEU N_e, is consistent with their current high census size.

In summary, the five analysed Ethiopian populations show a recent history consistent with that of African populations that experienced recent growth and potentially some gene flow from non-

African groups. Their distant history sheds light on the ancient population structure first described from genomic data by Li and Durbin (Li and Durbin 2011) and invoked by Eriksonn and Manica to explain the observed pattern of allele sharing with extinct hominids (Eriksson and Manica 2012), which might help in clarifying the early stages of our origin as a species.

4. General discussion

The rationale for the work presented in this thesis was to explore the genetic diversity of the modern Ethiopian populations, to elucidate their genetic history and to shed light on the evolutionary history of our species. From the latest to the earliest events, the main scientific questions concerned: describing the extent and dating the recent non-African gene flow back to Ethiopia, characterizing putative signature of high altitude adaptation, searching for the legacy of the out-of-Africa migrations in the genetic make-up of modern Ethiopians, and tracing the source of the human genetic diversity.

The first phase of the investigation involved an explorative survey of 13 populations chosen to represent Ethiopian and nearby diversity. This survey was carried out by genotyping 24 samples from each population on an Illumina Omni 1M SNP array (Chapter 2). The Ethiopians showed the highest heterozygosity levels in Africa and worldwide. The main outcome of the SNP array analyses was a strong correlation between linguistic and genetic stratifications. Indeed, the populations speaking Nilotic, Omotic, and Semitic plus Cushitic languages formed three distinct clusters. Furthermore, the Semitic and Cushitic and, in part, the Omotic populations showed a considerable proportion of non-African ancestry, accounting for up to 50% of their genomes. This component was estimated to have entered the Ethiopian gene pool around 3000 years ago. Although with caveats concerning the difficulty to distinguish a point migration event from a continuous gene flow, the retrieved genetic estimate was in good concordance with historical (Hansberry 1974) and linguistic (Kitchen et al. 2009) records related to putative migrations of people from the Levant into the Ethiopian region. The SNP array data also showed evidence for genetic adaptation to high altitude in Tygray and Amhara individuals, detected in a genomic region including the *DEC2* gene. This gene was previously known to be associated with the oxygen sensing machinery, but never found under selection in populations putatively adapted to high altitude. However further functional studies will be needed to confirm the actual link between this gene and the high altitude response. Another signal of putative selection in the admixed Ethiopian populations involved an excess of non-African genomic fragments containing the *SLC24A5* gene, responsible for a substantial proportion of the lighter European skin pigmentation (Pickrell et al. 2009). The detected excess might imply a preferential spread of the European haplotype in these populations.

Due to the reduced availability of North African samples and the low resolution provided by the SNP array data, a comparison between the African components of Ethiopian and North African

populations with the genomes of non-African populations led to inconclusive results. The aim of this comparison was to search for the putative African source of the out-of-Africa migrations, but the high amount of non-African gene flow in East and North African populations resulted in a loss of power of the proposed approach. Although it pointed to Egypt as the better candidate for the source, this unexpected result was considered to require more and better-quality support before being interpreted as challenging the prevailing view of East Africa as the most likely source. A higher number of less-admixed Egyptian samples, examined at a whole genome level, will therefore be needed to address this question.

Another problematic point from the SNP array results was the identification of the African geographical origin of human diversity. The newly-generated Ethiopian data were analysed for their LD patterns following the approach proposed by Henn and colleagues (Henn et al. 2011). The resulting LD decay patterns confirmed the South West African origin of the human diversity as proposed by Henn et al. However, concerns about the suitability for the proposed method of the ascertained SNP-array markers, as well as the putative confounding effects introduced by a small amount of non-African admixture on the overall r^2-based LD pattern, suggest caution when interpreting such results. Furthermore, the concept itself of geographic origin of genetic diversity appears weak, considering the migratory events that African populations might have experience over the last 200.000 years.

The SNP array data generated for the 13 Ethiopian populations was also used to search for a set of 22 autosomal markers that could best summarize the observed Ethiopian diversity. At least one such set, described in Appendix 4, performed better than any of a pool of 1000 randomly-generated sets. Given the successful outcome of this exercise, the set was shared with other labs to place a broader number of Ethiopian samples within the genetic space described by the populations studied in the present work.

Overall, the 13 chosen populations seemed to adequately represent the Ethiopian diversity, leaving almost no gaps in the genetic space that spans from North Africa to the East and West sub-Saharan populations, in accordance to the high cultural and linguistic diversity observed in the area.

In order to provide a higher resolution, unbiased representation of the Ethiopian genetic landscape, twenty-five individuals each from five of the most representative Ethiopian populations were collected and their genomes were re-sequenced on an Illumina HiSeq platform. The genotypes of the re-sampled populations were also compared with the ones

generated in the first phase of the project. All the re-sampled populations clustered with their homologous groups, hence showing a good reproducibility of the inclusion criteria adopted for the re-sampling campaign.

One striking difference between the previously-generated SNP array data and the re-sequencing data was illustrated by the heterozygosity values shown in Figure 3.5. Despite the one order of magnitude difference between the sets, readily explained by the higher number of low frequency variants in the sequencing data, the most remarkable difference is the relative position of the Ethiopian heterozygosity within the broader African and worldwide context. In contrast to values estimated from genotype data, the sequence data based heterozygosity estimates placed the Ethiopian populations at an intermediate position between the low non-African and the high sub-Saharan African values. The contrast between the SNP array and the sequencing result, consistent with an African origin of human diversity and with the partial non-African origin of the modern Ethiopian genome, is most likely a consequence of the SNP array ascertainment bias described in section 1.1.1. The heterozygosity values obtained from the SNP array data tend to magnify the non-African diversity as a consequence of the overrepresentation of non-African-specific variants in the array marker list. The intermediate positioning of the Ethiopians between non-African and sub-Saharan populations was also confirmed by the summary statistics of the number of variable sites and proportion of homozygous and heterozygous positions per population reported in Table 3.4. The derived site frequency spectrum (Figure 3.3) and pairwise F_{ST} values (Table 3.5) further confirmed this, with the notable exception of the Gumuz, who showed SFS and pairwise F_{ST} values consistent with a distinct evolutionary history. While the absence of non-African genomic components in the Gumuz could partially explain the observed differences from the other Ethiopian populations, the Gumuz are also distinct from the other sub-Saharan African populations. The PSMC analysis carried out on a high depth Gumuz genome (Figure 3.9) further described the ancestors of this population as potentially from a separate branch of the previously-described ancient African population stratification (Li and Durbin 2011; Eriksson and Manica 2012). The putative unique evolutionary history of the Gumuz, to be fully understood, needs further analyses aiming at comparing their haplotype structure with the one of the surrounding African populations. The sequencing data also provided another angle to the interpretation of the genetic back flow into the Ethiopian genomes detected from the SNPchip data. The non-African admixture was not detected by the doubleton sharing analyses, consistently with the high number of generations estimated since the time of admixture. A high level on intra-Ethiopian

admixture was instead underlined by the patterns of doubleton sharing (Figure 3.6), with Amhara and Oromo acting as a recipient of allele sharing from each other and from Somali and Wolayta, while Gumuz seem to be more isolated from the other Ethiopian populations.

More broadly speaking, future analyses will be needed to extract further biological information from the whole set of newly-sequenced samples presented here. The results described in Chapter 3 of this thesis must indeed be taken as preliminary. The choice of not investigating further the generated results was imposed by the time constraints of a three-year PhD project. However, the sequencing data that was produced for this project will be the focus of future work aimed at clarifying the points that could not be addressed with the SNP array results alone. In particular, genetic traces of the populations that performed the out-of-Africa migrations will be looked for in the African components of the Ethiopian genomes. In order to do so, the new genome sequences will have to be phased (Delaneau et al. 2012) and the resulting haplotypes partitioned into their African and non-African components using approaches like Hapmix (Price et al. 2009) and fineSTRUCTURE (Lawson et al. 2012). The non-African genomic components of the admixed Ethiopian populations can then be screened for signatures of recent gene flow back to African and for more detailed characterization of ancient diversity in the strictly African-specific ancestry components. The genetic signature of the out-of-Africa migration could indeed be represented, in Ethiopian populations, by genomic fragments that despite being of African origin, are more closely related to non-Africans than to most sub-Saharan populations. Such fragments, however, can potentially be separated from the ones introduced by recent back flow by means of looking at their heterozygosity values and TMRCA distributions. The out-of-Africa signature should indeed be characterized by higher heterozygosity and deeper TMRCA values than any non-African genome. A similar analysis should also be carried out on the African component of the Ethiopian genomes, to specify which genomic segments are peculiar to the region through clusters of F_{ST} outliers and proportion of identity by descent sharing. The same genetic diversity and TMRCA parameters can be calculated on the African fragments specific to the region, to find patterns of increased diversity and deeper coalescence, which could link the modern Ethiopian populations with the first modern humans that emerged in the area.

In addition to these demographic analyses, selection scans can be carried out on the re-sequenced Ethiopian populations and the results compared with the available phenotypes in search of possible functional explanations. The phenotypes collected, currently stored in secure UCL archives, can perhaps then be released in an openly-accessible database, like the

sequence data. These proposed analyses will be carried out in the months following the submission of this thesis and the results will be written up in one or more scientific articles.

In conclusion, the analyses performed in this thesis show that the high cultural and ethno-linguistic diversity in Ethiopia is mirrored in the genetic diversity of its populations. This diversity is structured in three main layers. The first, most recent, layer involves gene flow from Middle Eastern populations that took place around 3000 years ago. While being the easiest to detect, this genetic component is as well the biggest confounder for other analyses of earlier events. The second layer, relevant to all the non-African humans, is the putative signature of the out-of-Africa migrations. This level of investigation demonstrated, from a mitochondrial perspective, a crucial role of the Ethiopian populations in the out of Africa migrations, but further investigations of high resolution autosomal data are needed. The third and deepest layer is represented by the genetic traces that might link the Ethiopian populations with the emergence of the first modern humans documented by the rich fossil records of the region. These three levels of genetic information enclosed in the genome of the Ethiopian people witness the potential offered by the study of the genetic diversity in the region. The potential is there to shed light on the early stages of our evolutionary history from a genetic perspective, as well as on the processes that brought our species to spread over all the continents.

5. Bibliography

Alexander DH, Novembre J and Lange K (2009). "Fast model-based estimation of ancestry in unrelated individuals." Genome Research **19**(9): 1655-1664.

Alkorta-Aranburu G, Beall CM, Witonsky DB, Gebremedhin A, Pritchard JK and Di Rienzo A (2012). "The genetic architecture of adaptations to high altitude in ethiopia." PLoS Genet **8**(12): e1003110.

Beall CM (2002). "An Ethiopian pattern of human adaptation to high-altitude hypoxia." Proceedings of the National Academy of Sciences **99**(26): 17215-17218.

Beall CM (2007). "Detecting natural selection in high-altitude human populations." Respir Physiol Neurobiol **158**(2-3): 161-171.

Beall CM, et al. (2010). "Natural selection on EPAS1 (HIF2) associated with low hemoglobin concentration in Tibetan highlanders." Proceedings of the National Academy of Sciences **107**(25): 11459-11464.

Behar DM, et al. (2010). "The genome-wide structure of the Jewish people." Nature **466**(7303): 238-242.

Bentley DR (2006). "Whole-genome re-sequencing." Curr. Opin. Genet. Dev. **16**(6): 545-552.

Blench R (2006). Archaeology, Language, and the African Past, Rowman Altamira.

Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, Reynolds A, Ostrer H, Mezey JG and Bustamante CD (2012). "PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations." Hum. Biol. **84**(4): 343-364.

Campbell MC and Tishkoff SA (2008). "African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping." Annu Rev Genomics Hum Genet **9**: 403-433.

Cavalli-Sforza LL, Menozzi P and Piazza A (1994). The history and geography of human genes. Princeton, N.J., Princeton University Press.

Colonna V, Pagani L, Xue Y and Tyler-Smith C (2011). "A world in a grain of sand: human history from genetic data." Genome Biol **12**(11): 234.

Delaneau O, Marchini J and Zagury JF (2012). "A linear complexity phasing method for thousands of genomes." Nat Methods **9**(2): 179-181.

Ehret C (1995). Reconstructing Proto-Afroasiatic (Proto-Afrasian) : vowels, tone, consonants, and vocabulary. Berkley, Berkeley : University of California Press, 1995.

Eriksson A and Manica A (2012). "Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins." Proc. Natl. Acad. Sci. U. S. A. **109**(35): 13956-13960.

Frazer KA, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-861.

Freeman D and Pankhurst A (2003). Peripheral People. The Excluded Minorities of Ethiopia. London, Hurst and Company.

Gibbons A (2011). "Anthropology. A new view of the birth of Homo sapiens." Science **331**(6016): 392-394.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA and Bustamante CD (2011). "Demographic history and rare allele sharing among human populations." Proc. Natl. Acad. Sci. U. S. A. **108**(29): 11983-11988.

Green RE, et al. (2010). "A Draft Sequence of the Neandertal Genome." Science **328**(5979): 710-722.

Gronau I, Hubisz MJ, Gulko B, Danko CG and Siepel A (2011). "Bayesian inference of ancient human demography from individual genome sequences." Nat. Genet. **43**(10): 1031-1034.

Handley LJ, Manica A, Goudet J and Balloux F (2007). "Going the distance: human population genetics in a clinal world." Trends Genet. **23**(9): 432-439.

Hansberry WL (1974). Pillars in Ethiopian History. Washington D.C., Howard University Press.
Henn BM, Botigue LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J, Bustamante CD and Comas D (2012). "Genomic ancestry of North Africans supports back-to-Africa migrations." PLoS Genet **8**(1): e1002397.
Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodriguez-Botigue L, Ramachandran S, Hon L, Brisbin A, Lin AA, Underhill PA, Comas D, Kidd KK, Norman PJ, Parham P, Bustamante CD, Mountain JL and Feldman MW (2011). "Hunter-gatherer genomic diversity suggests a southern African origin for modern humans." Proc. Natl. Acad. Sci. U. S. A. **108**(13): 5154-5162.
Hinch AG, et al. (2011). "The landscape of recombination in African Americans." Nature **476**(7359): 170-175.
Huang da W, Sherman BT and Lempicki RA (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature protocols **4**(1): 44-57.
Johnson NA, Coram MA, Shriver MD, Romieu I, Barsh GS, London SJ and Tang H (2011). "Ancestral components of admixed genomes in a Mexican cohort." PLoS Genet **7**(12): e1002410.
Jurinke C, van den Boom D, Cantor CR and Koster H (2002). "The use of MassARRAY technology for high throughput genotyping." Adv. Biochem. Eng. Biotechnol. **77**: 57-74.
Kaplan I (1971). Area handbook for Ethiopia.
Keller A, et al. (2012). "New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing." Nat Commun **3**: 698.
Kitchen A, Ehret C, Assefa S and Mulligan CJ (2009). "Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East." Proceedings **276**(1668): 2703-2710.
Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E and Villems R (2004). "Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears." Am J Hum Genet **75**(5): 752-770.
Lahr M and Foley R (1994). "Multiple Dispersals and Modern Human Origin." Evolutionary Anthropology **3** (2): 48-60.
Lawson DJ, Hellenthal G, Myers S and Falush D (2012). "Inference of population structure using dense haplotype data." PLoS Genet **8**(1): e1002453.
Levine DN (1974). Greater Ethiopia. Chicago, The University of Chicago.
Li H and Durbin R (2011). "Inference of human population history from individual whole-genome sequences." Nature **475**(7357): 493-496.
Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL and Myers RM (2008). "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation." Science **319**(5866): 1100-1104.
Liu H, Prugnolle F, Manica A and Balloux F (2006). "A geographically explicit genetic model of worldwide human-settlement history." Am J Hum Genet **79**(2): 230-237.
Lovell A, Moreau C, Yotova V, Xiao F, Bourgeois S, Gehl D, Bertranpetit J, Schurr E and Labuda D (2005). "Ethiopia: between Sub-Saharan Africa and western Eurasia." Annals Of Human Genetics **69**(Pt 3): 275-287.
Lucotte G and Smets P (1999). "Origins of Falasha Jews studied by haplotypes of the Y chromosome." Hum. Biol. **71**(6): 989-993.
McDougall I, Brown FH and Fleagle JG (2005). "Stratigraphic placement and age of modern humans from Kibish, Ethiopia." Nature **433**(7027): 733-736.
Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D, Bekele E, Bradman N, Balding DJ and Tyler-Smith C (2012). "Ethiopian genetic diversity

reveals linguistic stratification and complex influences on the Ethiopian gene pool." Am J Hum Genet **91**(1): 83-96.

Pankhurst R (1998). The Ethiopians, Blackwell Publishers Ltd.

Passarino G, Semino O, Quintana-Murci L, Excoffier L, Hammer M and Santachiara-Benerecetti AS (1998). "Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms." Am J Hum Genet **62**(2): 420-434.

Pauling L (1964). "Molecular Disease and Evolution." Bull. N. Y. Acad. Med. **40**: 334-342.

Phillipson DW (1998). Ancient Ethiopia. Aksum: its antecedents and successors. London, British Museum Press.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW and Pritchard JK (2009). "Signals of recent positive selection in a worldwide sample of human populations." Genome Research **19**(5): 826-837.

Poloni ES, Naciri Y, Bucho R, Niba R, Kervaire B, Excoffier L, Langaney A and Sanchez-Mazas A (2009). "Genetic Evidence for Complexity in Ethnic Differentiation and History in East Africa." Annals of Human Genetics **73**(6): 582-600.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat. Genet. **38**(8): 904-909.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D and Myers S (2009). "Sensitive detection of chromosomal segments of distinct ancestry in admixed populations." PLoS Genet **5**(6): e1000519.

Pritchard JK, Stephens M and Donnelly P (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-959.

Prugnolle F, Manica A and Balloux F (2005). "Geography predicts neutral genetic diversity of human populations." Curr. Biol. **15**(5): R159-160.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and Sham PC (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am. J. Hum. Genet. **81**(3): 559-575.

Rasmussen M, et al. (2010). "Ancient human genome sequence of an extinct Palaeo-Eskimo." Nature **463**(7282): 757-762.

Reich D, et al. (2010). "Genetic history of an archaic hominin group from Denisova Cave in Siberia." Nature **468**(7327): 1053-1060.

Sanger F, Nicklen S and Coulson AR (1977). "DNA sequencing with chain-terminating inhibitors." Proc. Natl. Acad. Sci. U. S. A. **74**(12): 5463-5467.

Scally A, et al. (2012). "Insights into hominid evolution from the gorilla genome sequence." Nature **483**(7388): 169-175.

Scheinfeldt LB, Soi S, Thompson S, Ranciaro A, Woldemeskel D, Beggs W, Lambert C, Jarvis JP, Abate D, Belay G and Tishkoff SA (2012). "Genetic adaptation to high altitude in the Ethiopian highlands." Genome Biol **13**(1): R1.

Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL and Underhill PA (2002). "Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny." Am J Hum Genet **70**(1): 265-268.

Soares P, Alshamali F, Pereira JB, Fernandes V, Silva NM, Afonso C, Costa MD, Musilova E, Macaulay V, Richards MB, Cerny V and Pereira L (2011). "The expansion of mtDNA haplogroup L3 within and out of Africa." Mol. Biol. Evol.: in press.

Stephens M, Smith NJ and Donnelly P (2001). "A new statistical method for haplotype reconstruction from population data." Am. J. Hum. Genet. **68**(4): 978-989.

Stringer CB and Andrews P (1988). "Genetic and fossil evidence for the origin of modern humans." Science **239**(4845): 1263-1268.

Tang H, Peng J, Wang P and Risch NJ (2005). "Estimation of individual admixture: analytical and study design considerations." Genet. Epidemiol. **28**(4): 289-301.

The 1000 Genomes Project Consortium (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

The 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT and McVean GA (2012). "An integrated map of genetic variation from 1,092 human genomes." Nature **491**(7422): 56-65.

Thorne AG and Wolpoff MH (1992). "The multiregional evolution of humans." Sci. Am. **266**(4): 76-79, 82-73.

Tishkoff SA, et al. (2009). "The genetic structure and history of Africans and African Americans." Science **324**(5930): 1035-1044.

Toups MA, Kitchen A, Light JE and Reed DL (2011). "Origin of clothing lice indicates early clothing use by anatomically modern humans in Africa." Molecular Biology and Evolution **28**(1): 29-32.

Tzeng J, Lu HH and Li WH (2008). "Multidimensional scaling for large genomic data sets." BMC Bioinformatics **9**: 179.

Voight BF, Kudaravalli S, Wen X and Pritchard JK (2006). "A Map of Recent Positive Selection in the Human Genome." PLoS Biology **4**(3): e72.

Weir BS and Cockerham CC (1984). "Estimating F-Statistics for the Analysis of Population-Structure." Evolution **38**(6): 1358-1370.

White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G and Howell FC (2003). "Pleistocene *Homo sapiens* from Middle Awash, Ethiopia." Nature **423**(6941): 742-747.

Wright S (1946). "Isolation by Distance under Diverse Systems of Mating." Genetics **31**(1): 39-59.

Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, Kim Y, Sabeti P, Chen Y, Stalker J and Huckle E (2006). "Spread of an Inactive Form of Caspase-12 in Humans Is Due to Recent Positive Selection." The American Journal of Human Genetics **78**(4): 659-670.

Yi X, et al. (2010). "Sequencing of 50 human exomes reveals adaptation to high altitude." Science **329**(5987): 75-78.

Appendix 1 “A world in a grain of sand: human history from genetic data”

Colonna *et al. Genome Biology* 2011, **12**:234
http://genomebiology.com/2011/12/11/234

Genome Biology

REVIEW

A world in a grain of sand: human history from genetic data

Vincenza Colonna[1,2], Luca Pagani[1,3], Yali Xue[1] and Chris Tyler-Smith[1]*

Abstract

Genome-wide genotypes and sequences are enriching our understanding of the past 50,000 years of human history and providing insights into earlier periods largely inaccessible to mitochondrial DNA and Y-chromosomal studies.

To see a world in a grain of sand …

William Blake,
Auguries of Innocence

The genome of each individual is a temporary assemblage of DNA segments brought together for a single generation by a combination of chance, ancestry, recombination and natural selection. These segments have different histories because of recombination and can thus provide independent information about ancestry, the focus of this review. However, the ancestry of different segments is not entirely independent. Humans are not a single randomly mating population: we are subdivided, and these subdivisions into bands, tribes, clans, ethnic groups, nations and so on are of great interest to both scientists and non-scientists. Thus the thousands of different genomic segments in any individual do not trace back to ancestors randomly spread around the globe; segment ancestry is constrained by population history. Two non-recombining segments of the genome, mitochondrial DNA (mtDNA) and the Y chromosome, have been used for decades to study genetic histories [1,2]. Sometimes mtDNA and the Y chromosome share the same history, but often they do not, and such differences alert us to some of the complexities of the human past [3]. But mtDNA and the Y chromosome provide only two perspectives. Recent advances in technology provide access to most of the genome, and increasingly to the genomes of companion species. Here, we consider how this wider perspective is beginning to inform our view of human history. We will see that it is possible to probe much further back into the past, into a period in which the uniparental markers are uninformative yet key evolutionary events took place, and even to speculate about when humans might have begun to wear clothes or to start reconstructing the genomics of former populations before their contact with modern expansions.

Genome-wide data can be obtained by either genotyping samples or re-sequencing them. Genotyping provides information about the allelic state of positions in the genome (currently up to five million, mostly single nucleotide polymorphisms, SNPs) that have prior evidence of variability [4]; such studies are relatively low-cost and routine, and have been performed on massive numbers of samples. Whole-genome sequencing, by contrast, provides information about new as well as known variants; the technologies are still developing rapidly [5,6] and have provided our first glimpses of population-scale samples of hundreds of individuals [7]. Here, we discuss how samples have been chosen for studies of human history and some of the resulting sample sets, together with a number of developments in analytical approaches. We also focus on a few case studies, chosen to illustrate how whole-genome analyses compare with conclusions from mtDNA and the Y chromosome analysis, how studying other species informs our understanding of humans, and how genetic analyses themselves compare with conclusions from other sources of insights into history: archaeology, language, oral traditions or written records. We begin with examples based entirely on the analysis of modern human populations, and then move to studies that have used more diverse sources of genetic information: ancient DNA and other species associated with humans.

Sampling

If we wanted to sample an animal or plant species for genetic study, we would probably choose samples at random throughout its range. Human population

*Correspondence: cts@sanger.ac.uk
[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK
Full list of author information is available at the end of the article

sampling is never done like this. 'Related' individuals (meaning 'related in the last two to three generations') are almost always excluded; known recent migrants are also often excluded, or sometimes used as representatives of their source population; populations of particular interest to the samplers are often included. These considerations are illustrated by some of the most widely used sample sets available to the field (Figure 1a-c). We see, for example, the total lack of samples from Australia in these sets, the poor representation of India, and the predominantly migrant and admixed samples collected from the Americas. A representation of the variation in these populations (Figure 2a) reveals additional features of the samples: the high level of variation combined with sparse representation of African populations, for example. The samples chosen for the HapMap [8] and 1000 Genomes [7] projects (Figure 1b,c) reflected the medical, rather than evolutionary, emphases of these projects, and all the sampling choices were limited by political constraints. Nevertheless, these are the key samples for many inferences about human history. Individual studies may use particular additional samples, as we will see in some of the examples given, but will usually be interpreted in the context of these shared resources.

Advances in whole-genome analyses

Population genetics aims to understand the observed distribution of genetic variability and to infer histories of populations from genetic data. For this purpose, what really matters are the differences between sequences (variants). Developments in the past few years now enable millions of variants in thousands of individuals to be analyzed.

Differences can be summarized using statistics (such as allele frequencies or genetic distances) and used to quantify relationships among individuals or populations using clustering techniques, such as principal components analysis (PCA) [9]. With PCA, the information from n polymorphisms is summarized at the individual or population level with n artificial variables (components). Usually the first two or three (principal) capture much of the information and provide a picture of the relationships between the samples (Figure 2a). Model-based clustering techniques (STRUCTURE-like methods; Figure 2b) [10-12] go a little further, estimating the probability of an individual belonging to a certain genetic cluster, the ancestry coefficient. A model with k possible genetic clusters is assumed, and for each individual, k ancestry coefficients (summing to one) are calculated from the genotypes. Patterns of admixture can be visualized at the individual or population level when clusters do not coincide with populations, suggesting past demographic events such as migrations. We see, for example, the admixture in the 1000 Genomes American samples where three components are seen in most individuals (Figure 2b); k1 (light orange) represents a likely African origin, k2 (light blue) a European or West/South Asian contribution, and k5 (yellow) a Native American origin.

One way to estimate relationships that make use of individual whole-genome sequences is through the D statistic [13]. This statistic compares the sharing of derived alleles in two individuals with a third, and thus measures whether the two are equally distant from the third, or whether one is closer. This last situation is taken as evidence for greater genetic exchanges between them - for example, through admixture. More sophisticated inferences about past demographic events can be attained by fitting statistics derived from simulation under specific demographic models (for example, [14-16]) to empirical data. There are fast algorithms that can accommodate the growing mass of empirical data [16-18]. The underlying rationale is that because the parameters (such as population sizes or split times) in the simulation are known, a good fit to the empirical data indicates that the observed pattern of genetic variability might have been produced by that model. This approach, however, remains computationally intensive and it is unclear how fully the simple models used capture key elements of human history.

These analyses extract far more information from genome-wide data than from single loci such as mtDNA or the Y chromosome. Although demographic inferences have been made using single loci - for example, suggesting that 20 to 45 thousand years ago (KYA) most humans lived in South Asia [19] - such conclusions are highly dependent on the sampling strategy [20]. In any case, extant lineages from both mtDNA and the Y chromosome coalesce <200 KYA [21,22], which prevents inferences about earlier demography. By contrast, inference about effective population size back to several million years ago (MYA) has been made using whole-genome sequences [23]. This approach estimated the coalescence time of the maternal and paternal copies of each genomic segment and examined the distribution of effective population size at different time intervals inferred from these. It identified a significant reduction in population size in the past 100 KY, more marked in European and East Asian populations than in Africans, and a shared demography before that. The population size was larger between 100 KYA and 200 KYA, which might reflect population substructure at that time. Interestingly, considerable exchange between sub-Saharan Africans and Europeans/Asians was inferred until 20 to 40 KYA, consistent with some more standard models that estimated an unexpectedly low split time between Asians and Europeans of 23 KYA [16].

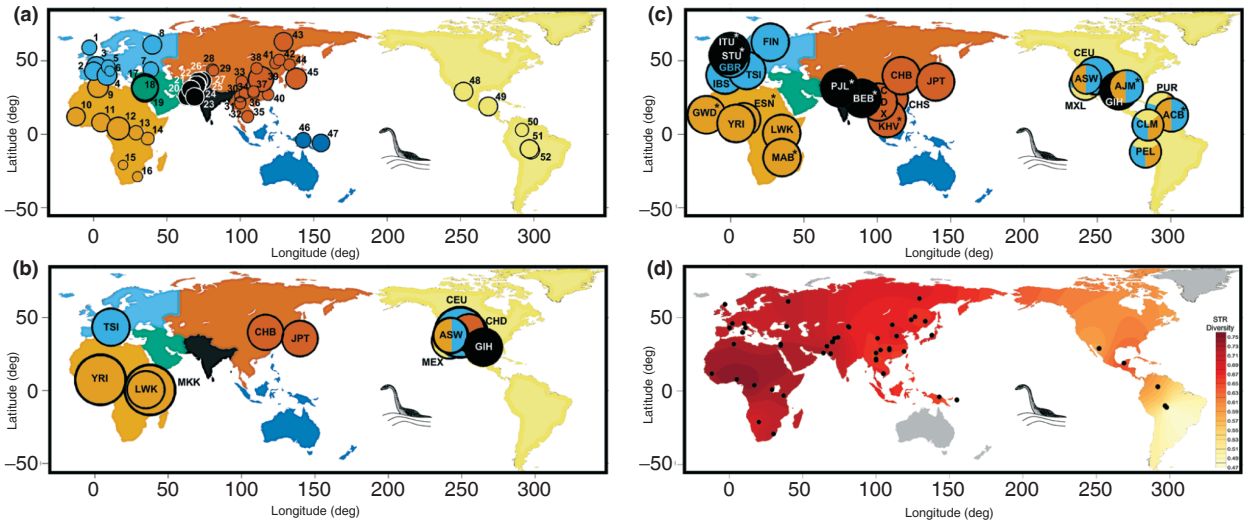

Figure 1. Widely used population samples and worldwide genetic diversity. (a) Human Genome Diversity Project (HGDP) [30], **(b)** HapMap [60], **(c)** 1000 Genomes Project [7], **(d)** Genetic diversity in the HGDP samples [26]. In (a-c), samples are represented by circles with an area proportional to the sample size, placed according to their sampling location, but colored according to the predominant continental ancestry. Numbers and acronyms identify individual samples and details are given in Additional file 1; asterisks in (c) indicate populations that are still being collected. In (d), diversity measurements [26] were interpolated between sampling sites (dots); Australia and Greenland are grayed out because of a lack of data. See Additional file 1 for further details.

Another great advantage of sequence data is that they provide an unbiased estimate of genomic variability. African individuals carry more SNPs than Europeans: 3.3 million each compared with 2.7 million according to one study [7]. However, somewhat counter-intuitively, the number of variants in a population sample of more than a few thousand is actually larger in Europeans [24,25]. This is because the European (or Asian) populations carry vast numbers of extremely rare variants that are seldom shared between individuals, a consequence of their recent explosive demographic growth.

Worldwide diversity patterns: the serial founder model and early expansion patterns

The broad pattern of global human genetic variation is well established: autosomes, the X chromosome, mtDNA and the Y chromosome all generally show higher genetic diversity in African populations than in non-Africans. In addition, non-African populations carry only a fraction of the common (and hence old) African genetic variants. Furthermore, phylogenetic trees from mtDNA, the Y chromosome and autosomal regions most commonly root in Africa, with non-African populations having subsets of the lineages. These conclusions, derived before the days of large-scale sequencing, were unsurprisingly reinforced by whole-genome sequences [7]. Such observations are readily explained by a recent and predominantly African origin for modern humans and expansion of a subgroup into the rest of the world, but detailed genetic analyses have allowed more sophisticated models to be developed.

Analyses of global genome-wide datasets - initially short tandem repeats (STRs) in the Human Genome Diversity Project (HGDP) panel (Figure 1a) - revealed a strong negative correlation between genetic diversity and migration distance from East Africa [26,27] (Figure 1d). The authors proposed a serial founder model to account for this relationship: a subgroup set out from a source population to colonize a neighboring region, expanded to form a secondary source population and the process was repeated in successive steps away from the origin. The correlation was further confirmed by whole-genome SNP data in the same panel [28] and by extensive re-sequencing data (thus largely free from ascertainment bias) in an independent sample set [29]. Although the model was established using autosomal variants, Y-STR diversity in the HGDP panel was consistent with it and showed corresponding decreases in Y-chromosomal coalescence time and effective population size, which further supported the model [30].

An extension of the model would be to identify the region within Africa that provides the best candidate for the origin; the above analyses simply assumed an origin in East Africa. Two such studies have pointed to south-western Africa origin [31,32]: for example, the ≠Khomani Bushmen from South Africa have the lowest linkage disequilibrium among the populations examined (a characteristic of an ancient population) [31]. This conclusion

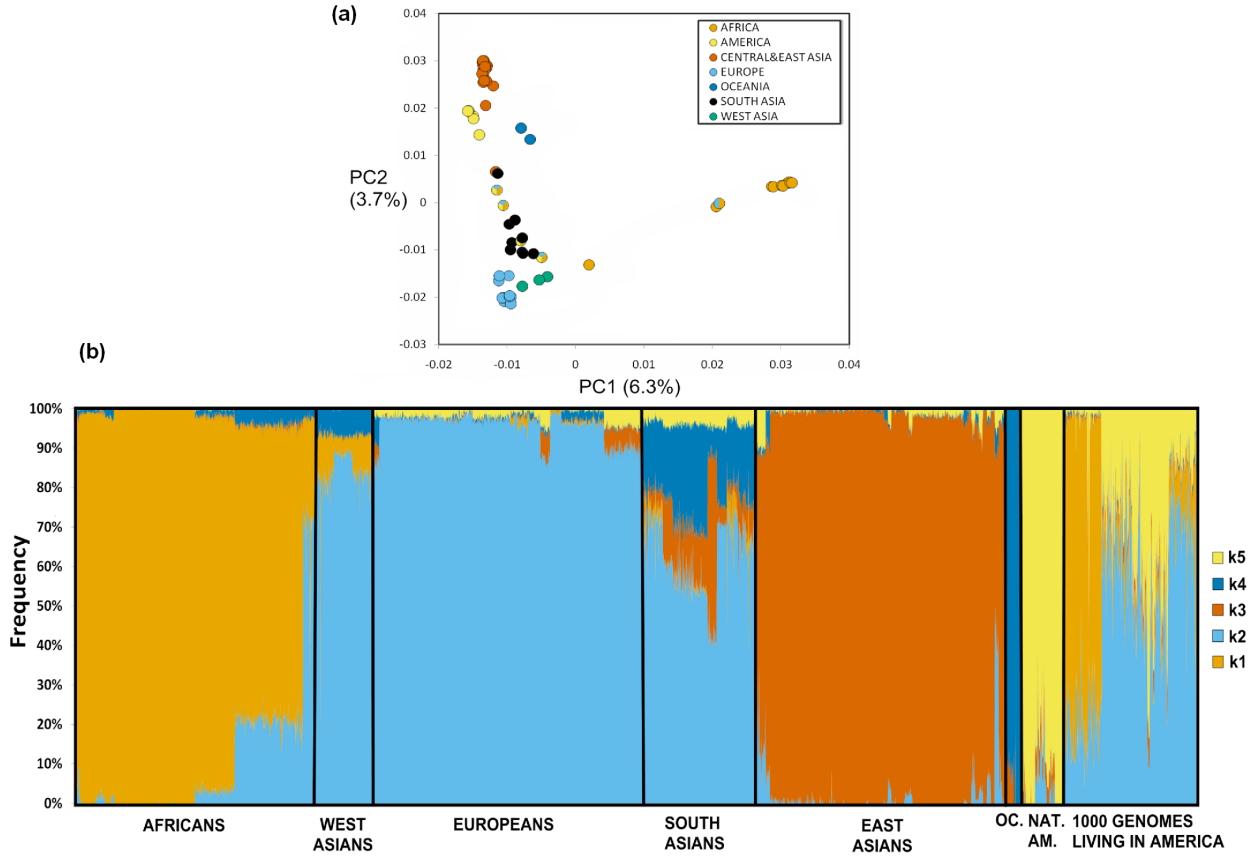

Figure 2. Representation of individual and population relationships in the combined HGDP and 1000 Genomes samples. (a) Principal components plot, where each dot represents a population, colored according to ancestry as in Figure 1. **(b)** STRUCTURE-like plot, where each thin vertical line represents an individual with inferred ancestry from five clusters, k1 to k5. Individuals are grouped by population, and the regions of origin are indicated. Oc, Oceania; Nat. Am., Native American. See Additional file 1 for further details.

must, however, be interpreted with caution because of the limited representation of East African populations in these comparisons, and the admixture and migration that have occurred in Africa since the origin, influencing the genetic properties of the modern populations analyzed.

This genetic model has stimulated new analyses of non-genetic datasets. An examination of 37 morphometric characteristics in 4,666 male skulls drawn from 105 populations worldwide revealed a decrease in phenotypic variability mirroring the loss in genetic diversity and suggested central/southern Africa as the origin [33]. Similarly, the number of phonemes used in a global sample of 504 languages was also well explained by a serial founder model of expansion from an inferred origin in southern/western Africa [34]. These strong patterns of human genetic, morphological and linguistic variation support a single African origin for most of human diversity, but do not preclude low levels of admixture with archaic humans [13,35] - a 'leaky replacement' model where most but not all archaic diversity was replaced by an expansion from Africa 50 to 70 KYA.

A genome sequence derived from a 100-year-old Australian Aboriginal hair sample has provided new insights into the early migration patterns [36]. Using a variant of the *D* statistic (designated D_{4P}), which assesses allele sharing between four individual genomes, the authors [36] investigated the sharing patterns between African, European, East Asian (Han Chinese) and Australian Aboriginal genomes. They observed more sharing between East Asian and European than between East Asian and Aboriginal genomes, and so proposed an initial split after the exit from Africa between the Aboriginal ancestors on the one hand, and the ancestors of modern Europeans and Chinese on the other, a conclusion supported by an independent study based on genotyping a much larger number of individuals [37].

The Jewish Diaspora

Population movements have, of course, continued since the initial expansion out of Africa, and one, initiated by the destruction of the First Temple in Jerusalem by Nebuchadnezzar II of Babylon (c. 586 BCE) and known

as the Jewish Diaspora, has been studied intensively using both historical and genetic data. The Diaspora led to the dispersal of Jewish people from the Levant to many parts of the world, and traditionally three main groups were recognized: Ashkenazim (Eastern European Jewish populations), Sephardim (Southern European Jewish populations) and Mizrahim (Middle Easterners) [38]. Other Jewish groups outside Israel include Ethiopian and Indian Jews. 'Jewishness' is inherited from the mother, so a shared Middle Eastern origin for mtDNA contrasted with a more diverse origin for Y chromosomes might be expected, at least in exogamous communities, and has indeed been reported [39,40].

Two independent studies have examined the relationships between Jewish and surrounding non-Jewish groups ('hosts') using genome-wide SNP data [38,41]. Both came to the conclusion that although the whole-genome data show detectable introgression from the local populations into the genetic pool of the Diaspora people, both autosomal and uniparental markers point towards a clear Middle Eastern origin for most of the individuals studied. In particular, the mtDNA and genomic markers show strong evidence for a Middle Eastern origin, matching the expectations from the matrilineal pattern of Jewishness (Figure 3). Furthermore, the genomic information provided for the first time a signature of the demographic expansion documented as the 'demographic miracle' of the Ashkenazi Jews during the past 500 years [38], whereas little evidence for inter-Diaspora gene flow was observed. The authors [38] write: 'Admixture with surrounding populations had an early role in shaping world Jewry, but, during the past 2,000 years, may have been limited by religious law as Judaism evolved from a proselytizing to an inward-looking religion.'

However, a major exception to this scenario was provided by the Ethiopian and Indian Jews. These two groups show a much greater extent of host genetic introgression, effectively clustering together with the local non-Jewish populations in both PCA and STRUCTURE-like analyses (Figure 3c). The genomic clustering explains observations from the mtDNA data [39,40] in which a high proportion of local maternal lineages were shared with the Ethiopian and Indian populations (Figure 3a), potentially as a result of more relaxed criteria of self-identification in these Jewish communities. This pictures the migration of these people as a star-shaped pattern centered on the Levant, with founder effects and genetic drift [42] acting differently on the individual branches.

To further emphasize the importance of social rules in shaping the genetic landscape of a population, we can make a parallel with the more recent (about 1,000 YA) Roma (Gypsy) expansion from a North Indian source population towards South-Western Europe. Although the available data still rely on a limited set of markers [43-45],

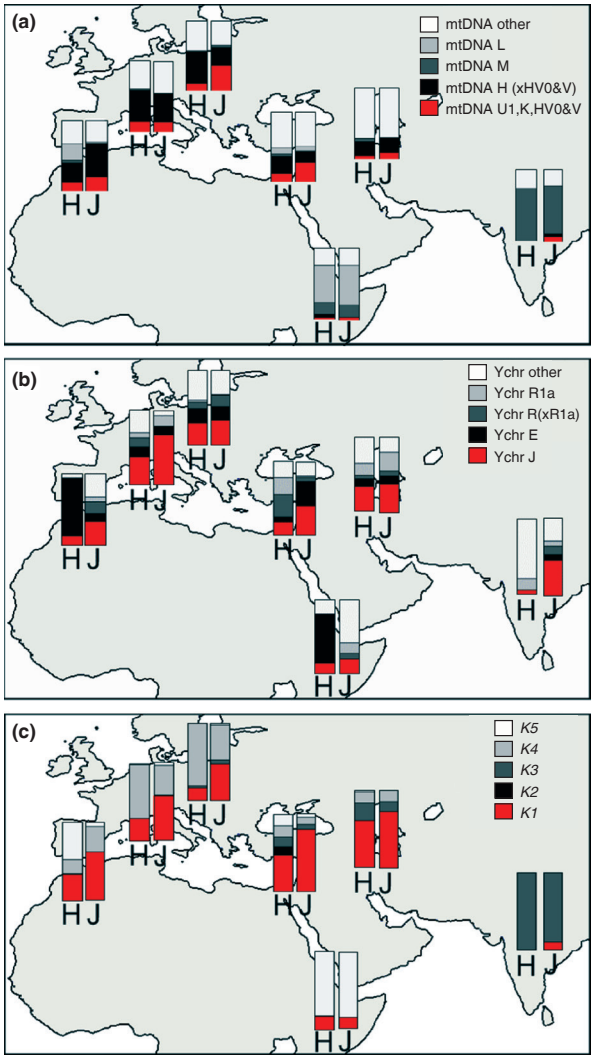

Figure 3. Genetic relationships between Jewish Diaspora (J) and nearby non-Jewish (host, H) populations. (a) mtDNA analysis; **(b)** Y-chromosomal analysis; **(c)** autosomal SNP genotype analysis. Lineages (a,b) or ancestry components (c) were divided into five classes, and in each panel one class (red) is present at higher frequency in most Jewish populations (J) than in the corresponding host population (H), illustrating a genetic heritage component shared by most Jewish populations. See Additional file 1 for further details.

the more flexible inheritance of Roma status is reflected in a serial dilution of the source (Indian) gene pool, characterized by stepwise admixture with local 'host' populations for both uniparental and autosomal markers. It is thus striking that the different migration pattern and relaxation in self-identification criteria allowed a greater extent of mixture of Roma with hosts in half of the time of the Jewish Diaspora.

The first Greenlanders

Greenland was among the last places on the earth to be reached by humans and is currently inhabited by the

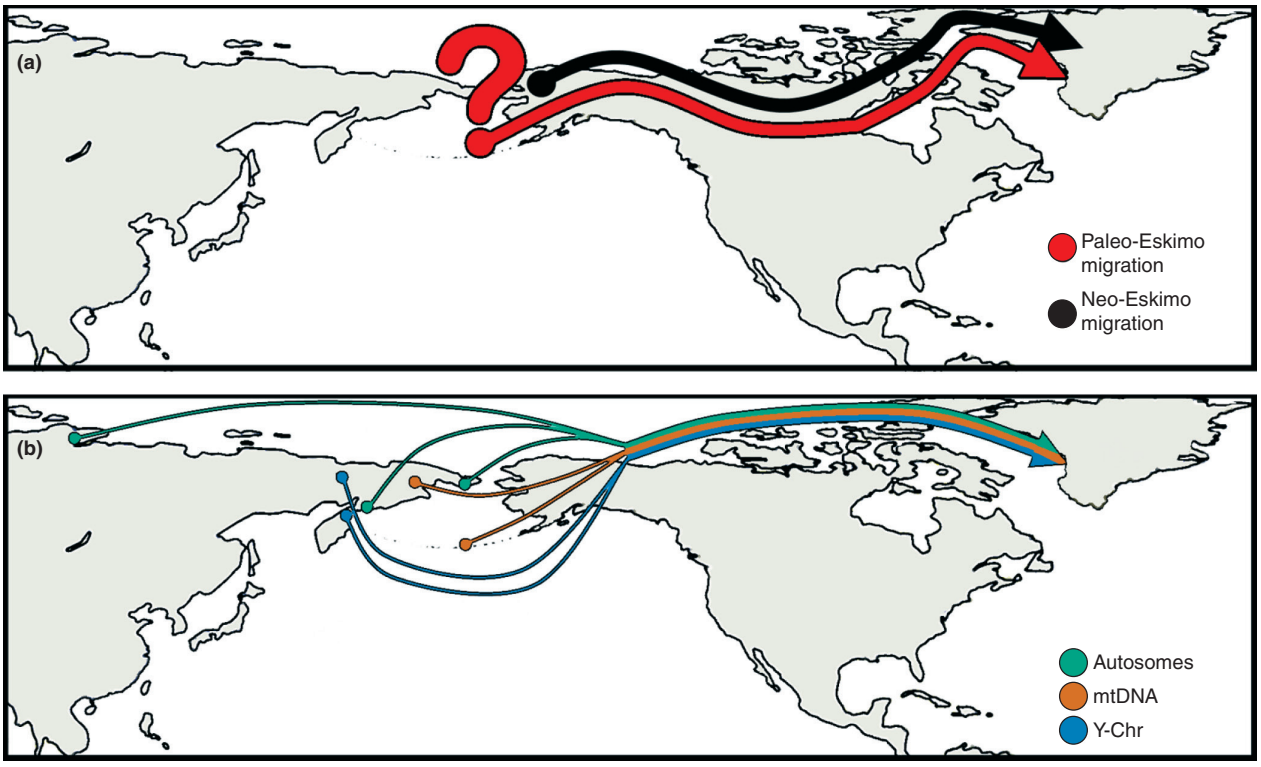

Figure 4. Archaeological and genetic evidence for the peopling of Greenland. (a) Archaeological evidence suggests two waves of migration to Greenland: Paleo-Eskimo (Saqqaq) about 4,500 YA (red) and Neo-Eskimo (Inuit) about 1,000 YA (black). The place of origin of the Saqqaq migration was unknown. **(b)** Genetic evidence suggests a Saqqaq origin in Siberia. The modern populations most similar to the Saqqaq mtDNA, Y chromosome and autosomes are all located in Siberia or the Aleut Islands. These maps represent initial and final locations, but not the route taken.

Inuit, who have their closest relatives in Canadian Inuit. Archaeological evidence suggests two waves of migration into Greenland from Siberia and Alaska (Old and New World Beringia, respectively; Figure 4a). The first migration started about 4.5 KYA and involved the Paleo-Eskimo populations, which included at least three different groups identified by distinct archaeological cultures (Saqqaq, Pre-Dorset and Dorset). The Inuit arrived in Greenland as a second wave (Neo-Eskimo) that took place about 1,000 YA [46]. Although both genetics and archaeology support a recent (1,000 YA) Beringian origin for the Neo-Eskimo populations [47], the origin of the Paleo-Eskimo groups was unknown, apart from the geographical proximity of Native American and Beringian populations to Greenland. The availability of an exceptionally well-preserved hair sample from an ancient Paleo-Eskimo individual from the Saqqaq culture, originating from the west coast of Greenland, allowed this issue to be investigated using genetics.

The hair dated to around 4,000 YA and provided both a high-coverage mtDNA sequence [48] and 20x coverage genomic sequence [49]. Analyses of these data suggested that this individual was not a likely ancestor of the current Greenland Inuit, confirming the theory of different waves of migration into Greenland, the last of which eventually replaced the previous populations. Indeed, the Paleo-Eskimo Saqqaq mtDNA was attributed to the D2a1 haplogroup, which has not been found in the Inuit [48]. The closest lineage to D2a1 is currently found in the Aleut and Sireniky-Yuit populations from the Aleut Islands and extreme East Siberia, respectively, suggesting an Old World Beringian origin for the maternal ancestry of the Saqqaq (Figure 4b). The Y lineage was Q1a*, now known from the Koryaks and Yukaghir of East Siberia [50]. This connection with Siberian populations was confirmed by the autosomal data. Indeed, clustering analysis of polymorphisms from the Saqqaq genome (in comparison with 35 populations from Europe, Asia and America) showed shared ancestry with three Siberian populations (Figure 4b) and excluded detectable admixture or shared origin with Western European and Native American populations. In addition to this ancestry information, analyses of functional variants in genomic DNA provided some insight into phenotypic traits of the Saqqaq individual [49], suggesting blood type A+, brown eyes, dry earwax and dark hair, with, ironically in view of the source material, a predisposition to baldness.

Lessons from associated species

The study of human DNA variation is not the only way to use genetics to decipher our history. Insights can also come from the study of organisms that have established long-term relationships with us (such as parasites and pathogens), providing an independent perspective on their host's evolutionary history. Broadly speaking, samples are collected in different geographical areas and their phylogeographic relationship is investigated to locate their probable origin (usually the place where the parasite's genetic diversity is greatest) and characterize their spread by describing and dating the branching pattern and/or analyzing population-genetic parameters.

As expected, in many cases a strong correlation has been found between parasite origins and spread and both ancient human migrations, such as the out-of-Africa movement [51], and more recent migrations, such as the Austronesian expansion [52] or along the Silk Road trade route [53]. In addition, louse phylogenetic history has also helped to reconstruct aspects of our appearance that do not fossilize, a challenging task using genotype information, and even more so for ancestors without genotype information. Chimpanzees have only head lice and gorillas only pubic, whereas humans have both. Head lice seem to have co-evolved with their respective hosts since the chimpanzee-human split 6 to 7 MYA. By contrast, the divergence time between gorilla and human pubic lice is substantially lower than the gorilla-human split (about 8 MYA), suggesting that pubic lice originated in gorillas and switched to the human lineage around 3.3 MYA [54]. This may indicate that, by that time, the human body was partially free from hair and a new niche was available. Despite the ancient loss of body hair, according to this interpretation, it took a long time for humans to introduce clothes, according to another study of lice. *Pediculus* head and clothing lice split around 70 to 80 KYA [55,56]. If the split occurred as soon as the new clothing niche became available, this date provides us with an estimate of the time of the introduction of clothes.

Conclusions and future directions

What can we learn from these examples about the information contributed by different genetic loci, and the comparison between genetic and non-genetic sources? It is useful to consider this question by time period. In the last about 50 KY, mtDNA, the Y chromosome and whole-genome data are all informative, and the agreements between them are more striking than the disagreements. For example, in both the Diaspora and Saqqaq genome studies, the genome-wide data extended, rather than overturned, the conclusions from the uniparental loci. This does not have to be the case, and is not always found: a family from England, for example, unexpectedly carried a Y chromosome typical of sub-Saharan Africa [57], but the rarity of such examples - the strong correlation between the histories of different genomic regions - provides evidence for marked geographical structure prevailing over this period. Nevertheless, some unresolved issues remain: genome-wide data suggest extensive gene flow between sub-Saharan Africans and non-Africans [23], whereas uniparental data do not [1,2].

Further back into the past, the uniparental loci provide less and less information because fewer lineages have survived to the present: all mtDNA and Y lineages outside Africa trace back to a single ancestor about 50 to 70 KYA, for example. So for understanding the early expansions, the main insights have come from genome-wide data (for example, [36]), and these are our only source of genetic data for the period >200 KYA.

In general, non-genetic data provide far more detailed information for the last few millennia, and the genetic inferences are largely consistent with them (such as the Diaspora). Genetics can add information, such as about the origins of the Paleo-Eskimos, that has not been available from other sources. Again, there are unresolved issues - for example, the genetic contact inferred between populations 20 to 40 KYA contrasts with the distinct regional archaeological records for this period [58]. And genetics can generate hypotheses, such as the adoption of clothing 70 to 80 KYA, that can now be tested by archaeologists.

As whole-genome sequencing becomes cheaper and more routine, it can more readily be used to address questions of evolutionary interest: the complex demography in Africa and the peopling of the Americas and the Pacific, for example. When did the ancestral populations split, how much subsequent gene flow was there, what patterns of migration can be identified? There may be admixed populations in which mtDNA and the Y chromosome from one ancestral source have been entirely lost by drift, but ancestral information can be reconstructed from the autosomal regions: the admixed American populations in the 1000 Genomes Project are allowing this possibility to be tested. Perhaps we can study the genomics of pre-contact Native American or Tasmanian populations, as proposed by the Taíno Genome Project [59] investigating the legacy of a pre-Colombian Puerto Rican population using 1000 Genomes data. Our history is encoded in our genomes, each much smaller than a grain of sand, and we are only now collecting the data and developing the tools to decipher this rich source of information.

Additional files

Additional file 1: Population samples and geographic coordinates used in Figures 1 and 2, together with additional information on the construction of Figures 2 and 3. [61]

Acknowledgements
Our work is supported by The Wellcome Trust (grant number 098051), and additionally an EMBO Short-term Fellowship (ASTF 324-2010) to VC, and the Cambridge European Trust, a Domestic Research Scholarship and Emmanuel College, Cambridge, UK (LP). We thank Jean McEwen for information about the 1000 Genomes samples.

Author details
[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [2]Institute of Genetics and Biophysics 'A. Buzzati-Traverso', National Research Council (CNR), Naples, Italy. [3]The Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Fitzwilliam Street, Cambridge CB2 1QH, UK.

Published: 21 November 2011

References

1. Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ: **Harvesting the fruit of the human mtDNA tree.** *Trends Genet* 2006, **22:**339-345.
2. Jobling MA, Tyler-Smith C: **The human Y chromosome: an evolutionary marker comes of age.** *Nat Rev Genet* 2003, **4:**598-612.
3. Underhill PA, Kivisild T: **Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations.** *Annu Rev Genet* 2007, **41:**539-564.
4. Ragoussis J: **Genotyping technologies for genetic research.** *Annu Rev Genomics Hum Genet* 2009, **10:**117-133.
5. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data.** *Nat Rev Genet* 2011, **12:**443-451.
6. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11:**31-46.
7. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.
8. The International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426:**789-796.
9. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2:**e190.
10. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155:**945-959.
11. Tang H, Coram M, Wang P, Zhu X, Risch N: **Reconstructing genetic ancestry blocks in admixed individuals.** *Am J Hum Genet* 2006, **79:**1-12.
12. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res* 2009, **19:**1655-1664.
13. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, *et al.*: **A draft sequence of the Neandertal genome.** *Science* 2010, **328:**710-722.
14. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: **Calibrating a coalescent simulation of human genome sequence variation.** *Genome Res* 2005, **15:**1576-1583.
15. Beaumont MA, Zhang W, Balding DJ: **Approximate Bayesian computation in population genetics.** *Genetics* 2002, **162:**2025-2035.
16. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD: **Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data.** *PLoS Genet* 2009, **5:**e1000695.
17. Marjoram P, Wall JD: **Fast "coalescent" simulation.** *BMC Genet* 2006, **7:**16.
18. McVean GA, Cardin NJ: **Approximating the coalescent with recombination.** *Philos Trans R Soc Lond B Biol Sci* 2005, **360:**1387-1393.
19. Atkinson QD, Gray RD, Drummond AJ: **mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory.** *Mol Biol Evol* 2008, **25:**468-474.
20. Gunnarsdottir ED, Li M, Bauchet M, Finstermeier K, Stoneking M: **High-throughput sequencing of complete human mtDNA genomes from the Philippines.** *Genome Res* 2011, **21:**1-11.
21. Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, Rosset S: **The dawn of human matrilineal diversity.** *Am J Hum Genet* 2008, **82:**1130-1140.
22. Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sellitto D, Scozzari R: **A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa.** *Am J Hum Genet* 2011, **88:**814-818.
23. Li H, Durbin R: **Inference of human population history from individual whole-genome sequences.** *Nature* 2011, **475:**493-496.
24. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR, Wheeler DA, Sabo A, Lusk C, Weiss KG, Akbar H, Cree A, Hawes AC, Newsham I, Varghese RT, Villasana D, Gross S, Joshi V, Santibanez J, Morgan M, Chang K, Iv WH, Templeton AR, Boerwinkle E, Gibbs R, Sing CF: **Deep resequencing reveals excess rare recent variants consistent with explosive population growth.** *Nat Commun* 2010, **1:**131.
25. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD: **Demographic history and rare allele sharing among human populations.** *Proc Natl Acad Sci U S A* 2011, **108:**11983-11988.
26. Prugnolle F, Manica A, Balloux F: **Geography predicts neutral genetic diversity of human populations.** *Curr Biol* 2005, **15:**R159-R160.
27. 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 U S A* 2005, **102:**15942-15947.
28. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM: **Worldwide human relationships inferred from genome-wide patterns of variation.** *Science* 2008, **319:**1100-1104.
29. Luca F, Hudson RR, Witonsky DB, Di Rienzo A: **A reduced representation approach to population genetic analyses and applications to human evolution.** *Genome Res* 2011, **21:**1087-1098.
30. Shi W, Ayub Q, Vermeulen M, Shao RG, Zuniga S, van der Gaag K, de Knijff P, Kayser M, Xue Y, Tyler-Smith C: **A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations.** *Mol Biol Evol* 2010, **27:**385-393.
31. Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodríguez-Botigué L, Ramachandran S, Hon L, Brisbin A, Lin AA, Underhill PA, Comas D, Kidd KK, Norman PJ, Parham P, Bustamante CD, Mountain JL, Feldman MW: **Hunter-gatherer genomic diversity suggests a southern African origin for modern humans.** *Proc Natl Acad Sci U S A* 2011, **108:**5154-5162.
32. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM: **The genetic structure and history of Africans and African Americans.** *Science* 2009, **324:**1035-1044.
33. Manica A, Amos W, Balloux F, Hanihara T: **The effect of ancient population bottlenecks on human phenotypic variation.** *Nature* 2007, **448:**346-348.
34. Atkinson QD: **Phonemic diversity supports a serial founder effect model of language expansion from Africa.** *Science* 2011, **332:**346-349.
35. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Pääbo S: **Genetic history of an archaic hominin group from Denisova Cave in Siberia.** *Nature* 2010, **468:**1053-1060.
36. Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, Kivisild T, Zhai W, Eriksson A, Manica A, Orlando L, De La Vega FM, Tridico S, Metspalu E, Nielsen K, Ávila-Arcos MC, Moreno-Mayar JV, Muller C, Dortch J, Gilbert MT, Lund O, Wesolowska A, Karmin M, Weinert LA, Wang B, Li J, *et al.*: **An Aboriginal Australian genome reveals separate human dispersals into Asia.** *Science* 2011, **334:**94-98.
37. Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC, Jinam TA, Phipps ME, Saitou N, Wollstein A, Kayser M, Pääbo S, Stoneking M: **Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania.** *Am J Hum Genet* 2011, **89:**516-528.
38. Atzmon G, Hao L, Pe'er I, Velez C, Pearlman A, Palamara PF, Morrow B, Friedman E, Oddoux C, Burns E, Ostrer H: **Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry.** *Am J Hum Genet* 2010, **86:**850-859.
39. Behar DM, Metspalu E, Kivisild T, Rosset S, Tzur S, Hadid Y, Yudkovsky G, Rosengarten D, Pereira L, Amorim A, Kutuev I, Gurwitz D, Bonne-Tamir B, Villems R, Skorecki K: **Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora.** *PLoS One* 2008, **3:**e2062.
40. Non AL, Al-Meeri A, Raaum RL, Sanchez LF, Mulligan CJ: **Mitochondrial DNA**

reveals distinct evolutionary histories for Jewish populations in Yemen and Ethiopia. *Am J Phys Anthropol* 2011, **144:**1-10.
41. Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, Rootsi S, Chaubey G, Kutuev I, Yudkovsky G, Khusnutdinova EK, Balanovsky O, Semino O, Pereira L, Comas D, Gurwitz D, Bonne-Tamir B, Parfitt T, Hammer MF, Skorecki K, Villems R: **The genome-wide structure of the Jewish people.** *Nature* 2010, **466:**238-242.
42. Bray SM, Mulle JG, Dodd AF, Pulver AE, Wooding S, Warren ST: **Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population.** *Proc Natl Acad Sci U S A* 2010, **107:**16222-16227.
43. Gusmao A, Gusmao L, Gomes V, Alves C, Calafell F, Amorim A, Prata MJ: **A perspective on the history of the Iberian gypsies provided by phylogeographic analysis of Y-chromosome lineages.** *Ann Hum Genet* 2008, **72:**215-227.
44. Gusmao A, Valente C, Gomes V, Alves C, Amorim A, Prata MJ, Gusmao L: **A genetic historical sketch of European Gypsies: The perspective from autosomal markers.** *Am J Phys Anthropol* 2010, **141:**507-514.
45. Mendizabal I, Valente C, Gusmão A, Alves C, Gomes V, Goios A, Parson W, Calafell F, Alvarez L, Amorim A, Gusmão L, Comas D, Prata MJ: **Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective.** *PLoS One* 2011, **6:**e15988.
46. Damas D: *Handbook of North American Indians. Volume 5.* Washington DC: Smithsonian Institution; 1984.
47. Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S: **mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion.** *Am J Hum Genet* 2000, **67:**718-726.
48. Gilbert MT, Kivisild T, Grønnow B, Andersen PK, Metspalu E, Reidla M, Tamm E, Axelsson E, Götherström A, Campos PF, Rasmussen M, Metspalu M, Higham TF, Schwenninger JL, Nathan R, De Hoog CJ, Koch A, Møller LN, Andreasen C, Meldgaard M, Villems R, Bendixen C, Willerslev E: **Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland.** *Science* 2008, **320:**1787-1789.
49. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert MT, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, *et al.*: **Ancient human genome sequence of an extinct Palaeo-Eskimo.** *Nature* 2010, **463:**757-762.
50. Malyarchuk B, Derenko M, Denisova G, Maksimov A, Wozniak M, Grzybowski T, Dambueva I, Zakharov I: **Ancient links between Siberians and Native Americans revealed by subtyping the Y chromosome haplogroup Q1a.** *J Hum Genet* 2011, **56:**583-588.
51. Tanabe K, Mita T, Jombart T, Eriksson A, Horibe S, Palacpac N, Ranford-Cartwright L, Sawai H, Sakihama N, Ohmae H, Nakamura M, Ferreira MU, Escalante AA, Prugnolle F, Björkman A, Färnert A, Kaneko A, Horii T, Manica A, Kishino H, Balloux F: ***Plasmodium falciparum*** **accompanied the human expansion out of Africa.** *Curr Biol* 2010, **20:**1283-1289.
52. Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu JY, Maady A, Bernhöft S, Thiberge JM, Phuanukoonnon S, Jobb G, Siba P, Graham DY, Marshall BJ, Achtman M: **The peopling of the Pacific from a bacterial perspective.** *Science* 2009, **323:**527-530.
53. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, Cui Y, Thomson NR, Jombart T, Leblois R, Lichtner P, Rahalison L, Petersen JM, Balloux F, Keim P, Wirth T, Ravel J, Yang R, Carniel E, Achtman M: ***Yersinia pestis*** **genome sequencing identifies patterns of global phylogenetic diversity.** *Nat Genet* 2010, **42:**1140-1143.
54. Reed DL, Light JE, Allen JM, Kirchman JJ: **Pair of lice lost or parasites regained: the evolutionary history of anthropoid primate lice.** *BMC Biol* 2007, **5:**7.
55. Kittler R, Kayser M, Stoneking M: **Molecular evolution of *Pediculus humanus* and the origin of clothing.** *Curr Biol* 2003, **13:**1414-1417.
56. Toups MA, Kitchen A, Light JE, Reed DL: **Origin of clothing lice indicates early clothing use by anatomically modern humans in Africa.** *Mol Biol Evol* 2011, **28:**29-32.
57. King TE, Parkin EJ, Swinfield G, Cruciani F, Scozzari R, Rosa A, Lim SK, Xue Y, Tyler-Smith C, Jobling MA: **Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy.** *Eur J Hum Genet* 2007, **15:**288-293.
58. Jobling MA, Hurles M, Tyler-Smith C: *Human Evolutionary Genetics.* New York and Abingdon: Garland Science; 2004.
59. **The Taino Genome Project** [https://sites.google.com/a/upr.edu/dna-lab/1000genomes/the-taino-genome-project]
60. The International HapMap 3 Consortium: **Integrating common and rare genetic variation in diverse human populations**. *Nature* 2010, **467:**52-58.
61. 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, **38:**904-909.

doi:10.1186/gb-2011-12-11-234
Cite this article as: Colonna V, *et al.*: A world in a grain of sand: human history from genetic data. *Genome Biology* 2011, **12:**234.

A world in a grain of sand: human history from genetic data

Vincenza Colonna, Luca Pagani, Yali Xue, Chris Tyler-Smith

Supplementary Information

Population samples and geographic coordinates used in Figures 1 and 2:

CODE	HAPMAP POPULATION	N	LAT	LONG
CEU	Utah residents (CEPH) with N and W European ancestry	180	39.8	-111.1
TSI	Toscans in Italy	100	43.4	11.1
ASW	African ancestry in Southwest USA	90	37.4	-117.4
LWK	Luhya in Webuye, Kenya	100	0.6	34.8
MKK	Maasai in Kinyawa, Kenya	180	0.4	37.6
YRI	Yoruba in Ibadan, Nigeria	180	7.4	3.9
GIH	Gujarati Indians in Houston, Texas	100	29.8	-95.4
CHB	Han Chinese in Beijing, China	90	39.9	116.4
CHD	Chinese in Metropolitan Denver, Colorado	100	39.7	-105.0
JPT	Japanese in Tokyo, Japan	91	35.7	139.7
MEX	Mexican ancestry in Los Angeles, California	90	34.0	-118.2

CODE	1000 GENOMES POPULATION	N	LAT	LONG
CEU	Utah residents (CEPH) with N and W European ancestry	100	39.8	-111.1
FIN	Finnish from Finland	100	62.6	27.0
GBR	British from England and Scotland	100	53.6	-1.6
TSI	Toscani in Italia	100	43.4	11.1
IBS*	Iberian populations in Spain	100	40.2	-3.7
ACB*	African Caribbean in Barbados	79	13.1	-59.6
ASW	African Ancestry in Southwest US	62	37.4	-117.4
AJM*	African American in Jackson, MS	79	32.3	-90.2
LWK	Luhya in Webuye, Kenya	100	0.6	34.8
YRI	Yoruba in Ibadan, Nigeria	100	7.4	3.9
MAB*	Malawian in Blantyre, Malawi	100	-15.8	35.0
ESN*	Esan in Nigeria	100	8.5	-11.7
GWD*	Gambian in Western Division, The Gambia	100	13.2	-16.6
GIH	Gujarati Indian in Houston, TX	100	29.8	-95.4
BEB*	Bengali in Bangladesh	100	24.4	89.7
ITU*	Indian Telegu in the UK	100	51.5	-0.2
STU*	Sri Lankan Tamil in UK	100	51.5	-0.2
PJL*	Punjabi in Lahore, Pakistan	100	31.5	74.3
CHB	Han Chinese in Beijing, China	100	39.9	116.4
CHS	Han Chinese South	100	24.6	108.1
JPT	Japanese in Toyko, Japan	100	35.7	139.7
CDX*	Chinese Dai in Xishuangbanna	100	22.0	100.8
KHV*	Kinh in Ho Chi Minh City, Vietnam	100	10.8	106.7
CLM	Colombian in Medellin, Colombia	70	6.2	-75.6
PEL*	Peruvian in Lima, Peru	70	-12.0	-77.1
MXL	Mexican Ancestry in Los Angeles, CA	70	34.0	-118.2
PUR	Puerto Rican in Puerto Rico	70	18.2	-66.3

CODE	HGDP POPULATION	N	LAT	LONG
1	Orcadian	16	59	-3
2	French_Basque	24	43	0
3	French	29	46	2
4	Sardinian	28	40	9
5	North_Italian	14	46	10
6	Tuscan	8	43	11
7	Adygei	17	44	39
8	Russian	25	61	40
9	Mozabite	30	32	3
10	Mandenka	24	12	-12
11	Yoruba	25	8	5
12	Biaka_Pygmies	36	4	17
13	Mbuti_Pygmies	15	1	29
14	San	7	-3	37
14	Bantu_N.E.	12	-21	20
16	Bantu Speakers	8	-29	30
17	Druze	48	32	35
18	Palestinian	51	32	35
19	Bedouin	49	31	35
20	**Makrani**	**25**	**26**	**64**
21	**Balochi**	**25**	**30.5**	**66.5**
22	**Brahui**	**25**	**30.5**	**66.5**
23	**Sindhi**	**25**	**25.5**	**69**
24	**Hazara**	**25**	**33.5**	**70**
25	**Pathan**	**25**	**33.5**	**70.5**
26	**Kalash**	**25**	**36**	**71.5**
27	**Burusho**	**25**	**36.5**	**74**
28	Uygur	10	44	81
29	Xibo	9	43.5	81.5
30	Naxi	10	26	100
31	Lahu	10	22	100
32	Dai	10	21	100
33	Tu	10	36	101
34	Yizu	10	28	103
35	Cambodians	11	12	105
36	Tujia	10	29	109
37	Miaozu	10	28	109
38	Mongola	10	45	111
39	Han	45	37.5	114
40	She	10	27	119
41	Daur	10	48.5	124
42	Oroqen	10	50.5	126.5
43	Yakut	25	63	129.5
44	Hezhen	10	47.5	133.5
45	Japanese	31	38	138
46	Papuan	17	-4	143
47	NAN_Melanesian	22	-6	155
48	Pima	25	29	252
49	Maya	25	19	269
50	Colombians	13	3	292
51	Karitiana	24	-10	297
52	Surui	21	-11	298

Figure 2:

Both PCA and STRUCTURE-like plots were constructed using the subset of genomic positions genotyped in common for the HGDP [28], HapMap [60] and 1000 Genomes samples [7]. The samples were pooled and the genomic positions pruned according to the global LD structure (as described in the ADMIXTURE user manual [12]. A Principal Components Analysis was then performed on the pruned pooled dataset using EIGENSTRAT [61], plotting PC1 against PC2. The variance explained by these PCs is shown in brackets (Figure 2a). A STRUCTURE-like analysis was performed on the same dataset using ADMIXTURE [12] and the k=5 output is displayed in Figure 2b.

Figure 3:

The mtDNA and Y-Chr data provided within [41] were plotted for Hosts and Jewish populations according to the seven geographic regions displayed on the map (Ethiopia, Morocco, Southern Europe, North-Eastern Europe, Middle East, Central Asia and India), corresponding to the main Jewish groups (Ethiopia, North Africa, Sephardim, Ashkenazim, Israel, Mizrahim and India) respectively. The genomic data from the same populations were analyzed using ADMIXTURE [12] with k=5 and the frequencies of the five clusters plotted as described above. In each plot, the red component is the one that best summarizes the Jewish Diaspora.

Appendix 2 “The proportion of human DNA within buccal swab extracts before and after whole genome amplification”

The proportion of human DNA within buccal swab extracts before and after whole genome amplification

Luca Pagani[1,2], Qasim Ayub[2]

AFFILIATIONS:

1: Division of Biological Anthropology, University of Cambridge, Cambridge, Cambridgeshire, UK

2: The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, Hinxton, Cambridgeshire, UK

ABSTRACT

BACKGROUND: The DNA extracted from the mucosal cells of the inner cheek (buccal swab) are a widely used and convenient non-invasive source for human studies, but its potentially reduced amount and low quality (elevated contamination and fragmented molecules) raise questions about its applications in genetic studies.
METHODOLOGY: Here we describe a quick method to assess the extent of contamination in both genomic and whole genome amplified (WGA) buccal swab DNA. This method is based on measurement of the total DNA concentration using both PicoGreen staining and gel electrophoresis, assessment of its quality by long-template PCR, and quantification of its human content by qPCR.
CONCLUSIONS: We show that 23.3 ± 3.4 µg of DNA via WGA from DNA extracted from buccal swabs with human DNA accounting for up to 42% on average. *When working with WGA DNA extracted from buccal swabs, we recommend that at least three times more DNA be used than when the DNA is derived from other sources such as a lymphoblastoid cell lines.*

INTRODUCTION

The mucosal cells of the inner cheek are a quick and widely-used non-invasive source of biological material, particularly for sampling DNA from humans. The cells collected after brushing the buccal mucosa with a small stick (Buccal Swab) can yield sufficient DNA to allow several kinds of genetic tests to be performed without requiring any painful, invasive procedure and eliminating the emotional and social issues that make many volunteers reluctant to donate blood [1].

Despite the convenience and wide use of this sampling method, many of the variables associated with it remain poorly understood. The average amount of DNA extracted from a buccal swab is low (400-600 ng per swab), and often does not allow large-scale genotyping [2] or whole genome sequence analysis to be performed. Additionally, the extent of contamination by DNA from food or other species that live on the mucosal surface has been little studied. While it is possible to circumvent the limitation in the amount of DNA through whole-genome amplification (WGA) of the extracted DNA, the contamination will still affect the proportion and concentration of human DNA in the sample. Therefore, an accurate estimation of the non-human contamination of buccal swab DNA prior to and after WGA is needed in order to correctly assess the amount of human DNA available for genetic analyses. Despite the availability of extraction methods [3] that decrease the extent of contaminants, it is not uncommon to deal with existing DNA collections extracted from buccal swabs or to be forced to rely on standard extraction kits during field-work,.

The present study was designed to estimate the ratio of human versus total DNA extracted from buccal swabs before and after WGA. The aim of this work is not only to provide a tool to extend the quality controls performed as part of any study using buccal samples, but also to indicate what can or cannot be expected when working on DNA extracted from these swabs.

To assess the overall integrity of the samples, a set of long template PCRs was performed on DNA from six buccal swabs and a control genomic sample. To measure the extent of the bacterial/food contamination, a direct measurement of the total dsDNA was performed together with a qPCR (with human-specific primers) using as template a gene known to be single copy within the human genome. The difference in number of cycles between each buccal swab DNA sample and the same amount of control DNA gave a measure of the ratio of human/total DNA per each sample (see materials and methods). The same steps were performed after WGA and the effect of WGA evaluated.

RESULTS AND DISCUSSION

The amount of DNA extracted from buccal swabs in the present study was evaluated in two independent ways. First, aliquots were analysed by gel electrophoresis and ethidium bromide staining, and compared with standards of known concentration. This method is imprecise (giving estimates only to the nearest 5 ng/µl), but robust to the presence of contaminants, and provides additional information about the size of the DNA. Second, more precise measurements were

made using PicoGreen fluorescence. The two sets of measurements (Table 1) were highly correlated (R^2 = 0.74; p value = 0.013), providing confidence in the PicoGreen numbers. In our experiments, a single buccal swab yielded 535 ± 232 (mean ± SD) ng of DNA. We evaluated the quality of this DNA by determining its ability to provide templates for PCR products of different sizes. All samples allowed the amplification of products of up to 1,200 bp (Figure 2, Table 1), but a 3,000 bp product was generated in only four out of six samples. We therefore conclude that the DNA is of good quality.

We next investigated the ratio of human:total DNA in the samples, assuming that lymphoblastoid cell line DNA was 100% human. As shown in Table 1, the human:total DNA ratios obtained differed among the samples, but were consistent across the quadruplicates performed for each sample, and in some cases the estimates were over 100%. Since it is impossible for a sample to contain more than 100% human DNA (i.e. BS1, BS3, BS4), it has to be concluded that the procedures used (PicoGreen measurement of DNA concentration + qPCR measurement of human component) together produce an error of at least 25%, or that the cell line DNA was not a reliable standard. This error is estimated using the number from the most extreme sample, BS3, which carries (33/133) = 25% of human DNA more than the maximum possible (100%). An explanation for this high error rate, which may also account for the high variability (average 106 ± 20) within the sample ratios, could be provided by the presence of chemical contaminants (e.g. ethanol, EDTA) following the DNA extractions, which might have affected the qPCR yield. A less efficient qPCR in the control sample would account for the high measurements. An alternative explanation could be a higher presence of initial DNA in the buccal samples (reflecting inaccuracy of PicoGreen measurements), but the correlation between PicoGreen measurements and band intensity on agarose gel electrophoresis seems to exclude this.

Finally, we repeated these analyses on WGA aliquots of the same samples. As expected, the WGA negative control (water) C- shows a DNA concentration > 0 (around one-half of the other samples, Table 2). This is explained by the self-amplification of the WGA primer mix, which can lead to the production of long DNA fragments (both ssDNA and dsDNA). In the other samples, starting from 28.43 ng, 23.3 ± 3.4 µg DNA was obtained. The DNA quality was slightly reduced after WGA: although amplification of ≤1,200 bp fragments was successful in all samples, amplification of the 3,000 bp fragment failed in all (Table 2) probably due to the reduced size of WGA DNA molecules.

With one (WG6) exception all the remaining WGA amplified samples contain much less human DNA (from ½ in WG5 to ¼ in WG1) in comparison with the WGA control DNA prepared from a lymphoblastoid cell line (C+). It is important to note that the WGA qPCR produced a distribution of values comparable to those obtained with the starting buccal swab extracts. Since the WGA samples should be less contaminated by chemicals which could have altered the performance of the qPCR, it must be concluded that the high error rate (25%) experienced with the buccal swabs samples was not caused by the nature of the samples and should, therefore, be taken into account also for the WGA measures.

Assuming the control (C+) sample to be 100% human DNA, the maximum human: total WGA DNA was 15% (Table 2). The human:total DNA ratio measured in the buccal swabs samples (despite 25% error) decreased considerably with the exception on WG6. However, if the WGA C+ percentage (±SD) of human DNA after qPCR is considered as “relative 100%”, the average relative percentage of human DNA in WGA samples is 42.0% (±24.0%). However, removing the outlier WG6 sample from the analysis brings the above values down to a more likely 32% (± 9.5%We find that, on average, >500 ng DNA can be extracted from a buccal swab, and that this can readily be amplified 800-fold to provide WGA DNA of good quality that permits PCR amplification of >1 kb single-copy fragments. However, we find that substantial contamination is present in the WGA buccal swab DNAs. We conclude that, when using such DNAs, approximately three times more should be supplied than when WGA DNA from other sources is used.

MATERIALS AND METHODS

Sample Collection and Ethics

Six buccal (inner cheek mucosa cells) anonymised swabs were collected from two consenting donors (namely the two authors of this paper), at least one hour after the consumption of food or drink, using Qiagen's Gentra Buccal Cell Kit. Cells were lysed and the DNA extracted from the samples immediately after collection following the kit manufacturer's instructions. The control genomic DNA (NA19240), prepared from a lymphoblastoid cell line, was obtained from the Coriell Institute (Camden NJ, USA).

Ethical approval for this research was not required since tissue which is being held whilst it is processed with the intention to extract DNA is not considered 'relevant material' under the Human Tissue Act providing it is held for a matter of hours or days (ref. Section 60, Human Tissue Authority (HTA) Code of Practice 9 – Research). Extracted DNA is not considered 'relevant material' (ref. Section 50, HTA Code of Practice 9 – Research, and Section 53 of the Human Tissue Act 2004). The Human Tissue Act 2004 does not apply to cell-lines (ref. HTA Supplementary List of Materials). Tissue obtained by non-invasive swabbing was used for the purposes of protocol optimisation only, and no genetic information or analysis of human material was undertaken (Section 45 of the Human Tissue Act does not apply). Written consent was considered unnecessary for this study and the Human Tissue Act 2004 does not specify the format in which consent should be given or recorded, except for anatomical examination or public display which must be in writing (ref. Section 56, HTA Code of Practice 1 – Consent). Samples were taken from willing, healthy and informed volunteers (namely the two authors of this paper), not NHS patients, so specific NHS Research Ethics Committee approval for the taking of samples from NHS patients was not required.

DNA Concentration

After extraction, the total dsDNA concentration was measured twice using PicoGreen [4], a direct method that measures the fluorescent emission after intercalation of a specific dye into dsDNA. The results shown in Table 1 were confirmed by gel electrophoresis (Figure 1).

DNA Quality

To assess the quality of the extracted DNA, a set of PCRs amplifying single-copy fragments from 152 bp to 3,000 bp (protocols available on request) were performed on the six DNA samples, using cell line DNA as a positive control (C+) and water blank as the negative control (C-).

Human DNA quantitation

A qPCR experiment was performed to determine the relative ratio of human:total DNA present in each buccal swab extract. The human-specific primers for the RNAse-P single copy gene and the universal fluorescent probe 12 were used for the analysis. During qPCR the samples were amplified together with cell line control DNA of predetermined concentration and the amount of starting human DNA template was estimated by measurement of the fluorescence emitted by the probe bound within the amplified region. Since the initial concentration of total DNA template

added to each reaction was the same, as estimated by PicoGreen measurement, a difference in the threshold cycles between the buccal swab sample and the control DNA was used to estimate the ratio of the human versus microbial contaminant DNA with the following formula:

$$\text{Ratio (Human/Total DNA)} = 2^{-(\text{sCycles} - \text{cCycles})}$$

With:

sCycles = number of cycles needed by the sample to reach the threshold fluorescence intensity
cCycles= number of cycles needed by the control to reach the threshold fluorescence intensity

Whole Genome Amplification

Whole-genome amplification with the GE GenomiPhi HY DNA Amplification Kit (cat. No. 25-6600-25) was performed on 28.43 ng of the six buccal swabs extracted DNA and the control samples. A higher amount of initial DNA was taken (rather than the 10 ng suggested by the protocol) to increase the total amount of initial human DNA. A 1:50 dilution of these WGA samples was used for PCR amplification as described earlier.[5].

Acknowledgements

The authors would like to thank Emma Gray and Katy Stirrups from The Wellcome Trust Sanger Institute for their kind help with PicoGreen and qPCR measurements respectively and Carol Smee from the same Institute for advices on Ethic consent.

References

1. Elizabeth Milne, Frank M. van Bockxmeer, Laila Robertson, Joanna M. Brisbane, Lesley J. Ashton et al. (2006) Buccal DNA Collection: Comparison of Buccal Swabs with FTA Cards. Cancer Epidemiol Biomarkers 15(5):1056.

2. David López Herráez , Mark Stoneking (2008) High fractions of exogenous DNA in human buccal samples reduce the quality of large-scale genotyping. Anal Biochem. 383(2):329-31.

3. Maloney B, Ray B, Hayden EP, Nurnberger JI Jr, Lahiri DK (2009) Development and validation of the high-quality 'rapid method for swab' to genotype the HTTLPR serotonin transporter (SLC6A4) promoter polymorphism. Psychiatr Genet. 19(2):72-82.

4. Enger O. (1996) Use of the fluorescent dye PicoGreen for quantification of PCR products after agarose gel electrophoresis. Biotechniques 21(3):372-4.

5. Frank B. Dean, John R. Nelson, Theresa L. Giesler, and Roger S. Lasken (2001)
Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. Genome Research 11:1095-1099

Financial Disclosure

The present study was funded by the Wellcome Trust Genome Research Limited (reg no. 2742969). The authors would like to thank the Cambridge European Trust, the UK Domestic Research Scholarship and Cambridge University Emmanuel College for providing funding for Luca Pagani's PhD project.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interest

The authors declare no competing interests.

Abbreviations

List of abbreviations included in the paper: DNA: Deoxyribonucleic Acid; WGA: Whole Genome Amplification; PCR: Polymerase Chain Reaction; qPCR: quantitative Polymerase Chain Reaction; SD: Standard Deviation

Figures

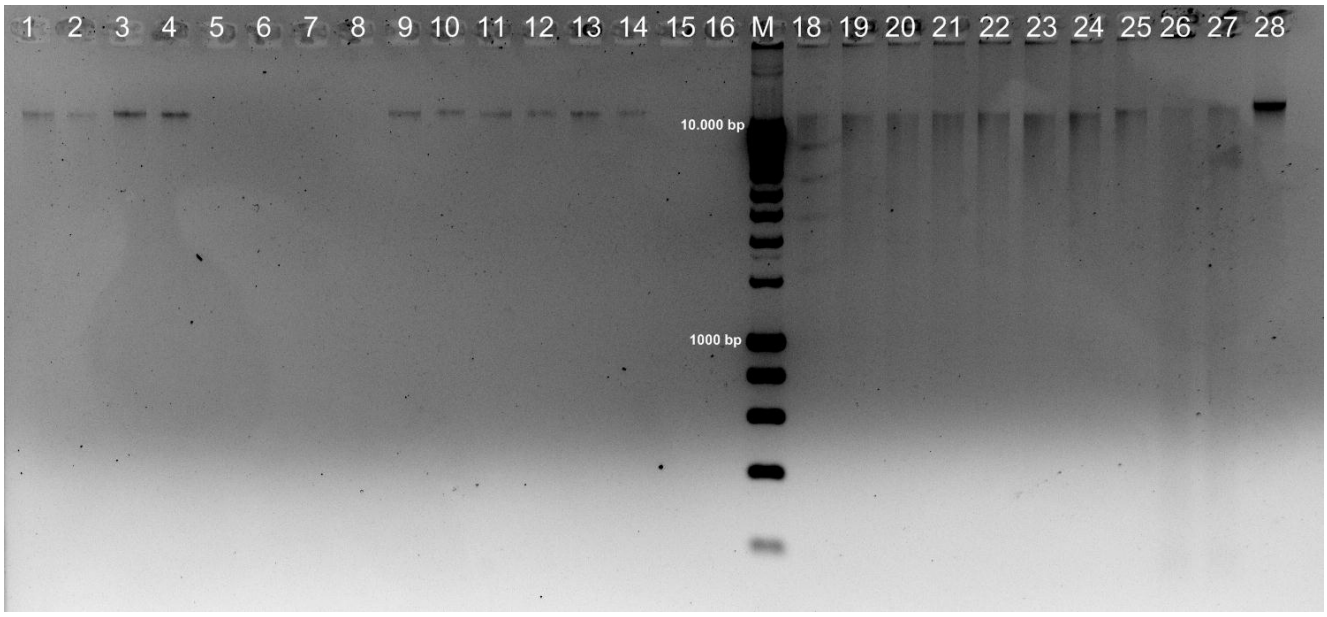

Figure 1 Gel Electrophoresis of genomic and WGA DNAs.

Lanes (Samples/Dilutions): 1. BS1/1:10; **2.** BS1/1:10; **3.** BS2/1:10; **4.** BS2/1:10; **5.** BS3/1:10; **6.** BS3/1:10; **7.** BS4/1:10; **8.** BS4/1:10; **9.** BS5/1:10; **10.** BS5/1:10; **11.** BS6/1:10; **12.** BS6/1:10; **13.** C+/1:10; **14.** C+/1:10; **15.** C-/1:10; **16.** C-/1:10; **M.** Ladder; **18.** WG1/1:50; **19.** WG2/1:50; **20.** WG3/1:50; **21.** WG4/1:50; **22.** WG5/1:50; **23.** WG6/1:50; **24.** WGC+/1:50; **25.** WGC-/1:50; **26.** BS3/1:1; **27.** BS4/1:1; **28.** C+/1:1.

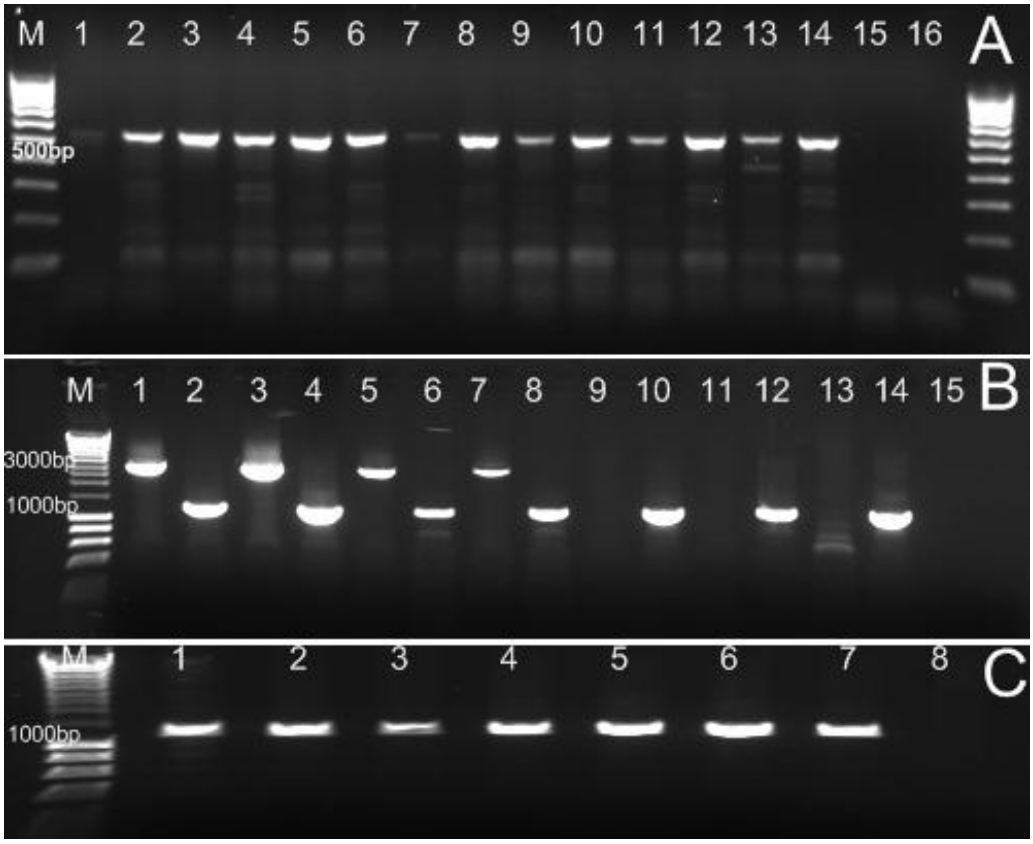

Figure 2 PCR amplifications.

A. 500bp PCR. M. Ladder; **1.** BS1; **2.** WG1; **3.** BS2; **4.** WG2; **5.** BS3; **6.** WG3; **7.** BS4; **8.** WG4; **9.** BS5; **10.** WG5; **11.** BS6; **12.** WG6; **13.** C+; **14.** WGC+; **15.** C-; **16.** WGC **B. 3000bp PCR. M.** Ladder; **1.** BS1; **3.** BS2; **5.** BS3; **7.** BS4; **9.** BS5; **11.** BS6; **13.** C+; **15.** C-. **1200bp PCR. 2.** BS1; **4.** BS2; **6.** BS3; **8.** BS4; **10.** BS5; **12.** BS6; **14.** C+; **16.** C-.**C. 1200bp PCR. M.** Ladder; **1.** WG1; **2.** WG2; **3.** WG3; **4.** WG4; **5.** WG5; **6.** WG6; 7**.** WGC+; **8.** WGC-.

Table 1 **Buccal Swabs Samples**

Sample	DNA concentration (ng/µl)		Total DNA (ng)	PCR length (bp)				qPCR %Human ± SD
	pGreen	EtBr		162	500	1200	3000	
BS1	35.6	20	712	**Y**	**Y**	**Y**	**Y**	130±29
BS2	44.1	40	881	**Y**	**Y**	**Y**	**Y**	89±8
BS3	11.4	15*	455	**Y**	**Y**	**Y**	**Y**	133±23
BS4	14.0	15*	562	**Y**	**Y**	**Y**	**Y**	102±45
BS5	16.8	20	335	**Y**	**Y**	**Y**	**N**	92±9
BS6	13.4	15	268	**Y**	**Y**	**Y**	**N**	91±8
C+	20.2	20	n/a	**Y**	**Y**	**Y**	**N**	n/a
C-	n/a	0	n/a	**N**	**N**	**N**	**N**	n/a

Table 2 **28.43ng of total DNA in 50 reaction volume WGA 1:50 diluted afterwards**

Sample	DNA concentration (ng/ul)		PCR length (bp)				%Human ± SD	
	pGreen	EtBr	162	500	1200	3000	qPCR	As %of WGC+
WG1	7.8	10*	**n/a**	**Y**	**Y**	**N**	3.7 ± 0.9	25.0
WG2	8.9	10*	**n/a**	**Y**	**Y**	**N**	5.8 ± 1.3	39.1
WG3	8.0	10*	**n/a**	**Y**	**Y**	**N**	3.6 ± 0.4	24.4
WG4	8.7	10*	**n/a**	**Y**	**Y**	**N**	4.4 ± 0.4	29.4
WG5	10.6	10*	**n/a**	**Y**	**Y**	**N**	6.8 ± 1.1	45.9
WG6	9.9	10*	**n/a**	**Y**	**Y**	**N**	13.0 ± 1.5	87.9
WGC+	11.5	10*	**n/a**	**Y**	**Y**	**N**	14.8 ± 0.6	100.0
WGC-	5.4	10*	**n/a**	**N**	**N**	**N**	0	0

* smeared DNA

Appendix 3 “Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations”

Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations

Emilia Huerta-Sánchez[1,2,*], Michael DeGiorgio[1,*], Luca Pagani[3,4,*], Ayele Tarekegn[5], Rosemary Ekong[6], Tiago Antao[3], Alexia Cardona[3], Hugh E. Montgomery[7], Gianpiero L. Cavalleri[8], Peter A. Robbins[9], Michael E. Weale[10], Neil Bradman[6], Endashaw Bekele[5], Toomas Kivisild[3], Chris Tyler-Smith[4], Rasmus Nielsen[1,2,11,12]

[1]Department of Integrative Biology, University of California, Berkeley, CA, USA
[2]Department of Statistics, University of California, Berkeley, CA, USA
[3]Division of Biological Anthropology, University of Cambridge, Cambridge, UK
[4]Wellcome Trust Sanger Institute, Hinxton, UK
[5]Addis Ababa University and Center of Human Genetic Diversity, P.O. Box 1176, Addis Ababa, Ethiopia
[6]The Centre for Genetic Anthropology, Department of Genetics, Evolution and Environment, University College London, London, UK
[7]Institute for Human Health and Performance, University College London, London, UK
[8]Molecular and Cellular Therapeutics, The Royal College of Surgeons in Ireland, Dublin, Ireland
[9]Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford, UK
[10]Department of Medical and Molecular Genetics, King's College London, London, UK
[11]Department of Biology, University of Copenhagen, Copenhagen, Denmark
[12]Beijing Genomics Institute, Shenzhen, China

*These authors contributed equally
Corresponding authors:
Emilia Huerta-Sanchez (emiliahs@stat.berkeley.edu)
Michael DeGiorgio (mdegiorgio@berkeley.edu)
3060 VLSB
Berkeley, CA 94720
Running title: Adaptation to high altitude in Ethiopians
Keywords: adaptation to high-altitude, natural selection

Summary

Background: The Tibetan and Andean Plateaus and Ethiopian highlands are the largest regions to have long-term high-altitude residents. Such populations are exposed to lower barometric pressures, and hence atmospheric partial pressures of oxygen. Such 'hypobaric hypoxia' may limit physical functional capacity, reproductive health and even survival. As such, selection of genetic variants advantageous to hypoxic adaptation is likely to have occurred. Identifying signatures of such selection is likely to help understanding of hypoxic adaptive processes. Here, we seek evidence of such positive selection using five Ethiopian populations, three of which are from high-altitude areas in Ethiopia. As these populations may have been recipients of Eurasian gene flow, we correct for this admixture.

Results: Using single nucleotide polymorphism genotype data from multiple populations, we find the strongest signal of selection in a gene *BHLHE41* (also known as *DEC2* or *SHARP1*). Remarkably, a major role of this gene is regulation of the same hypoxia response pathway on which selection has most strikingly been observed in both Tibetan and Andean populations. Because it is also an important player in the circadian rhythm pathway, *BHLHE41* might also provide insights into the mechanisms underlying the recognized impacts of hypoxia on the circadian clock.

Conclusion: These results support the view that Ethiopian, Andean, and Tibetan populations living at high altitude have adapted to hypoxia differently, with convergent evolution affecting different genes from the same pathway.

Highlights

- Selection scan reveals *BHLHE41* (also known as *SHARP1* or *DEC2*) as a candidate gene involved in high altitude adaptation in Ethiopians.
- *BHLHE41* is a highly relevant candidate gene with a regulatory role in the hypoxia response pathway, the same pathway affected in Tibetans and Andeans, providing an important example of convergent evolution in humans.

Introduction

Barometric pressure falls with ascent to altitude, and with it the partial pressure of inspired oxygen. The resulting reduction in systemic oxygen availability (hypobaric hypoxia) impairs physical performance, and can be detrimental to health and survival [1, 2]. It may also impair placental function and neonatal survival [3]. Over generations, such effects are likely to have exerted pressure for the selection of beneficial genetic variants in humans who have settled the three major high-altitude regions of the world: the Tibetan Plateau, the Andean Plateau, and the Ethiopian highlands. However, the genetic targets of selection may, in part, have differed between these populations: for example, hemoglobin concentration and oxygen saturation vary between them [4]. Indeed, hemoglobin concentrations are largely independent of altitude (up to 4000 meters [m]) in Tibetans, but rise with altitude in the high-altitude Andean population [4]. Tibetan individuals also have lower arterial oxygen saturations than Andeans at the same altitude [4]. Both traits are highly heritable [3, 5]. In contrast, amongst male Ethiopians of mostly Amharic ethnicity living at altitudes of 3500m, both hemoglobin concentration and arterial oxygen saturation remain almost the same as those of U.S. sea level males [6] and show small changes compared to lowland Amhara individuals [7]. These observations suggest that the genetic adaptation to high altitude in Ethiopians may differ from that in Tibetans and Andeans.

At the cellular level, activation of the HIF hypoxia-sensing pathway is the key

response to a reduced oxygen environment, primarily through the activity of the HIF-1α and HIF-2α transcription factors. In the case of Tibetans, two genes within the HIF hypoxia-sensing pathway exhibiting strong signatures of positive selection in response to high altitude have been identified (e.g., *EPAS1* and *EGLN1*; [8-14]). In Andeans, *EGLN1* is the only gene so far identified which also has strong signatures of positive selection in Tibetans [9]. In Tibetans, variants in these genes are associated with a minimal increase in hemoglobin concentration at high altitude.

Recently, Scheinfeldt et al. (2012) searched for signatures of positive selection across the genome in one high-altitude Ethiopian population—the Amhara. Using a number of statistical methods for detecting selection (including per-SNP F_{ST} between pairs of populations, locus specific branch length (LSBL), integrated haplotype score (iHS), cross population composite likelihood ratio (XP-CLR), and SNP-phenotype association), they proposed several genes as candidate targets of positive selection [15]. More recently, Alkorta-Aranburu et al. (2012) analyzed both the Amhara and the Oromo, two populations that inhabit high altitude regions in Ethiopia. In that study, they conducted many population comparisons designed to detect selection in high-altitude Amhara and Oromo and look for enrichment of hypoxia genes in their results [7]. They also employ genotype-phenotype associations to propose some candidate genes.

However, studying high altitude adaptation in Ethiopia is challenging for at least

two reasons. First, because of its geographic location it is likely that there has been gene flow from sub-Saharan Africa, northern Africa, and the Middle East into Ethiopia. Indeed, Ethiopian populations share a non-negligible proportion of their genetic material with other non-African populations [16]—perhaps as high as 40-50% [17]. Also, there has likely been substantial gene flow between low and high-altitude populations within Ethiopia and nearby regions. Second the Ethiopian highlands are in a lower altitude than the Andean and Tibetan plateaus.

Here we undertake an analysis of the Amhara, Tigray and Oromo genotype data published in Pagani et al. (2012) [17], which are individuals sampled at intermediate altitudes. Archeological studies in Ethiopia, show that humans have inhabited regions of more than 2000m for thousands of years [18, 19]. Both the Amhara and the Oromo populations have inhabited regions of more than 2500m for many generations. Alkora-Aranburu et al. (2012) [7] assume 5000 years as a reasonable estimate for the Amhara high-altitude settlement in Ethiopia. By contrast, the estimates of when the Oromo settled in regions of high altitude are far more recent at approximately 500 years [20, 21]. Genetic differentiation between the high- and low-altitude Amhara or between the high- and low-altitude Oromo is almost negligible as measured by principle component analysis (PCA) and F_{ST} analyses (Figure S3 and Text S2 in [7]) which suggests continual gene flow between the high- and low-altitude populations. Therefore, we expect the signal of positive selection to remain detectable in the populations considered

here, even if their current environment does not expose them to such extreme selection pressure.

Both the Amhara and Tigray populations share the same Semitic language group, cluster in PCA [17] and live at similar elevations. We therefore scanned both a separate and a pooled sample of Amhara-Tigray individuals for signals of positive selection, and compared our findings with those from previous studies on high-altitude adaptation in Ethiopians. For completeness, we carried out the same analysis with Oromo samples, both separately and when pooled with the Amhara-Tigray groups (see *Methods* for exact altitude of the sampled populations). The low-altitude populations employed in the study consist of the Afar and the Anuak from Pagani et al. (2012) [17]. Finally, we conducted a simulation study to assess the effects of admixture on our ability to detect true signatures of positive selection in admixed populations.

To detect selection, we used a statistical method that, for each gene, ranks signals based on a previously developed score, termed the population branch statistic (PBS; [11]). The PBS method has been proven effective in detecting selected loci amongst high-altitude Tibetan populations. It employs three populations, such that a population's PBS value corresponds to the magnitude of the allele frequency change at a given locus relative to the divergence from the other two populations (see *Methods*). In this study, we sought to identify signals of positive selection specifically in Ethiopians hypothesized to be high-altitude

adapted. Therefore, we applied the test on the Amhara, Tigray, and Oromo separately as well as the Amhara-Tigray or Amhara-Tigray-Oromo pooled data. Briefly, we computed the PBS for each gene region after correcting for non-African admixture (see *Methods* section). When allele frequencies were corrected for admixture from European or Middle Eastern populations, the top candidate is *BHLHE41*, a gene of functional relevance to both hypoxic-response and circadian rhythm pathways.

Results

Non-African admixture in Ethiopians

The Ethiopian populations considered in this study share a moderate degree of genetic similarity with some non-African populations. Figure 1 shows Ethiopian populations clustering between sub-Saharan Africans and non-Africans. Notably, African Americans also lie between sub-Saharan Africans and non-Africans, and they are known to have about 20% of European ancestry on average [22]. This clustering pattern is consistent with the results reported for the same samples, where it was shown that this pattern is probably due to European admixture (between 40 and 50%) into Ethiopians dated at 3000 years ago [17]. Though the current understanding of the demographic history for Ethiopians remains incomplete, such admixture could result in spurious signals of positive selection if the admixture proportion differs among Ethiopian populations. In fact, when the

selection scan was performed without correcting for admixture, the first and second most significant genes for the Tigray-Amhara scan were *MYEF2* and *SLC24A5* (Table S1), which both display strong signatures of positive selection in Europeans, and are suggested to relate to lighter skin pigmentation [23]. Therefore we propose that the higher European admixture proportion observed for these genes in the high-altitude compared to the low-altitude Ethiopians led to the potentially spurious signal of altitude adaptation. Indeed, if we calculate the average European admixture in the *MYEF2-SLC24A5* region (see *Methods* section), then we find it to be about 22% in the low-altitude Afar, but 48% and 44% in the high-altitude Amhara and Tigray, respectively. We cannot rule out, however, that the European alleles were indeed differentially selected in the high-altitude population due to an unknown selective pressure. However, we will in the following use an admixture-corrected version of the PBS statistic by Yi *et al.* (2010) [11] to detect selection (see *Methods* section for details on this method).

Simulation results: correcting for admixture leads to fewer false positives

To assess whether correcting for admixture leads to fewer false positives, we performed two types of simulations (see *Methods* for details). In the first scenario, we simulated two admixed populations: one with selection at a given locus and with a higher admixture proportion, representing the non-African admixed highland Ethiopians, and the other without selection and a lower admixture proportion, representing non-African admixture in the lowland

Ethiopians (see Figure S1*A*). In the second scenario, the non-African population itself experiences positive selection at a locus before the admixture event (e.g., *SLC24A5* locus in Europeans), but the admixed populations experience no selective event (see Figure S1*B*). This scenario evaluates the false positive rate for our PBS statistic with or without the admixture correction.

In Figure S2, we plot receiver operator characteristic (ROC) curves under the first simulation scenario (Figure S1*A*), and identify signals of selection with the PBS statistic. The null distribution of the PBS statistic is derived from the same demographic models in Figures S1*A* S1*B* without any selective event (*i.e.*, neutrality). The ROC curves reveal that correction for admixture improves the sensitivity (*i.e.*, the true positive rate as defined by the proportion of true selection signals correctly identified as selection signals by the PBS) and lowers the false positive rate for detecting selection in the admixed population. However, in practice, one is often only concerned with a method's performance at reasonably low false positive rates. For two of the parameters values displayed in Figure S2, Figure S3 focuses on more realistic false positive rates, varying from 0.0 to 0.05. It shows that within this range correcting for admixture affords approximately a 20% increase in sensitivity when compared to not correcting for admixture. In addition, Figure S4 shows that correcting for admixture correctly down weights the false signal of selection in the admixed population that arises from an adaptive event in the non-African group (*i.e.*, under the setting displayed in

Figure S1*B)*. It is worth noting, however, that even without correction, the PBS statistic is mostly robust to admixture at the level simulated here.

Selection scans

Selection scan in Amhara-Tigray after correcting for admixture

Figure 2*A* plots the PBS values for each gene, and groups them with PBS values from other genes containing identical numbers of SNPs. Table S2 lists the top 25 genes after correcting for admixture when the Amhara and the Tigray populations are combined.

At the top of the list is *BHLHE41* (also known as *DEC2, or SHARP1*), which is also an extreme outlier with respect to the empirical distribution of PBS values for genes with comparable numbers of SNPs (Figure S5*A*). This gene is a biologically plausible candidate for selection, being involved in hypoxic-response pathways, and having a physical interaction with *HIF-1α*. Figure 3 illustrates the known and predicted interactions for BHLHE41 from the STRING database [24], including several components of the hypoxia pathway. The HIF-1α/ARNT1 protein heterodimer plays a critical role in the hypoxia-induced transcription of vascular endothelial growth factor (*VEGF*) [25], and BHLHE41 negatively regulates *VEGF* expression by its interaction with HIF-1α/ARNT1 activation [26]. In addition, the promoter region of *BHLHE41* contains a hypoxia response element that is bound and transcriptionally regulated by HIF-1α, generating an apparent negative feedback loop [27]. More recently, experiments in breast

cancer cell lines have confirmed the direct interaction of BHLHE41 with HIF-1α, and demonstrated that BHLHE41 is a global inhibitor of HIF-1α and HIF-2α's transcriptional activity via down regulation of HIF-1α and HIF-2α protein expression [28]. BHLHE41 is proposed to facilitate the delivery of the HIF proteins to the proteasome [28]. *BHLHE41* also ranked highly when the analysis is performed using only the Tigray or only the Amhara population (Tables S3 and S4, respectively).

BHLHE41 is also a component of the circadian clock pathway [29,30], and a mutation in *BHLHE41* is associated with a short-sleep phenotype in humans [31]. Interestingly, the genes at the 18th and 19th position (respectively *DKFZp779M0652* and *SLC35C1*) in Table S2 are within 100kb of *CRY2*, a gene that is also a member of the circadian clock pathway. CRY2 may indirectly suppress *HIF-1α*/*ARNT1* activity through the transcriptional regulation of the circadian *PER1* gene, which is known to interact with *HIF-1α* via the PAS domain [32]. Extensive crosstalk between circadian clock and hypoxia pathways has been previously elucidated [33]. Another gene, *SMURF2* (*SMAD*-specific E3 ubiquitin-protein ligase 2), plays a role in the vascular inflammatory response in the presence of hypoxia in endothelial cells through an up-regulation of *TGF-β* signaling [34]. In addition, *CASP1* (caspase 1) is in the hypoxia response pathway, and has been implicated in the pathogenesis of many disorders including cardiovascular disease. Interestingly, the alcohol dehydrogenase genes *ADH6, ADH1A*, *ADH1B*, and *ADH1C* are differentiated between the low- (Afar

and Anuak) and the high-altitude populations (Tigray and Amhara). They have previously been observed to display strong signals of positive selection, concurrent with the introduction of agriculture and fermentation in human societies [35], and were also identified in the recent study of Amhara high-altitude populations [7].

The ranking in Table S2 discussed in the previous section is based on PBS statistics calculated from the aggregation of all the SNPs in the immediate region (*i.e.*, within 50kb upstream or downstream) of the gene. If we instead calculate PBS for each SNP separately, and then rank genes that are within 50kb of each SNP, we again retrieve *BHLHE41, SMURF2*, and *CASP1* (Figure 4 [Amhara-Tigray comparison], and Figure S6) as having a group of SNPs with PBS values above the 0.10% cutoff of the empirical distribution of all PBS values. These three genes thus rank highly when either 'per SNP' or 'per genic-region' analyses are performed.

Selection scan in the Oromo

Applying the same methodology to the Oromo population, and employing the same low-altitude (the Afar) and outgroup (the Anuak) populations as controls, *BHLHE41* again emerges as the most significant locus, with or without admixture correction (see Figure 2*B* and Supplementary Table S5). In the Oromo, *BHLHE41* is also an extreme outlier with respect to the empirical distribution of PBS values for genes with comparable numbers of SNPs (Figure S5*B*). If we

instead calculate PBS for each SNP separately, and rank genes that are within 50kb of each SNP, then we still retrieve *BHLHE41* (see Figure 4 [Oromo comparison], Figure S7). Interestingly, unlike the Amhara and Tigray, neither the alcohol genes nor the pigmentation genes show strong differentiation between the Oromo and the low-altitude Afar. If we pool the Oromo with the Amhara and the Tigray, *BHLHE41* remains the top candidate (Figure 4 [Amhara-Tigray-Oromo comparison], Figures S5*C* S8 and Table S6).

The identification of the *BHLHE41* gene in the Oromo population supports the hypothesis that it arose by a single early selective event affecting an ancestral population, rather than by two independent selective events. The genetic differentiation between the Oromo and the Amhara (F_{ST} = 0.01 [7]; F_{ST} = 0.02, [17]) is sufficiently small to support a scenario in which selection occurred in an ancestral Amhara-Tigray-Oromo population. Alternatively, it is possible that we observe selection on *BHLHE41* in the Oromo due to recent gene flow followed by selection for the variant, causing its frequency to increase. Unlike the Amhara, records point to a recent (~500 years) settlement of the Ethiopian highlands by the Oromo population [20, 21]. This recent estimate would support the scenario of selection aided by gene flow. In fact, in the genic region of *BHLHE41*, the F_{ST} between the Oromo and the Amhara and between the Oromo and the Tigray is smaller than between the Oromo and the Afar despite the latter pair belonging to the same language group (Cushitic). Furthermore, the F_{ST} between Amhara and Oromo in that gene region (F_{ST} = 0.01) is smaller than the median across all

gene regions (median F_{ST} = 0.03), and the same observation holds true between the Oromo and the Tigray populations. These F_{ST} values suggest that selection has acted to reduce genetic differentiation between high-altitude adapted populations in this locus. If true, then this is the first example in humans of natural selection acting to reduce F_{ST} between populations in a genomic region. Populations of Ethiopia, however, have a complex demographic history and more studies are needed to reconcile these observations.

Comparison with other studies

Scheinfeldt et al. (2012), performed a scan for positive selection in the Amhara, one of the high-altitude Ethiopian populations analyzed here, albeit sampled from a different location in Ethiopia [15]. Overall, they found no enrichment for HIF pathway genes, but did propose a number of candidate genes: *VAV3*, *CBARA1*, *THRB*, *ARNT2*, *PIK3CB*, *ARHGAP15*, and *RNF216*. All of these genes have SNPs with LSBL values in the top 0.10% of our analysis. If we intersect their full gene list from the top 0.10% of SNPs (listed in Table S2 in Scheinfeldt et al. (2012) [15]) with our hypoxia-related genes (see *Methods* section for definition of this hypoxia set), we obtain *DDIT4*, *NARFL*, *RYR2*, *RYR1*, *ARNT2*, and *GATA6*. For an equivalent analysis, we examined our Amhara sample without correcting for admixture, and retrieved the SNPs with PBS values in the top 0.10%, revealing a collection of hypoxia-related genes: *PPARA*, *ANGPT2*, *RYR2*, *SFRP1*, *ITPR2*, *TP53*, *CHRNB2*, *FLT1*, and *PYGM*. From our top-ranking list, only *RYR2* overlapped with their LSBL list. They do, however, identify *PPARA*

from the XP-CLR test of selection, and if we extend the interval to include genes within 100kb of the top 0.10% of SNPs, then *VAV3* and *THRB* also appear in common. In our dataset, we could not identify any SNPs in or near *CBARA1*, *ARNT2*, or *PIK3CB* with PBS values in the top 0.10%.

The more recent Alkorta-Aranburu et al. (2012) study made multiple Ethiopian high- and low-altitude population comparisons (Table S22 in [7]) and they included results with the same PBS statistic we apply here [7]. They concluded that with the PBS metric the most significant enrichment of hypoxia-related genes emerged when comparing a mixed high- and low-altitude Amhara population to the Masai (Luhya as outgroup). Their results are marginally significant when comparing high-altitude Amhara to low-altitude Amhara, which suggests, analogous to what we find between Amhara and Tigray, that the population groups are so closely related that they both still harbor the selected loci at similar frequencies. However, their candidate gene list with extreme PBS values does not include *BHLHE41*, and none of their PBS candidates are significantly associated with hemoglobin concentration or oxygen saturation phenotypes. Interestingly one SNP within a hypoxia-related gene, *RORA*, does associate with hemoglobin concentration in the Amhara despite lacking a signature of positive selection.

Expected number of hypoxia-related genes

We performed a permutation test (see *Methods* section) to assess the number of hypoxia-related genes that would be expected by chance to be found in a random selection of the same number of genes that we observe in the top 0.10% of SNPs, and found that the expectation varies from six to 12 genes depending on the high-altitude population considered (Table S7). Our analysis identified between seven and 16 genes in the hypoxia-related set, which is not a statistically significant enrichment. For a comparison of gene lists that include our admixture correction and all the high-altitude populations, we calculated a PBS score for each SNP in 14 cases: before and after correcting for admixture, for the Amhara, Tigray and Oromo separately as well as to their pooled data. We then identified genes within 50kb of the top SNPs. Table S8 displays the results of intersecting the genes within the top-ranked 0.10% of SNPs with the hypoxia-related gene set, with the exception of *BHLHE41* (which was not included in the gene set definition, but retrospectively appears to be highly relevant for hypoxic response). Notably, *BHLHE41* and *RYR2* are the only genes that appear under all the 14 scenarios considered.

Discussion

In this study, we have compared Ethiopian populations to identify genes that are likely involved in the Ethiopians' adaptation to high altitude. While the current sampled populations live at intermediate elevations, it is likely that they are descendants of high-altitude adapted populations that lived at elevations greater

than 2500m, and therefore we expect the signals of selection to remain detectable in the modern populations. In addition, the current altitude of residence (approximately 1800m) is not free of selective pressure. Standard barometric pressure at 1800m is 604 mmHg at 2000m, a reduction of nearly 20% when compared to sea level, which has clear biological effects. Indeed, low birth-weight is three times more common at elevations greater than 2000m even in the United States, with a threshold effect of 1500m [36].

One challenging aspect of our analysis is that Ethiopians have a complex demographic history, involving, among other events, admixture with non-Africans [17]. If admixture is not corrected for, then genes such as *SLC24A5*, which are involved in lighter skin pigmentation in Europeans, show strong signals of positive selection in the high-altitude populations. Accounting for admixture in the pooled Amhara-Tigray samples results in a decrease in the strength of these potentially spurious signals of high-altitude adaptation, and, instead, yields the strongest signal from the *BHLHE41* gene. Furthermore, in the Oromo, the strongest signal is also in the *BHLHE41* gene. This is a functionally relevant candidate gene with a major regulatory role in the same hypoxia-sensing pathway that has undergone selection in Tibetan and Andean populations. It is transcriptionally regulated by HIF-1α, binds HIF-1α, and represses many of the hypoxia-induced transcriptional targets including *VEGF,* likely due to the increased degradation of HIF-1α and HIF-2α proteins by BHLHE41 [26-28]. In

addition, it is a component of the circadian clock pathway [29, 30], and a mutation in *BHLHE41* is associated with a short-sleep phenotype in humans [31].

A role in hypoxic responses offers a clear target for selection in response to altitude, but a role in the regulation of circadian cycles is perhaps less clear. However, extensive circadian-hypoxia pathway crosstalk occurs [33]. Indeed, hypoxia-mediated changes in circadian rhythms have been suggested to be a key driver of the sleep fragmentation and poor sleep quality seen in lowlanders at high altitude [37]. In agreement, sleep quality is better in the native high-altitude populations of Tibet and the Andes [38-41].

The genes highlighted in our analyses were not detected, however, in previous studies of Ethiopian highland populations [7, 15], and a number of differences between our analyses and theirs could account for this. The first relates to the choice of low-altitude reference group. We compared the high-altitude Amhara, Tigray and Oromo populations against the Afar, a low-altitude population with a similar genetic and linguistic (Afro-Asiatic language group) background. Scheinfeldt et al. (2012) used the low-altitude Omotic group, which is less closely related to the Amhara, as shown using comparable samples in Pagani et al. (2012). By contrast, Alkorta-Aranburu et al. (2012) used multiple low-altitude groups (Table S22 in Alkorta-Aranburu et al. (2012)). In one of their PBS three-population combinations, they compare the low-altitude Amhara and high-altitude Amhara and find marginal enrichment for hypoxia-related genes, but it may be

that continued gene-flow between the two groups have kept the putative selected mutations at similar frequencies. Accordingly, when they pooled the low- and high-altitude Amhara, they find stronger enrichment in hypoxia-related genes. The second difference is that we chose the Anuak as our outgroup; in Scheinfeldt et al. (2012), their outgroup population was the more diverged Yoruban population, whereas Alkorta-Aranburu et al. (2012) employed various outgroup populations not including the Anuak. The third difference is that we found it important to correct for population admixture (Figures S2-S4). If SNP frequencies are left uncorrected, then much of the selection signal could derive from differences in admixture proportions.

We agree with the conclusions of Scheinfeldt et al. (2012) and Alkorta-Aranburu et al. (2012) that high-altitude adaptation can take place by distinct genetic alterations, as there is no overlap with the candidate genes from this study and those of previous Tibetan and Andean studies [7, 15]. However, Tibetan and Andean environments are considerably more extreme at elevations greater than 4000m, and phenotypic differences exist in hemoglobin concentrations and oxygen saturation. Thus, the underlying genetic differences may reflect different biological adaptation mechanisms. Furthermore, the high-altitude Ethiopian populations are the least isolated of the three global high-altitude populations, increasing the difficulty of uncovering signatures of adaptation. Nevertheless, at the pathway level, we demonstrated broadly shared biological processes targeted by selection in each of the adapted high-altitude populations, indicative

of convergent evolution. The top gene revealed by our analyses, *BHLHE41*, is an excellent candidate for further studies as it has an important function in the hypoxia response pathway. Given its role in the circadian clock, it also provides justification to explore the relationship between hypoxic conditions and the circadian cycles in future studies.

Methods and Materials

Data

We analyzed genetic data of individuals from Ethiopia available from Pagani et al. (2012) [17]; namely, three high-altitude populations (26 Amhara, 21 Tigray, and 21 Oromo individuals) and two low-altitude populations (12 Afar and 23 Anuak individuals). The Amhara and Tigray are members of the Semitic and the Afar and the Oromo of the Cushitic linguistic groups, both belonging to the Afro-Asiatic linguistic family. The Anuak are members of the Nilotic language group, a member of the Nilo-Saharan linguistic family. The Amhara, Tigray, Oromo, Afar, and Anuak samples were collected at altitudes of 1829m, 1695m, 1758m, 400m, and 500m, respectively. Though the samples were not collected at extremely high altitudes, the Amhara and Oromo populations have been residing in regions of Ethiopia higher than 2500m for many generations [7, 18-21]. The Tigray individuals were chosen so that they had both parents and all grandparents living at greater than 2000m. Using these five populations, we performed five selection scans without correcting for admixture and another five selection scans after correcting for admixture. In one scan we combined the two high-altitude

populations (Amhara and Tigray) and in another scan we combined the three high altitude populations (Amhara, Tigray and Oromo). In the other three scans we considered the Amhara, Tigray and Oromo separately. All consent information can be found in Pagani et al. (2012) [17].

Multidimensional scaling

Between each pair of individuals i and j in our dataset, we computed the allele sharing distance. For a particular site *k* in the genome, the pair of individuals i and j have a distance, denoted as $d_{i,j}^{k}$, of 0.0 if they both have the same genotypes, they have a distance of 0.5 if one has a homozygous and the other has a heterozygous genotype, and they have a distance of 1.0 if they are both homozygous but for different alleles. Assume that there are L sites in the genome for which neither individual i nor j is missing any genotype, and that these sites are indexed as $k = 1, 2, \ldots, L$. Then we compute the allele sharing distance between individuals *i* and *j* as

$$d_{i,j} = \frac{1}{L}\sum_{k=1}^{L} d_{i,j}^{k},$$

such that $0 \leq d_{i,j} \leq 1$ for $i \neq j$ and that $d_{i,j} = 0$ for $i = j$. We then construct a matrix of allele sharing distances between all pairs of individuals and apply classical multidimensional scaling to obtain components displayed in Figure 1.

Admixture analyses

We compared unrelated individuals from the Ethiopian (Pagani et al., 2012) populations to unrelated individuals from the HapMap phase 3 populations [42, 43] as well as to unrelated individuals from the HGDP populations [44, 45]. From these comparisons, and from previous results in Pagani et al., (2012) [7], we observe that the individuals in our Ethiopian dataset are probably admixed (Figure 1), and it was therefore necessary to control for admixture within our analyses because admixture can mimic signals of positive selection. To correct for admixture, we employ the European population as a proxy to represent the non-African population that contributed genetic material to the Ethiopian populations. For each population (*i.e.*, Amhara, Tigray, Oromo, Afar, Anuak, and European), we calculated allele frequencies at each locus. To control for admixture, we followed Bhatia et al. (2012) [46]. We assumed that the low- and high-altitude Ethiopian populations were a mixture between the Anuak (the outgroup population) and the European population (the nine unrelated CEU individuals from the Complete Genomics dataset; Drmanac et al., 2009 [47]). This assumption is reasonable given that the Afar, the Amhara, and the Tigray cluster between Europeans and the Anuak (see Figure 1). Under this assumption, at a given locus *k,* we can calculate the pseudo un-admixed allele frequency for each population by

$$p^{unadmixed} = \frac{p^{observed} - \alpha p^{European}}{1 - \alpha},$$

where α is the proportion of European admixture (Bhatia et al., 2012). We employed the value of α that minimizes F_{ST}^{k} between the Anuak and the

corresponding population (e.g., Afar, Tigray, Amhara, Oromo, or the combined Amhara-Tigray or Amhara-Tigray-Oromo populations). To compute F_{ST}^{k} we used the formula derived in Reynolds et al. (1983) [48].

To calculate the mean European admixture in the *MYEF2-SLC24A5* region for the Afar, Amhara and Tigray, we averaged the values of α across all SNPs within the region ranging from 50kb upstream of *SLC24A5* to 50kb downstream of *MYEF2*.

We did not find the same *MYEF2-SLC24A5* region to be under selection in the Oromo population. However, results from Pagani et al. (2012) show that the Oromo also have received some non-African admixture, and thus we also applied the correction to the Oromo population.

Population branch statistic

To detect regions under selection we calculated the population branch statistic (PBS; [11]). This test of selection takes three populations that have an evolutionary relationship illustrated in Figure S9. We assume that the Anuak are an outgroup population to the Afar and the high-altitude (HA) populations—the Amhara, the Tigray, and the Oromo. Under a scenario of only genetic drift, we expect the HA populations and the Afar to be more genetically similar than the HA-Anuak or the Afar-Anuak. If however, there has been local adaptation in the HA populations, then the regions targeted by positive selection would be highly

diverged between the Afar and the HA, and the Afar-Anuak would be more similar than the HA-Anuak comparison (Figure S9). Therefore, we should only detect genes that have been targeted by selection in the HA populations. Because we know that one of the selective pressures is lower oxygen level, we expect to observe genes involved in the response to hypoxia.

We obtained RefSeq gene annotations from http://genome.ucsc.edu/ and we used the longest RefSeq identifier for the analysis. For a given gene, F_{ST} between a pair of populations, which is a measure of genetic differentiation between a pair of populations, was calculated using the formula from Reynolds et al. (1983) [48] across all SNPs within the genic region, such that each SNP is located within 50kb of the transcription start and end of the gene. For each gene (+/- 50kb), we computed F_{ST} between population pairs Amhara-Afar, Amhara-Anuak, Tigray-Afar, Tigray-Anuak, (Amhara-Tigray)-Afar, (Amhara-Tigray)-Anuak, Oromo-Afar, Oromo-Anuak, (Amhara-Tigray-Oromo)-Afar, (Amhara-Tigray-Oromo)-Anuak, and Afar-Anuak. Using Anuak as our outgroup, Afar as our lowland population, and either Amhara, Tigray, Oromo, Amhara-Tigray, or Amhara-Tigray-Oromo as our highland population. We calculated the population branch statistic (PBS) of the high altitude population using the following formula:

$$PBS_{HA} = \frac{T^{HA,LA} + T^{HA,Outgroup} - T^{LA,Outgroup}}{2},$$

where $T^{HA,LA} = -\log(1 - F_{ST}^{HA,LA})$ is an estimate of the divergence time between the high altitude (HA) and low altitude (LA) populations. Similarly, $T^{HA,Outgroup}$ is an

estimate of the divergence time between the HA and outgroup populations and $T^{LA,Outgroup}$ is an estimate of the divergence time between the LA and outgroup populations. We calculated PBS [11] for each (highland, lowland, outgroup) triple at each gene. Additionally, for each SNP used in the F_{ST} calculations, we required that at least 10 alleles (*i.e*., five individuals) were observed in each population within a (highland, lowland, outgroup) triple. We computed PBS before and after correction of admixture, and the results are shown in Table S1 and Table S2 respectively. These two tables correspond to the case in which the two high-altitude populations (the Amhara and the Tigray) were combined. In both cases, the Anuak population was used as the outgroup population. We also performed the analysis requiring at least 20 alleles (10 individuals), and *BHLHE41* dropped from first place to second place. In the Oromo population, *BHLHE41* remained at the top of the list before and after correcting for admixture (Table S5 shows results after correcting for admixture).

Simulations with selection and admixture

Based on our observed results and those from Pagani et al. (2012), it is likely that the low- and high-altitude Ethiopian populations are admixed with non-African populations. Therefore, we wanted to investigate the effect that admixture has on our ability to identify selection signals, as it may mimic the effect of positive selection in the high-altitude population. Moreover, it is possible that the non-African population has itself experienced recent selection; therefore, it was also important to understand the influence of admixture in this case. We

performed simulations to determine the impact of admixture on the PBS statistic, to assess whether correcting for admixture improves inferences of positive selection, and to quantify how often we identify selection when it is in fact admixture from a population that has recently experienced positive selection. We considered two selection scenarios on the same demographic model. First, we considered the case in which the high-altitude population is targeted by positive selection and has a genetic contribution from a non-African population that has experienced a bottleneck in its demographic history. Second, we considered the case in which the high-altitude population receives a genetic contribution from a non-African population that has experienced both a bottleneck and positive selection. For an illustration of the models see Figure S2. We employed the software SFS_CODE [49] to simulate under the two selection scenarios. We simulated a region of 10kb with per-generation per-site mutation and recombination rates of 10^{-3}. We investigated a population-scaled selection coefficient *S* of 150 and 250, admixture proportions into the low- (α_1) and high-altitude (α_2) populations of (α_1, α_2) = (0.1, 0.2), (0.1, 0.3), (0.1, 0.4), and (0.2, 0.4), a time at which admixture occurs T_{ADM} of 1.5 and 3 thousand years ago (kya), with all other demographic parameters remaining fixed (see Figure S1). For each simulation we sampled 25 individuals from each population. The motivation for these recent admixture times stem from Pagani et al. (2012) [17], who found that admixture into these populations was recent (approximately 2.5 to 3kya). In addition, under the scenario in which selection occurs in the high-altitude population (Figure S1*A*), we used a time of selection T_{SEL} of 1.5kya and

3kya, and under the scenario in which selection occurs in the non-African population (Figure S1*B*), we used a time of selection of 5kya. These time estimates are comparable to those estimated in Pagani et al. (2012) [17].

We obtained 10^3 simulated datasets under each parameter combination. For the scenarios in which selection occurs in the high-altitude population, we calculated the proportion of true positives under a specified false positive rate that was based on the corresponding neutral scenario (see Figure S2). For two parameter combinations, we ran 10^4 simulations, and plotted the true positive rate as a function of false positive rate in the range from 0.0 to 0.05 (see Figure S3). Figure S3*A* corresponds to T_{ADM} = 1.5kya, S = 150, T_{SEL} = 1.5kya, α_1 = 0.1, and α_2 = 0.4. Figure S3*B* corresponds to T_{ADM} = 1.5kya, S = 150, T_{SEL} = 1.5kya, α_1 = 0.2, and α_2 = 0.4. For the scenario with selection in the non-African population, we calculated the proportion of times (out of 10^3) that a simulation is falsely called a positive for a specified false positive rate (see Figure S4).

Expected number of hypoxia genes

Due to multiple testing, the probability of finding a SNP in or near a hypoxia-related gene increases as the number of tests increase. We restricted our set of genes to those that are within 50kb of any SNP to estimate the total number of all possible genes that could be captured with the given data. Then, focusing on the top 0.10% of SNPs in our study, we identified the number of genes that were within 50kb of these SNPs. To approximate the number of hits for hypoxia genes

that we should expect by chance, we sampled the same number of unique genes that we identified using the top 0.10% of SNPs from the set of all possible genes. The reason for sampling the same number of unique genes that we identified using the top 0.10% of SNPs is because we wanted to maintain the SNP structure found in the real data. Finally, we counted the number of genes in the intersection of that random set and the hypoxia-related gene set (see *Definitions of hypoxia gene set* section). We repeated this experiment 10^3 times to derive an empirical null distribution and compute the expected number of hits by chance. Table S7 contains the median number of hypoxia-related genes we would expect to see based on this empirical null distribution under each of the different population scenarios.

Definitions of hypoxia gene set

The AmiGO tool (http://amigo.geneontology.org) was accessed and used to list all genes within the Gene Ontology biological process term "response to hypoxia" plus all descendent terms (GO:0001666 "response to hypoxia", GO:0071456 "cellular response to hypoxia" and GO:0070483 "detection of hypoxia"). This resulted in a set of 152 unique human genes (see Table S9).

Acknowledgments

This work was supported by research grants from the U.S. NSF to E.H.S. (DBI-0906065) and to M.D. (DBI-1103639) and the U.S. NIH (R01HG003229) to R.N.

References

1. Leon-Velarde, F., Maggiorini, M., Reeves, J.T., Aldashev, A., Asmus, I., Bernardi, L., Ge, Ri.-Li., Hackett, P., Kobayashi, T., Moore, L.G., et al. (2005).

Consensus statement on chronic subacute high altitude diseases. *High Alt. Med. Biol*. **6**, 147-157.

2. Monge, C.C., Arregui A., and Leon-Velarde, F. (1992). Pathophysiology and epidemiology of chronic mountain sickness. *Int J Sports Med*. 13, S79-S81.

3. Niermeyer, S., Andrade Mollinedo, P., and Huicho L. (2009). Child health and living at high altitude. Arch Dis Child. 94, 806-811.

4. Beall, M.C. (2006). Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. *Integr. Comp. Biol.* 46, 8-24.

5. Moore, L.G., Young, D., McCullough, R.E., Droma, T., and Zamudio, S. (2001). Tibetan protection from intrauterine growth restriction (IUGR) and reproductive loss at high altitude. *Am J Hum Biol.* 13, 635-44.

6. Beall, C.M., Decker, M.J, Brittenham G.M., Kushner, I., Gebremedhin, A., and Strohl, K.P. (2002). An Ethiopian pattern of human adaptation to high-altitude hypoxia PNAS 99, 17215-17218.

7. Alkorta-Aranburu, G., Beall, C.M., Witonsky, D.B., Gebremedhin, A., Pritchard, J.K., and Di Rienzo, A. (2012). The Genetic Architecture of Adaptations to High Altitude in Ethiopia. *PLoS Genet*. **8:**e1003110.

8. Beall, C.M., Cavalleri, L.G., Deng, L., Elston, R.C., Gao, Y., Knight, J., Li, C., Li, J.C., Liang, Y., McCormack, M., et al. (2010). Natural selection on *EPAS1* (*HIF2α*) associated with low hemoglobin concentration in Tibetan highlanders. *Proc. Natl. Acad. Sci. USA.* **107,** 11459-11464.

9. Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R., Scherer, S.W., Julian, C.G., Wilson, M.J., Lopez-Herraez, D., et al. (2010). Identifying signature of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet*. **6:**e1001116.

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

11. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329,** 75-78.

12. Peng, Y., Yang, Z., Zhang, H., Cui, C., Qi, X., Luo, X., Tao, X., Wu, T., Ouzhuluobu, Basang, et al. (2011). Genetic Variations in Tibetan Populations and High-Altitude Adaptation at the Himalayas. *Mol Biol Evol.* **28**, 1075-1081.

13. Wang, B., Zhang, Y.B., Zhang, F., Lin, H., Wang, X., Wan, N., Ye, Z., Weng, H., Zhang, L., Li, X., et al. (2011). On the Origin of Tibetans and Their Genetic Basis in Adapting High-Altitude Environments. *PLoS ONE* **6**:e17002.

14. Xu, S., Li, S., Yang, Y., Tan, J., Lou, H., Jin, W., Yang, L., Pan, X., Wang, J., Shen, Y., et al. (2011). A Genome-Wide Search for Signals of High-Altitude Adaptation in Tibetans. *Mol. Biol. Evol.* **28**, 1003-1011.

15. Scheinfeldt, L.B., Soi, S., Thompson, S., Ranciaro, A., Wolderneskel, D., Beggs, W., Lambert, C., Jarvis, J.P., Abate, D., Belay, G., et al. (2012). Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biol*. **13,** R1 doi:10.1186/gb-2012-13-1-r1.

16. Semino, O., Santachiara-Benerecetti A.S., Falaschi, F., Cavalli-Sforza, L.L, and Underhill, P.A. (2002). Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am. J. Hum. Genet*. **70,** 265-268.

17. Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Romero Gallego, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., et al. (2012). Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool*. The Am. J. Hum. Genet*. **91**, 83-96.

18. Pleurdeau, D. (2006). Human Technical Behavior in the African Middle Stone Age: The Lithic Assemblange of Porc-Epic Cave (Dire Dawa, Ethiopia). Afr Archaeol Rev 177–197. doi: 10.1007/s10437-006-9000-7.

19. Brandt, S. A. (1986). The upper pleistocene and early holocene prehistory of the horn of Africa. The African Archaeological Review 4, 41–82.

20. Hassen, M. (1990). The Oromo of Ethiopia: a history, 1570–1860. Great Britain: Cambridge University Press.

21. Lewis, H.S. (1966). The Origins of the Galla and Somali. The Journal of African History 7, 27–46. doi: 10.1017/S0021853700006058.

22. Tang, H., Jogersen, E., Gadde, M., Kardia, S.L., Rao, D.C., Zhu, X., Schork, N.J., Hanis, C.L., and Risch, N. (2006). Racial admixture and its impact on BMI and blood pressure in African and Mexican Americans. *Human Genet.* **119**, 624-633.

23. Lamason, R.L., Mohideen, M.-A.P.K., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynec, M.J., Mao, X., Humphreville, V.R., Humbert, J.E., et al. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310,** 1782-1786.

24. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., et al. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*. **37**, D412-6.

25. Forsythe, J.A., Jiang, B.H., Iyer, N.V., Agani, F., Leung, S.W., Koos, R.D., Semenza, G.L. (1996). Activation of vascular endothelial growth factor gene transcription by hypoxia-inducible factor 1. *Mol. Cell. Biol.* **16,** 4604-4613.

26. Sato, F., Bhawal, U.K., Kawamoto, T., Fujimoto, K., Imaizumi, T., Imanaka, T., Kondo, J., Koyanagi, S., Noshiro, M., Yoshida, H., et al. (2008). Basic-helix-loop-helix (bHLH) transcription factor DEC2 negatively regulates vascular endothelial growth factor expression. *Genes Cells*. **13,** 131-144.

27. Miyazaki, K., Kawamoto, T., Tanimoto, K., Nishiyama, M., Honda, H., Kato, Y. (2002). Identification of functional hypoxia response elements in the promoter region of the DEC1 and DEC2 genes. *J. Biol. Chem*. **277,** 47014-47021.

28. Montagner, M., Enzo, E., Forcato, M., Zanconato, F., Parenti, A., Rampazzo, E., Basso, G., Leo, G., Rosato, A., Bicciato, S., et al. (2012). SHARP suppresses breast cancer metastasis by promoting degradation of hypoxia-inducible factors. *Nature* **487**, 380-384.

29. Honma, S., Kawamoto, T., Takag, Y., Fujimoto, K., Sato, F., Noshiro, M., Kato, Y., Honma, K. (2002). *Dec1* and *Dec2* are regulators of the mammalian molecular clock. *Nature* **419,** 841-844.

30. Kato, Y., Noshiro, M., Fujimoto, K., Kawamoto, T. (2010). Roles of Dec1 and Dec2 in the core loop of the circadian clock, and clock outputs metabolism. *Metabolism Clinical and Experimental* **61**, s34-s42.

31. He, Y., Jones, C.R., Fijiki, N., Xu, Y., Guo, B., Holder, J.L. Jr., Rossner, M.J., Nishino, S., Fu, Y.H. (2009). The transcriptional repressor DEC2 regulates sleep length in mammals. *Science* **325,** 866-870.

32. Koyanagi, S., Kuramoto, Y., Nakagawa, H., Aramaki, H., Ohdo, S., Soeda, S., Shimeno, H. (2003). A molecular mechanism regulating circadian expression of vascular endothelial growth factor in tumor cells. *Cancer Res.* **63,** 7277-7283.

33. Chilov, D., Hofer, T., Bauer, C., Wnger, R.H., and Gassmann M. (2001). Hypoxia affects expression of circadian genes PER1 and CLOCK in mouse brain. *FASEB J.* **15,** 2613-2622.

34. Akman, H.O., Zhang, H., Siddiqui, M.A.Q., Solomon, W., Smith, E.L.P., Batuman, O.A. (2001). Response to hypoxia involves transforming growth factor-β2 and Smad proteins in human endothelial cells. *Blood.* **98,** 3324-3331.

35. Peng, Y., Shi, H., Qi, X., Xia, C., Zhong, H., Ma, R.Z., Su, B. (2008). The ADH1B Arg47His polymorphism in East Asian populations and expansion of rice domestication in history. *BMC Evol. Biol*. **20**, 1-15.

36. Yip, R. (1987). Altitude and birth weight. The Journal of Pediatrics, Volume 111, Issue 6, Part 1, 869–876.

37. Mortola, J.P. (2007). Hypoxia and circadian patterns. *Respiratory Physiology & Neurobiology* **158**, 274-279.

38. Coote, J.H., Stone, B., and Tsang, G. (1992). Sleep of Andean high altitude natives. *Eur. J. Appl. Physiol.* **64,** 178-181.

39. Coote, J.H., Tsang, G., Baker, A., and Stone, B. (1993). Respiratory changes and structure of sleep in young high-altitude dwellers in the Andes of Peru. *Eur. J. Appl. Physiol.* **66,** 249-253.

40. Plywaczewski R, *et al.* (2003) Sleep structure and periodic breathing in Tibetans and Han at simulated altitude of 5000 m. *Resp. Physiol. Neurobi.* **136,** 187-197.

41. Plywaczewski, R., Wu, T.-Y., Wang, X.-Q., Cheng, H.-W., Sliwinski, P., Zielinski, J. (2003). Sleep structure and periodic breathing in Tibetans and Han at simulated altitude of 5000 m. *Resp. Physiol. Neurobi.* **136,** 187-197.

42. International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52-58.

43. Pemberton, T.J., Wang, C., Li, J.Z., Rosenberg, N.A. (2010). Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *Am. J. Hum. Genet*. **87,** 45-464.

44. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100-1104.

45. Rosenberg, N.A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet*. **70,** 841-847.

46. Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., et al. (2011). Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *Am. J. Hum. Genet*. **89,** 368-381.

47. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2009). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78-81.

48. Reynolds, J., Weir, B.S., and Cockerham, C.C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*. 1**05,** 767-779.

49. Hernandez, R.D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bionformatics* **24**, 2786-2787.

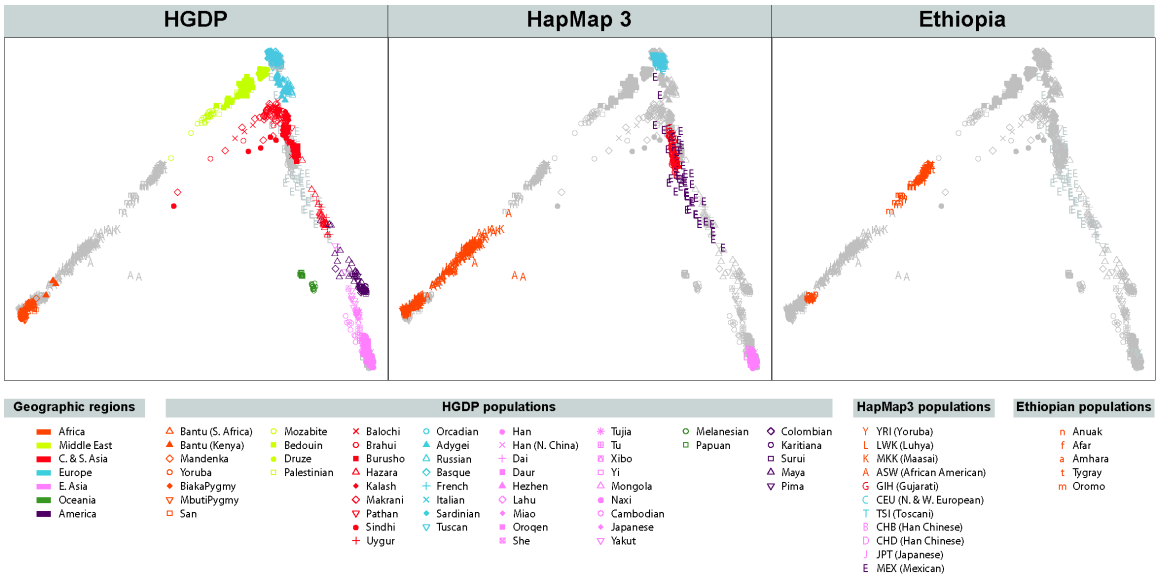

Figure 1: Multidimensional scaling for the HGDP, HapMap 3, and Ethiopian populations. Note that the Anuak individuals lie between the Yoruban and African American individuals, and the Afar, Amhara, Tigray and Oromo individuals lie between the African American and Middle Eastern individuals.

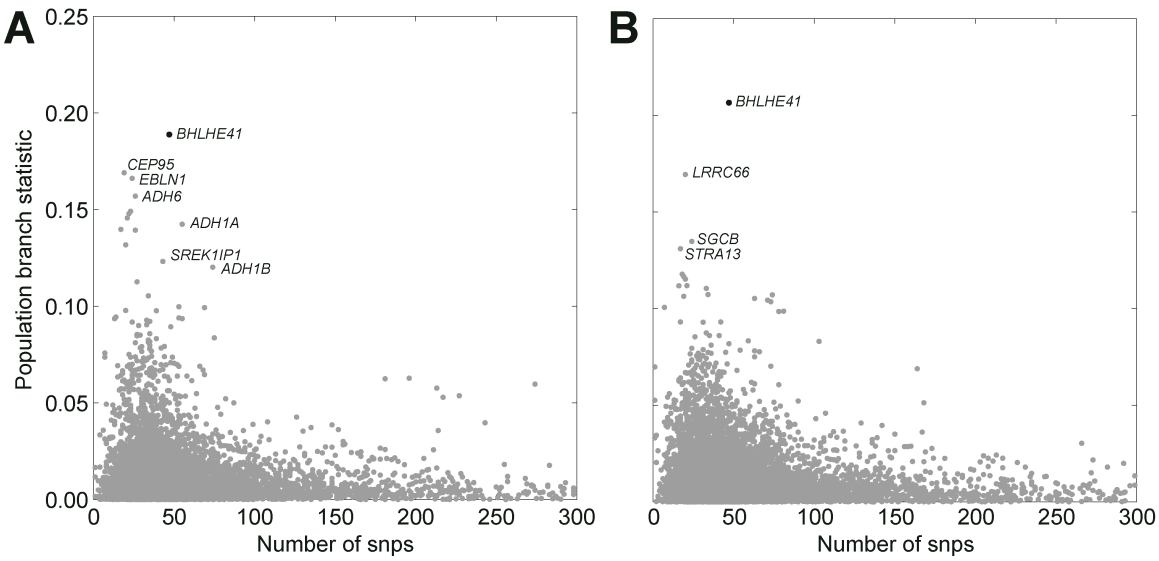

Figure 2: The empirical distribution of PBS values per gene region as a function of the number of SNPs in the gene. The *y*-axis is the corresponding PBS value of the gene region with a given number of SNPs (the *x*-axis). The *x*-axis has been truncated at 300 SNPs.

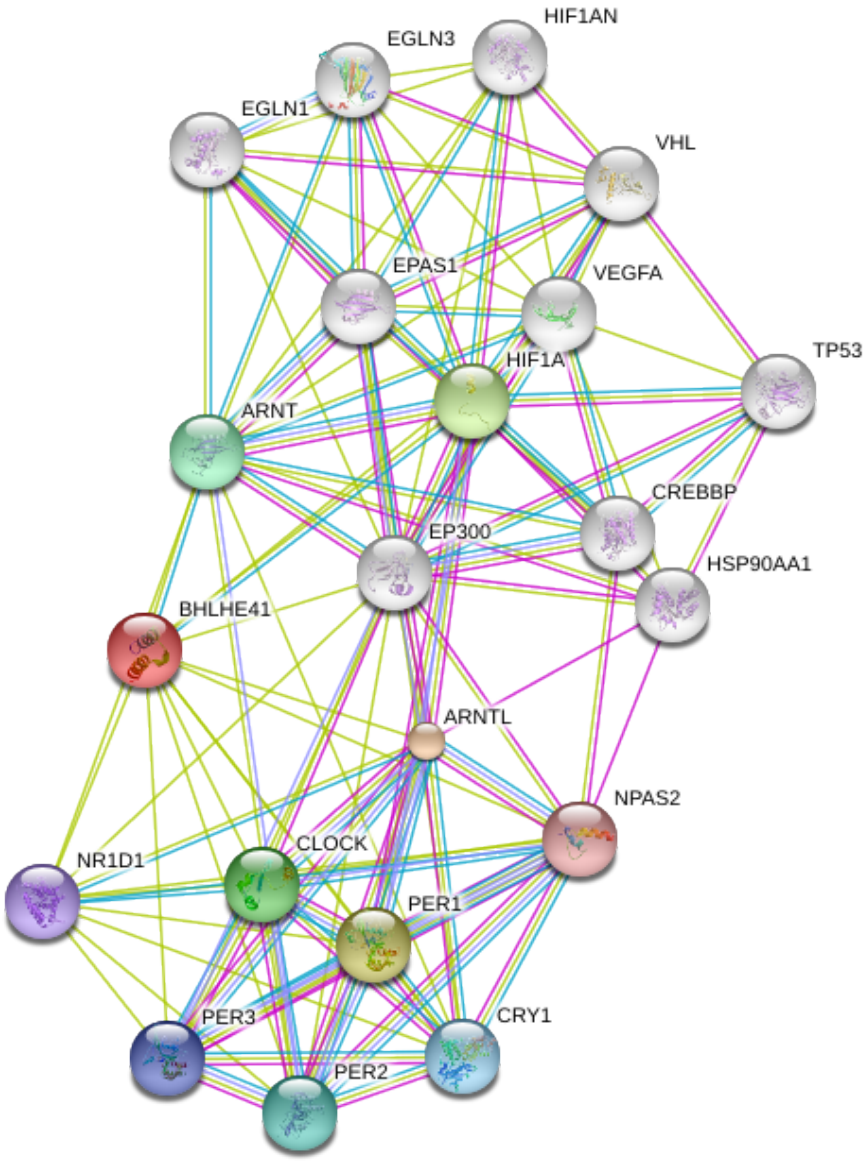

Figure 3: STRING 9.0 (Jensen et al., 2009) database of interactions with BHLHE41, including up to 10 direct links with other genes and 10 genes separated by two links from BHLHE41. Colored links connected to BHLHE41 are from PubMed co-occurrence (yellow) and co-membership in pathways from the NCI-Nature Pathways Interaction Database (light blue).

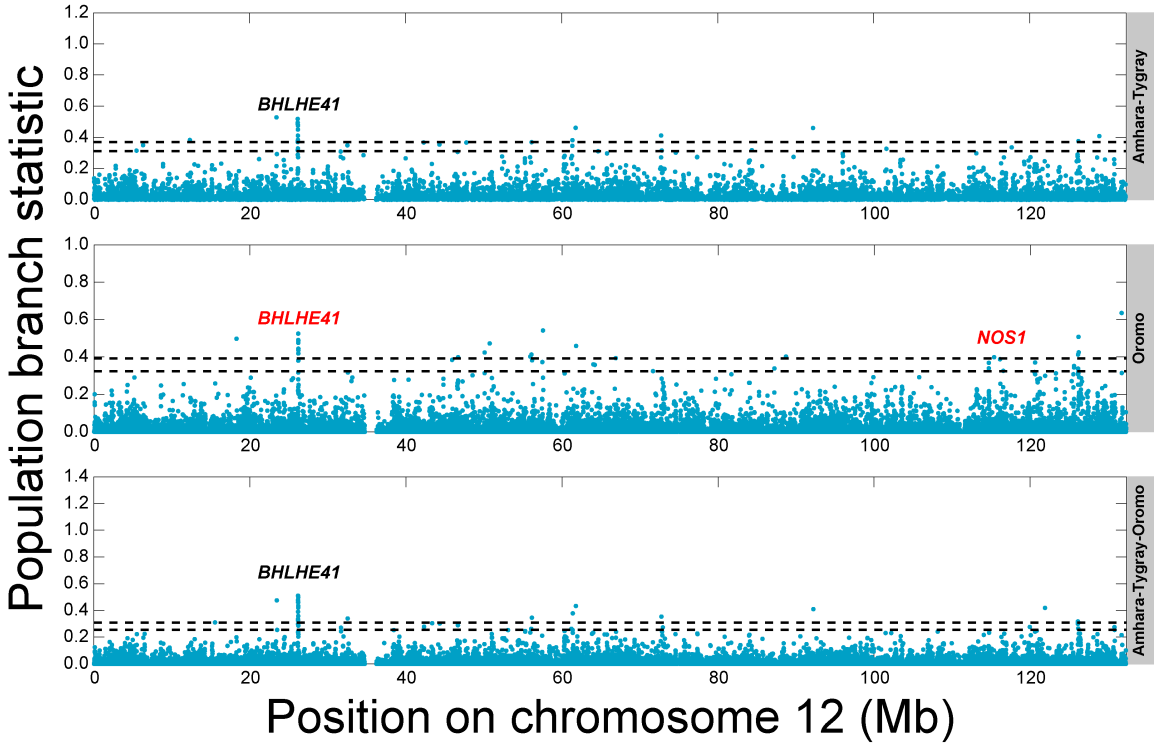

Figure 4: Per-SNP population branch statistic results. Gaps represent regions of the genome that were not covered. Names of genes in black contain at least one SNP in the top 0.10% of SNPs that is located inside the gene. Genes in red contain at least one SNP in the top 0.10% that is located within 50kb of the gene but not within the gene. The top and bottom horizontal dotted lines are the 0.05% and 0.10% empirical cutoffs, respectively. The names in the shaded gray area are the population(s) considered (Amhara-Tigray, Oromo and Amhara-Tigray-Oromo). Figures S6-S8 display all 22 autosomes for each population comparison shown here.

Supplementary material: Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations

Emilia Huerta-Sánchez[1,2,*], Michael DeGiorgio[1,*], Luca Pagani[3,4,*], Ayele Tarekegn[5], Rosemary Ekong[6], Tiago Antao[3], Alexia Cardona[3], Hugh E. Montgomery[7], Gianpiero L. Cavalleri[8], Peter A. Robbins[9], Michael E. Weale[10], Neil Bradman[6], Endashaw Bekele[5], Toomas Kivisild[3], Chris Tyler-Smith[4], Rasmus Nielsen[1,2,11,12]

[1]Department of Integrative Biology, University of California, Berkeley, CA, USA
[2]Department of Statistics, University of California, Berkeley, CA, USA
[3]Division of Biological Anthropology, University of Cambridge, Cambridge, UK
[4]Wellcome Trust Sanger Institute, Hinxton, UK
[5]Addis Ababa University and Center of Human Genetic Diversity, P.O. Box 1176, Addis Ababa, Ethiopia
[6]The Centre for Genetic Anthropology, Department of Genetics, Evolution and Environment, University College London, London, UK
[7]Institute for Human Health and Performance, University College London, London, UK
[8]Molecular and Cellular Therapeutics, The Royal College of Surgeons in Ireland, Dublin, Ireland
[9]Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford, UK
[10]Department of Medical and Molecular Genetics, King's College London, London, UK
[11]Department of Biology, University of Copenhagen, Copenhagen, Denmark
[12]Beijing Genomics Institute, Shenzhen, China

*These authors contributed equally

Table S1: Top genes without admixture correction for the combined AmharaTigrayan population

Gene name	Chr	Description	PBS
MYEF2	15	myelin expression factor 2	0.199
SLC24A5	15	solute carrier family 24, member 5	0.186
OAZ2	15	ornithine decarboxylase antizyme 2	0.159
ADH6	4	alcohol dehydrogenase 6 isoform 1	0.159
ADH1A	4	class I alcohol dehydrogenase, alpha subunit	0.156
ATP6V1G3	1	ATPase, H+ transporting, lysosomal, V1 subunit	0.152
CEP95	17	Homo sapiens centrosomal protein 95kDa (CEP95), mRNA.	0.150
TRIP4	15	thyroid hormone receptor interactor 4	0.145
NR1H3	11	nuclear receptor subfamily 1, group H, member 3	0.144
KIAA0101	15	hypothetical protein LOC9768 isoform 1	0.142
ZNF609	15	zinc finger protein 609	0.139
ACP2	11	acid phosphatase 2, lysosomal isoform 1	0.139
BHLHE41	12	basic helix-loop-helix domain containing, class	0.138
SHE	1	Src homology 2 domain containing E	0.138
DDX5	17	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	0.134
DKK4	8	dickkopf homolog 4 precursor	0.134
VDAC3	8	voltage-dependent anion channel 3 isoform b	0.134
POLB	8	polymerase (DNA directed), beta	0.133
EPGN	4	epithelial mitogen	0.133
POLG2	17	polymerase (DNA directed), gamma 2, accessory	0.133
ADH1B	4	class I alcohol dehydrogenase, beta subunit	0.131
GMFG	19	glia maturation factor, gamma	0.126
ZEB1	10	zinc finger E-box binding homeobox 1 isoform b	0.125
TDRD10	1	tudor domain containing 10 isoform b	0.124
MILR1	17	Homo sapiens mast cell immunoglobulin-like receptor 1 (MILR1), mRNA.	0.122

Table S2: Top genes with admixture correction for the combined Amhara-Tigray population

Gene name	Chr	UCSC description	Nearby Genes	PBS
BHLHE41	12	basic helix-loop-helix domain containing, class	**RASSF8**,SSPN	0.1889
CEP95	17	Homo sapiens centrosomal protein 95kDa (CEP95), mRNA.		0.1691
EBLN1	10	Homo sapiens endogenous Bornavirus-like nucleoprotein 1 (EBLN1), mRNA	BMI1	0.1661
ADH6	4	alcohol dehydrogenase 6 isoform 2	ADH4,ADH1A	0.1569
DDX5	17	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	**PICAM**1,POLG2, MIR5047	0.1492
POLG2	17	polymerase (DNA directed), gamma 2, accessory	DDX5	0.1479
VPS36	13	Homo sapiens vacuolar protein sorting 36 homolog (S. cerevisiae) (VPS36), mRNA	THSD1	0.1457
ADH1A	4	class I alcohol dehydrogenase, alpha subunit		0.1424
CKAP2	13	cytoskeleton associated protein 2 isoform 2	THSD1	0.1398
MILR1	17	Homo sapiens mast cell immunoglobulin-like receptor 1 (MILR1), mRNA.	POLG2,DDX5, **PECAM1**	0.1394
THSD1	13	thrombospondin type I domain-containing 1		0.1318
SREK1IP1	5	Homo sapiens SREK1-interacting protein 1 (SREK1IP1), mRNA.		0.1232
ADH1B	4	class I alcohol dehydrogenase, beta subunit		0.1202
SMURF2	17	SMAD specific E3 ubiquitin protein ligase 2	POLG3,DDX5,MIR5047,**PECAM**1	0.1126
SLC35C1	11	Homo sapiens solute carrier family 35, member C1 (SLC35C1), transcript variant 1, mRNA.	CRY2,CHST1, MAPK8IP1	0.1054
CARD16	11	caspase-1 dominant-negative inhibitor pseudo-ICE	**CASP1**	0.0998
ADH1C	4	alcohol dehydrogenase 1C, gamma polypeptide		0.0993
NR4A2	2	nuclear receptor subfamily 4, group A, member 2	GPD2	0.0979
NANS	9	N-acetylneuraminic acid phosphate synthase	HEMGN, ANP32B	0.0976
OR4C13	11	olfactory receptor, family 4, subfamily C		0.0944
GPR111	6	Homo sapiens G protein-coupled receptor 111 (GPR111), mRNA.		0.0939
CASP1	11	caspase 1 isoform alpha precursor		0.0936
LYPLA1	8	lysophospholipase I		0.0936
SHE	1	Src Homology 2 domain containing E		0.0928
ZNF598	16	zinc finger protein 598		0.092

Table S3: Top genes with admixture correction for the Tigrayan population

Gene name	Chr	Description	PBS
PCYOX1	2	prenylcysteine oxidase 1	0.238
ADH6	4	class I alcohol dehydrogenase, beta subunit	0.228
C2orf42	2	ypothetical protein LOC54980	0.224
GATSL2	7	Homo sapiens GATS protein-like 2 (GATSL2), mRNA.	0.210
ADH1A	4	class I alcohol dehydrogenase, alpha subunit	0.203
EPGN	4	epithelial mitogen	0.202
IL29	19	interleukin 29	0.197
BHLHE41	12	basic helix-loop-helix domain containing, class	0.193
PSMA5	1	proteasome alpha 5 subunit	0.190
KRTCAP3	2	keratinocyte associated protein 3	0.188
ISL1	5	islet-1	0.184
LRFN1	19	leucine rich repeat and fibronectin type III	0.177
PCBP1	2	poly(rC) binding protein 1	0.177
CARD16	11	caspase-1 dominant-negative inhibitor pseudo-ICE	0.174

Table S4: Top genes with admixture correction for the Amhara population

Gene name	Chr	Description	PBS
STX3	11	syntaxin 3	0.221
CEP95	17	Homo sapiens centrosomal protein 95kDa (CEP95), mRNA.	0.208
NANS	9	N-acetylneuraminic acid phosphate synthase	0.206
EBLN1	10	Homo sapiens endogenous Bornavirus-like nucleoprotein 1	0.191
POLG2	17	polymerase (DNA directed), gamma 2, accessory	0.190
BHLHE41	12	basic helix-loop-helix domain containing, class	0.188
VPS36	13	vacuolar protein sorting 36	0.184
DDX5	17	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	0.182
OR10V1	11	olfactory receptor, family 10, subfamily V	0.182
MILR1	17	Homo sapiens mast cell immunoglobulin-like receptor 1, mRNA.	0.178
CKAP2	13	Homo sapiens cytoskeleton associated protein 2, transcript variant 1,	0.177
MRPL16	11	Homo sapiens mitochondrial ribosomal protein L16	0.174
THSD1	13	thrombospondin type I domain-containing 1	0.165
SMURF2	17	SMAD specific E3 ubiquitin protein ligase 2	0.161
NEK5	13	NIMA-related kinase 5	0.156

Table S5: Top genes with admixture correction for the Oromo population

Gene name	Chr	Description	PBS
BHLHE41	12	basic helix-loop-helix domain containing, class	0.207
LRRC66	4	leucine rich repeat containing 66	0.169
SGCB	4	sarcoglycan, beta	0.135
STRA13	17	stimulated by retinoic acid 13	0.131
RFNG	17	radical fringe homolog	0.118
GPS1	17	G protein pathway suppressor 1 isoform 1	0.118
RAC3	17	ras-related C3 botulinum toxin substrate 3 (rho	0.117
DCXR	17	Homo sapiens dicarbonyl/L-xylulose reductase (DCXR), transcript variant 2, mRNA.	0.117
LRRC45	17	leucine rich repeat containing 45	0.115
C8orf22	8	hypothetical protein LOC492307	0.112
LARP1B	4	La ribonucleoprotein domain family member 2	0.112
ZCCHC6	9	zinc finger, CCHC domain containing 6	0.110
RASSF6	4	Ras association (RalGDS/AF-6) domain family 6	0.107
ELP3	8	elongation protein 3 homolog	0.107
DUS1L	17	PP3111 protein	0.106

Table S6: Top genes with admixture correction for the combined Amhara-Tigray-Oromo populations

Gene name	Chr	Description	PBS
BHLHE41	12	basic helix-loop-helix domain containing, class	0.197
EBLN1	10	Homo sapiens endogenous Bornavirus-like nucleoprotein 1 (EBLN1), mRNA	0.137
CEP95	17	Homo sapiens centrosomal protein 95kDa (CEP95), mRNA.	0.118
ADH6	4	alcohol dehydrogenase 6 isoform 2	0.116
POLG2	17	polymerase (DNA directed), gamma 2, accessory	0.113
DDX5	17	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	0.109
MILR1	17	Homo sapiens mast cell immunoglobulin-like receptor 1 (MILR1), mRNA	0.107
ADH1A	4	class I alcohol dehydrogenase, alpha subunit	0.100
ZCCHC6	9	zinc finger, CCHC domain containing 6	0.096
VPS36	13	vacuolar protein sorting 36	0.096
CKAP2	13	cytoskeleton associated protein 2 isoform 1	0.090
POP7	7	processing of precursor 7, ribonuclease P/MRP	0.087
SMURF2	17	SMAD specific E3 ubiquitin protein ligase 2	0.084
ACTL6B	7	actin-like 6B	0.083
ADH1B	4	class I alcohol dehydrogenase, beta subunit	0.082
STRA13	17	stimulated by retinoic acid 13	0.079
OTX1	2	orthodenticle homeobox 1	0.078
SHE	1	Src Homology 2 domain containing E	0.078
NR4A2	2	Nuclear receptor subfamily 4, group A, member 2	0.076
ISCA1	9	HESB like domain containing 2	0.075
GNB2	7	guanine nucleotide-binding protein, beta-2	0.075
GIGYF1	7	GRB10 interacting GYF protein 1	0.075
RAC3	17	ras-related C3 botulinum toxin substrate 3 (rho	0.073
DCXR	17	Homo sapiens dicarbonyl/L-xylulose reductase (DCXR), transcript variant 2, mRNA.	0.073
SSPN	12	sarcospan isoform 1	0.073

Table S7: Enrichment of hypoxia genes. The expected number of hypoxia-related genes (as defined in the *Methods* section) under each of the scenarios considered.

Populations	Median Number of Hypoxia Related Genes
Oromo-Tygray-Amhara with correction	7
Amhara-Tygray with correction	6
Amhara-Oromo with correction	8
Oromo-Tygray with correction	7
Amhara with correction	6
Tygray with correction	6
Oromo with correction	7
Oromo-Tygray-Amhara no correction	11
Amhara-Tygray no correction	9
Amhara-Oromo no correction	11
Oromo-Tygray no correction	11
Amhara no correction	7
Tygray no correction	6
Oromo no correction	12

Table S8: Genes located close to SNPs in the top 0.10% of SNPs. A gene under the "In genes" column indicates that at least one SNP in the top 880 SNPs (0.10%) is contained in that gene.

Top .1% (880) SNPs		
Populations	**Hypoxia Related Genes**	
	In genes	**Within 50kb of a gene**
Oromo-Tygray-Amhara with correction	BHLHE41, ACE, ENG, RYR2, SFRP1, SLC8A1	BHLHE41, ACE, CASP1, CHRNB2, ENG, EPO, HIF3A, MMP2, PDGFB, PLOD1, RYR2, SFRP1, SLC8A1
Amhara-Tygray with correction	BHLHE41, ENG, NF1, RYR2, SFRP1	BHLHE41, AQP1, CASP1, CHRNB2, ENG, HIF3A, NF1, PDGFB, PLOD1, RYR2, SFRP1
Amhara-Oromo with correction	BHLHE41, ENG, RYR2, SDC2, SFRP1, SLC8A1	BHLHE41, CASP1, ENG, HIF3A, MMP2, PDIA2, RYR2, SDC2, SFRP1, SLC8A1
Oromo-Tygray with correction	BHLHE41, ACE, ENG, RYR2	BHLHE41, ACE, CASP1, CHRNB2, ENG, EPO, PDGFB, PDIA2, PLOD1, RYR2
Amhara with correction	CD38, NF1, PPARA, RYR2, SMAD9	BHLHE41, CA9, CASP1, CD38, CHRNB2, NF1, PPARA, RYR2, SMAD9, TRH
Tygray with correction	BHLHE41, ACE, EPAS1, NF1, RYR2	BHLHE41, ACE, CASP1, CHRNB2, EPAS1, NF1, PDGFB, PLOD1, RYR2
Oromo with correction	RYR2, SLC8A1	BHLHE41, ANGPTL4, EGLN3, JAG2, MB, NOS1, PDIA2, RYR2, SFRP1, SLC8A1
Oromo-Tygray-Amhara no correction	BHLHE41, ANGPT2, ENG, RYR2, SFRP1	BHLHE41, ANGPT2, AQP1, CA9, CHRNB2, ENG, EPO, PDIA2, PLAT, PTK2B, PYGM, RYR2, SFRP1
Amhara-Tygray no correction	BHLHE41, ANGPT2, ENG, RYR2, SFRP1	BHLHE41, ANGPT2, AQP1, CA9, CASP1, CHRNB2, ENG, EPO, FLT1, PYGM, RYR2, SFRP1
Amhara-Oromo no correction	BHLHE41, ANGPT2, DDIT4, ENG, ITPR2, RYR2, SDC2, SFRP1, SLC8A1	BHLHE41, ANGPT2, CA9, CHRNB2, DDIT4, ENG, EPO, FLT1, ITPR2, PDIA2, PTK2B, PYGM, RYR2, SDC2, SFRP1, SLC8A1
Oromo-Tygray no correction	BHLHE41, ANGPT2, ENG, RYR2	BHLHE41, ANGPT2, CHRNB2, ENG, EPO, JAG2, PDIA2, PLAT, PTK2B, PYGM, RYR2
Amhara no correction	ANGPT2, ITPR2, PPARA, RYR2, SFRP1	BHLHE41, ANGPT2, CA9, CHRNB2, FLT1, ITPR2, PPARA, PYGM, RYR2, SFRP1
Tygray no correction	BHLHE41, RYR2, SCNN1G	BHLHE41, CASP1, CHRNB2, EPO, PLAT, RYR2, SCNN1G
Oromo no correction	ANGPT2, DDIT4, ITPR1, RYR2, SLC8A1	BHLHE41, ANGPT2, ANGPTL4, ASCL2, CHRNB2, DDIT4, EGLN3, EPO, ITPR1, JAG2, PDIA2, PYGM, RYR2, SLC8A1

Table S9: Hypoxia gene set

ABAT	CRYAA	LONP1	SCNN1B
ACE	CRYAB	MB	SCNN1G
ACSL6	CST3	MMP14	SDC2
ACTN4	CXCR4	MMP2	SERPINA1
ADA	CYP1A1	MT3	SFRP1
ADAM17	DDIT4	MT-CYB	SHH
ADIPOQ	DPP4	MT-ND4	SLC11A2
ADM	E2F1	MT-ND5	SLC2A8
ALAS2	ECE1	NARFL	SLC8A1
ALDOC	EDN1	NF1	SMAD3
ANG	EDNRA	NOS1	SMAD4
ANGPT2	EGLN1	NOS3	SMAD9
ANGPT4	EGLN2	NPPC	SOCS3
ANGPTL4	EGLN3	P2RX3	SOD2
APOLD1	ENG	PAM	SOD3
AQP1	EP300	PDE5A	STAT5B
ARNT	EPAS1	PDGFA	TFRC
ARNT2	EPO	PDGFB	TGFB1
ASCL2	ERCC3	PDIA2	TGFB2
ATP1B1	FLT1	PDLIM1	TGFB3
BCL2	GATA6	PGF	TGFBR1
BIRC2	GPR182	PLAT	TH
BMP2	HIF1A	PLAU	THBS1
BNIP3	HIF3A	PLD2	TLR2
CA9	HMOX1	PLOD1	TLR4
CAPN2	HMOX2	PLOD2	TRH
CASP1	HSD11B2	PML	TXN2
CAV1	HSP90B1	PPARA	UBE2B
CCL2	HYOU1	PRKAA1	UBQLN1
CCNB1	IL18	PRKCQ	UCN3
CD24	ITGA2	PSEN2	UCP2
CD38	ITPR1	PTK2B	UCP3
CHRNA4	ITPR2	PYGM	USF1
CHRNA7	JAG2	RHOA	VCAM1
CHRNB2	KCNA5	RYR1	VEGFA
CITED2	KCNK3	RYR2	VHL
CLDN3	KCNMA1	SCAP	VLDLR
CREBBP	LCT	SCFD1	XRCC1

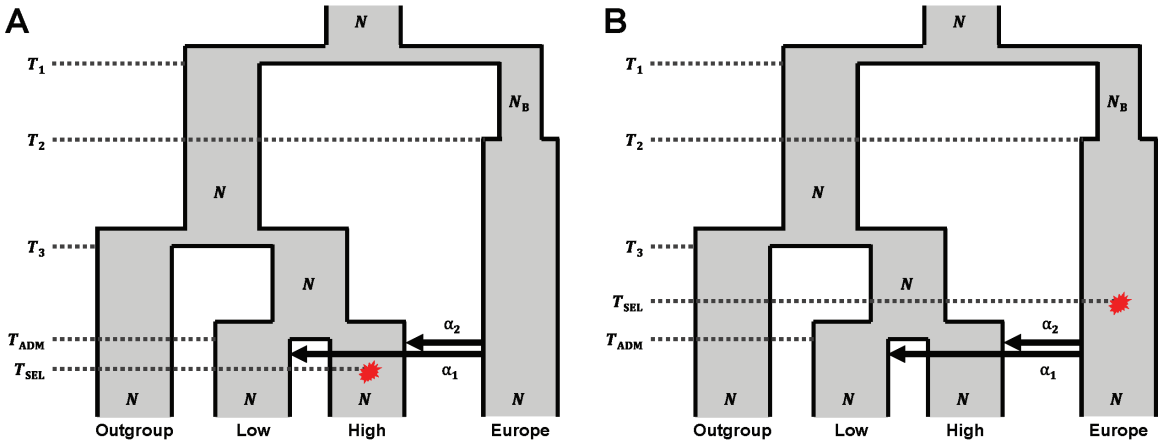

Figure S1: Demographic model and selection scenarios used to assess the performance of the corrected and uncorrected PBS test statistics. At time T_1 in the past, the African and the non-African populations split. The non-African populations experienced a bottleneck with small population size N_B for T_1 - T_2 years and then recovered to size N. At time T_3, the African population split into two populations. At time T_{ADM}, the low- and the high-altitude populations are created using genetic contributions from an ancestral population and the non-African populations. The contributions from the non-African population into the low- and high-altitude populations are α_1 and α_2, respectively. Aside from the non-African population during its bottleneck, all populations have size N. The parameters T_1, T_2, T_3, and N_B were kept fixed at 60kya, 40kya, 10kya, and 0.5N respectively. For the pair (α_1, α_2), we considered combinations of (0.1, 0.2), (0.1, 0.3), (0.1, 0.4), and (0.2, 0.4). For the time of admixture T_{ADM}, we considered values of 1.5kya and 3kya. The population-scaled selection parameters considered are S = 150 and 250. The timing of selection T_{SEL} in the high-altitude population occurred at either 3kya or 1.5kya. Finally, the timing of selection in the non-African population was set to T_{SEL} = 5kya.

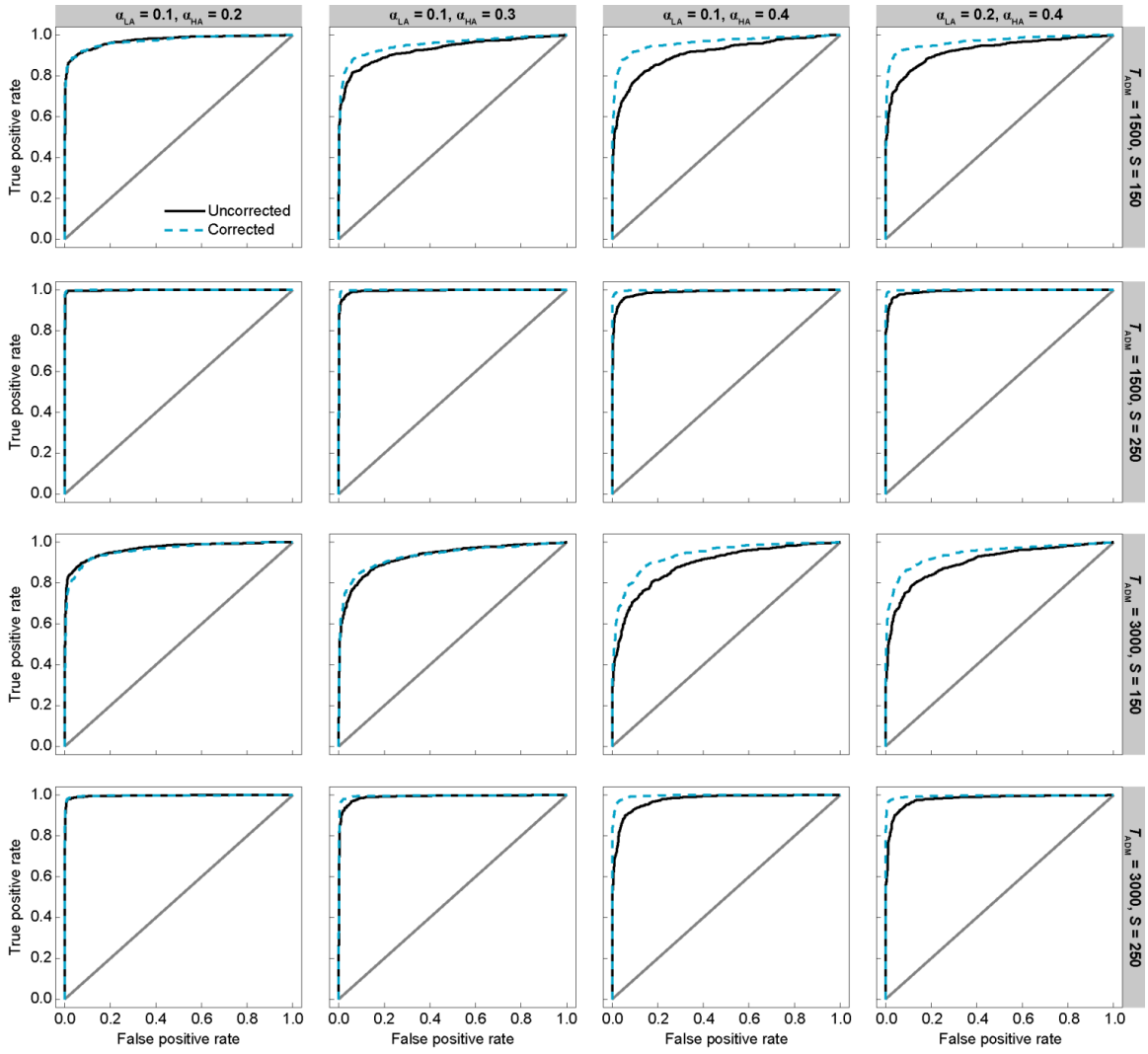

Figure S2: Receiver operator characteristic curves for the selection scenario in Figure S1*A*. We consider divergence times of T_{ADM} =1.5kya and 3kya. We consider population-scaled selection parameters of S = 150 and 250. Each column represents a different set of admixture fractions for the low- and high-altitude populations.

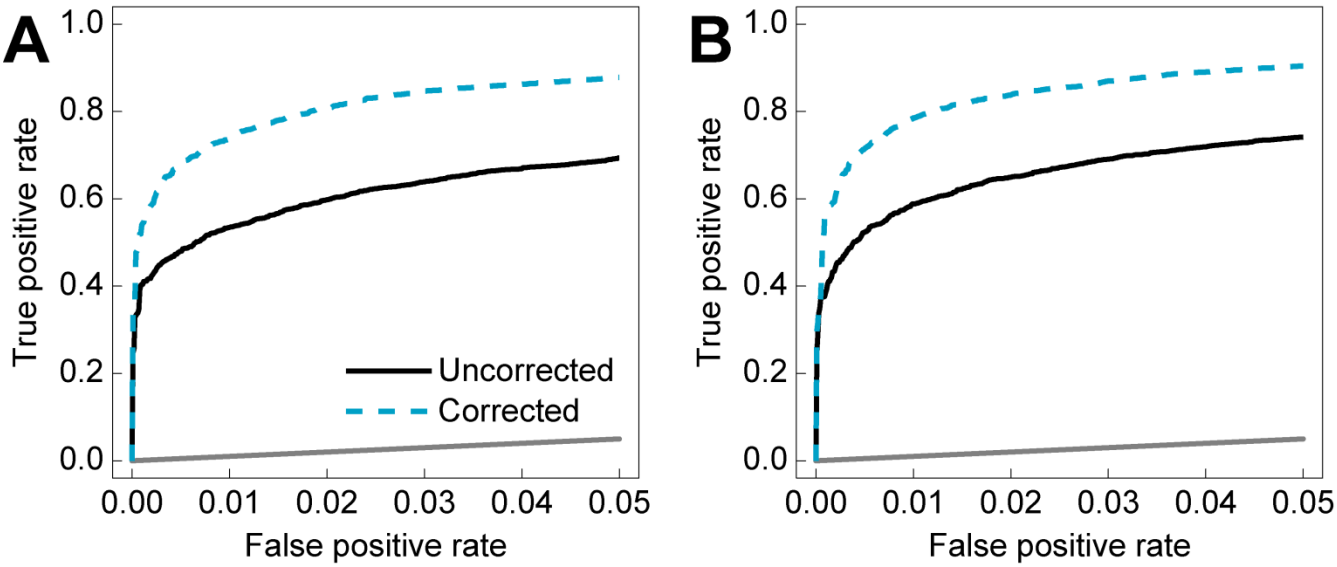

Figure S3: Two cases in Figure S2 zoomed in for false positive rates in the range from 0 to 0.05. (*A*) Receiver operator characteristic (ROC) curve in which admixture occurs at time T_{ADM} = 1.5kya with fraction α_1 = 0.1 for the low- and α_2 = 0.4 for the high-altitude population, and in which selection occurs at time T_{SEL} = 1.5kya with strength S = 150. (*B*) ROC curve in which admixture occurs at time T_{ADM} = 1.5kya with fraction α_1 = 0.2 for the low- and α_2 = 0.4 for the high-altitude population, and in which selection occurs at time T_{SEL} = 1.5kya with strength S = 150.

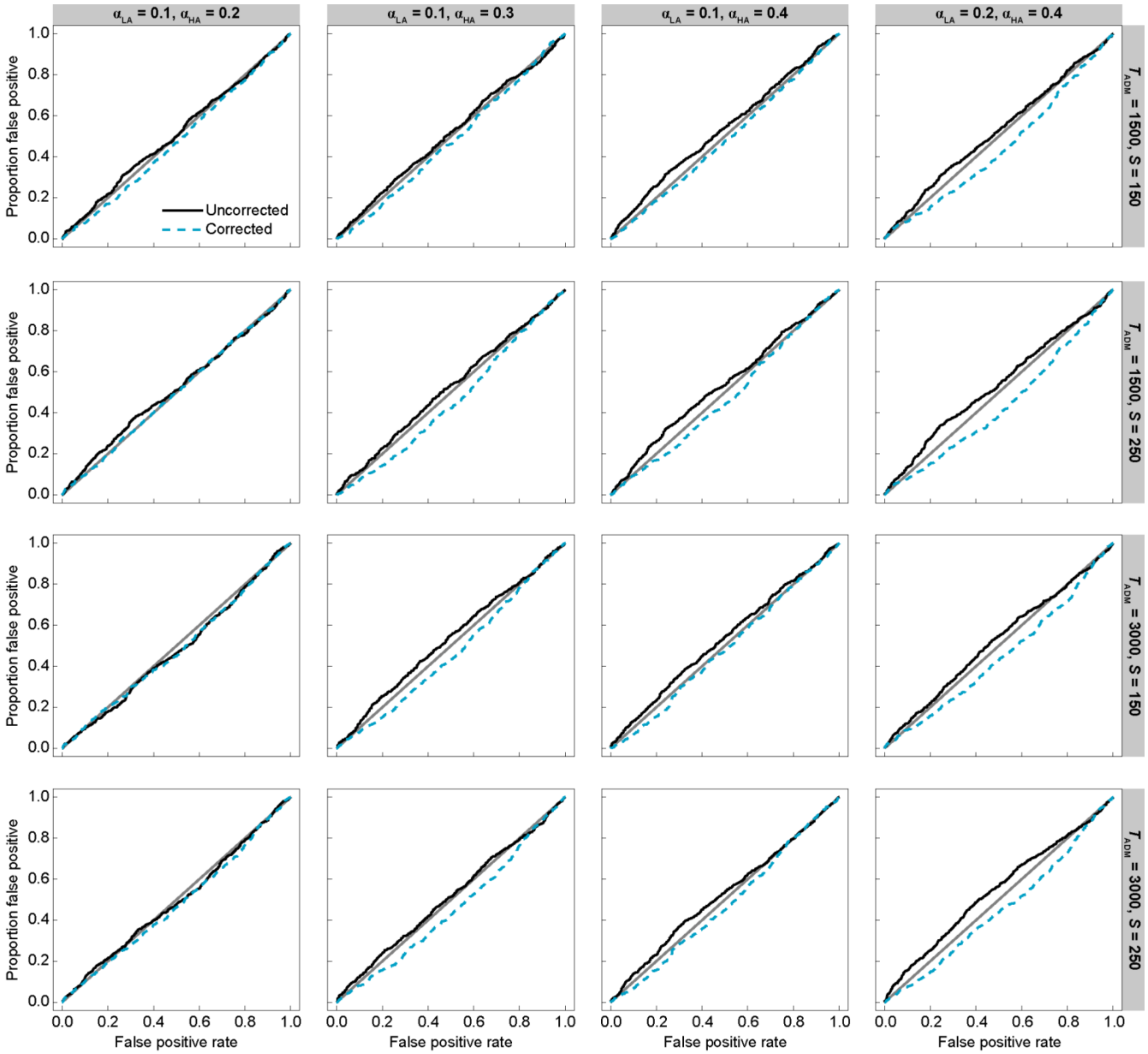

Figure S4: Proportion of times that a simulation is falsely called a positive as a function false positive rate for the selection scenario in Figure S1*B*. We consider divergence times of T_{ADM} =1.5kya and 3kya. We consider population-scaled selection parameters of S = 150 and 250. Each column represents a different set of admixture fractions for the low- and high-altitude populations.

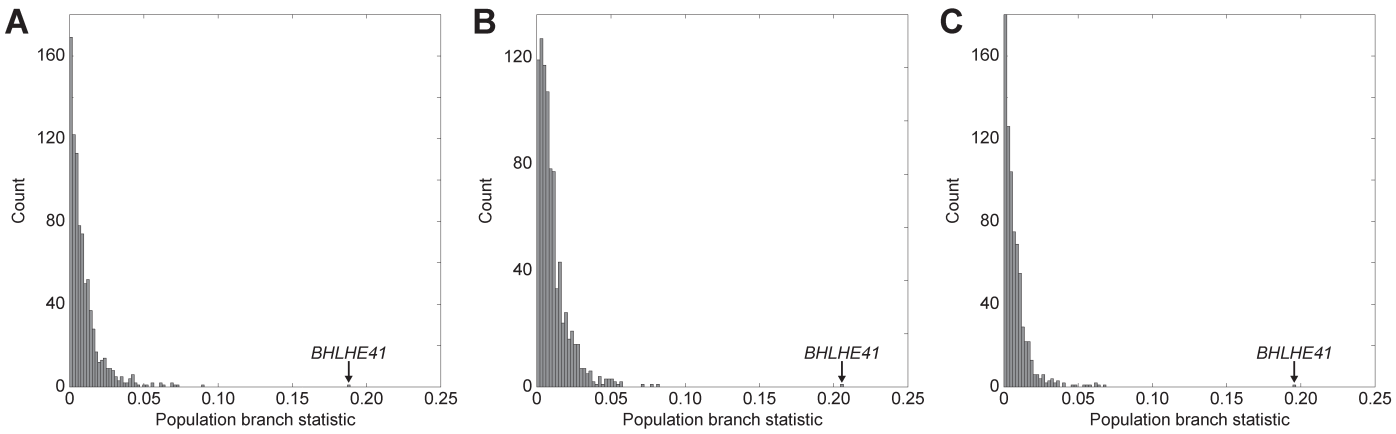

Figure S5: The empirical distribution of PBS values for genes with 45 to 49 SNPs. *BHLHE41* contains 47 SNPs. Panel *A* corresponds to the Amhara-Tygray selection scan, panel *B* corresponds to the Oromo selection scan, and panel *C* corresponds to the Amhara-Tygray-Oromo selection scan.

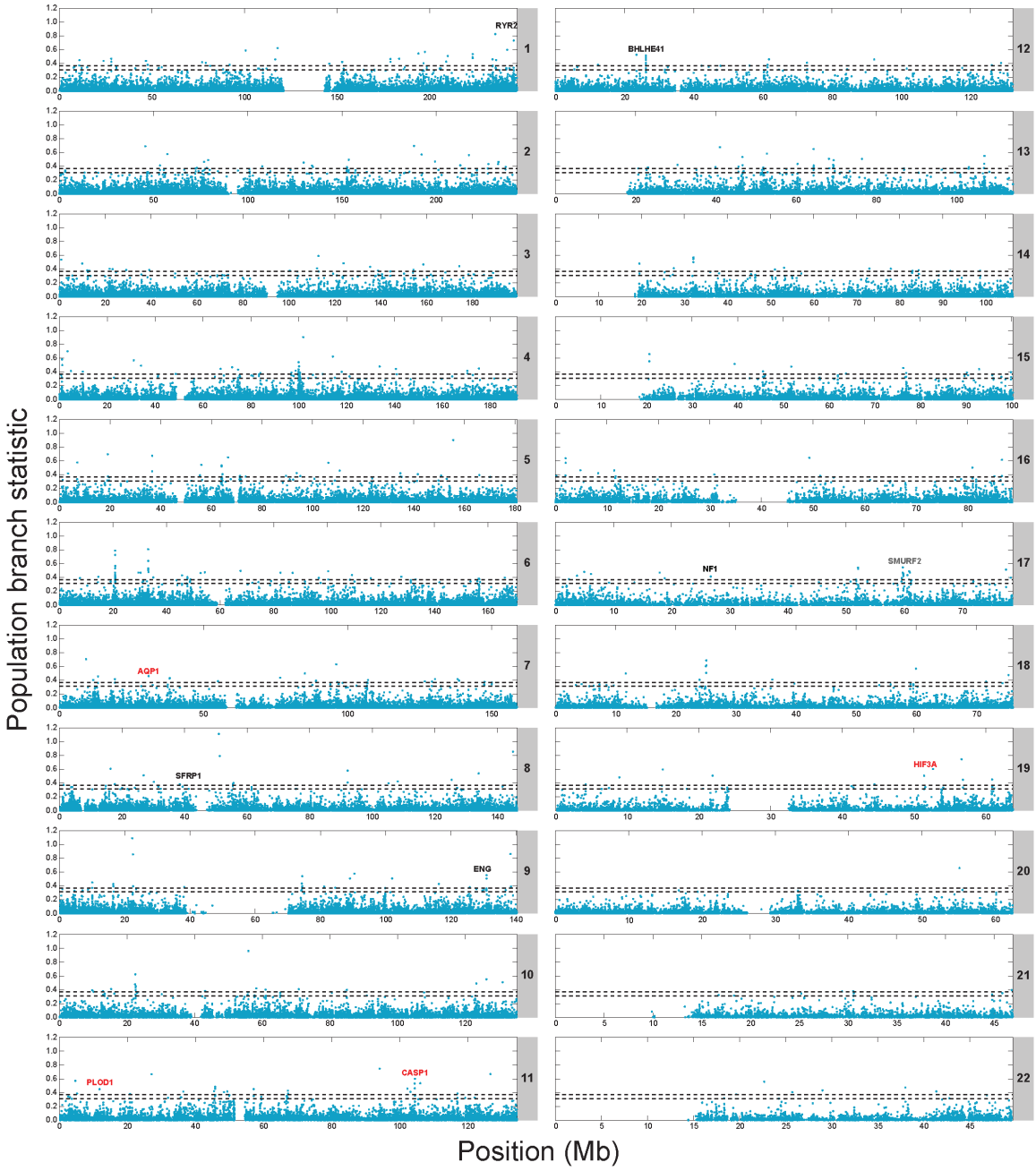

Figure S6: Per-SNP population branch statistic. Results for the analyses of the Tigray and Amhara populations combined. Gaps represent regions of the genome that were not covered. Names of genes in black contain at least one SNP in the top 0.10% of SNPs that is located inside the gene. Genes in red contain at least one SNP in the top 0.10% that is located within 50kb of the gene but not within the gene. The numbers in the shaded gray area are the chromosome number. The top and bottom horizontal dotted lines are the 0.05% and 0.10% empirical cutoffs, respectively.

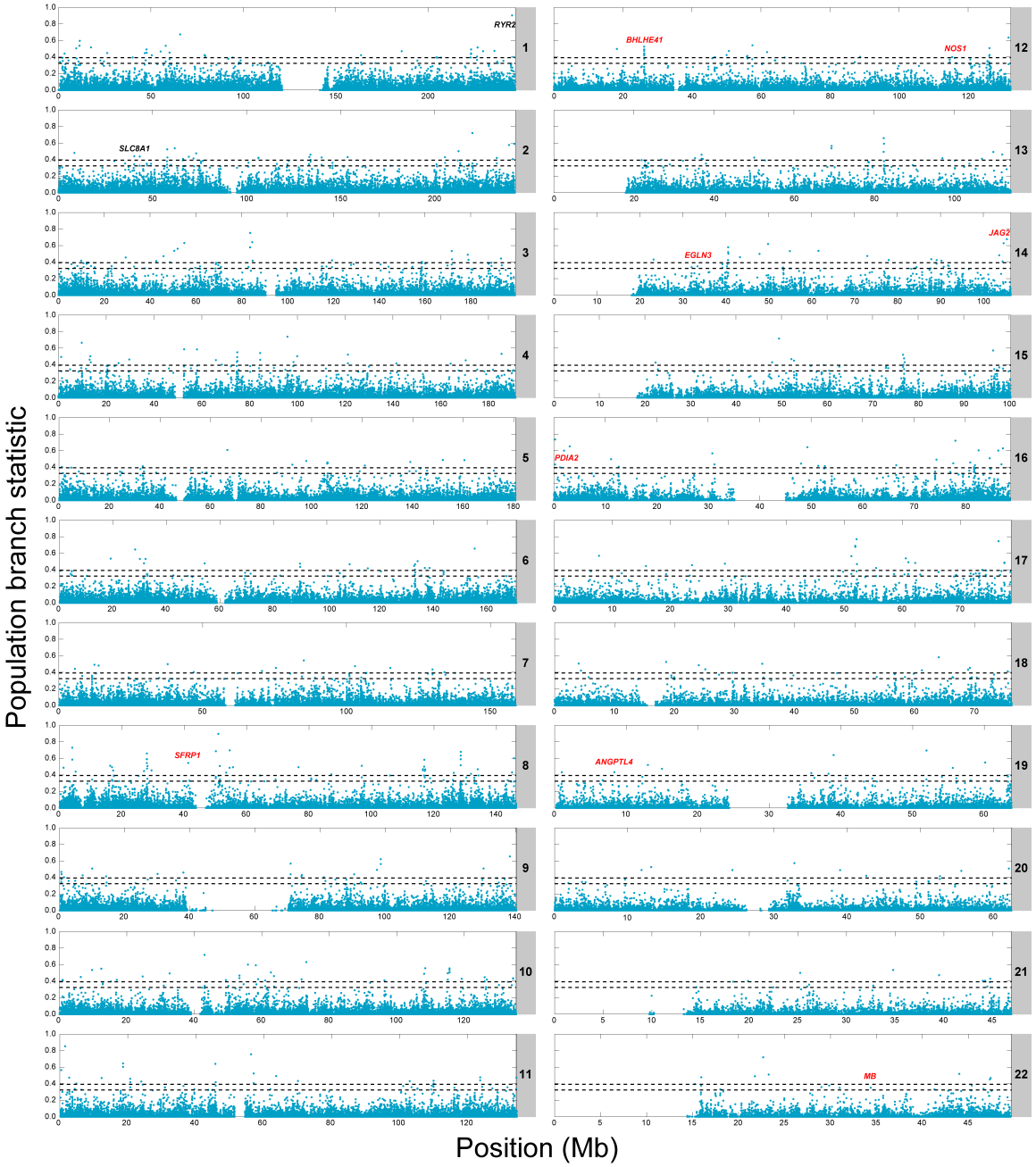

Figure S7: Per-SNP population branch statistic. Results for the analyses of the Oromo population. Gaps represent regions of the genome that were not covered. Names of genes in black contain at least one SNP in the top 0.10% of SNPs that is located inside the gene. Genes in red contain at least one SNP in the top 0.10% that is located within 50kb of the gene but not within the gene. The numbers in the shaded gray area are the chromosome number. The top and bottom horizontal dotted lines are the 0.05% and 0.10% empirical cutoffs, respectively.

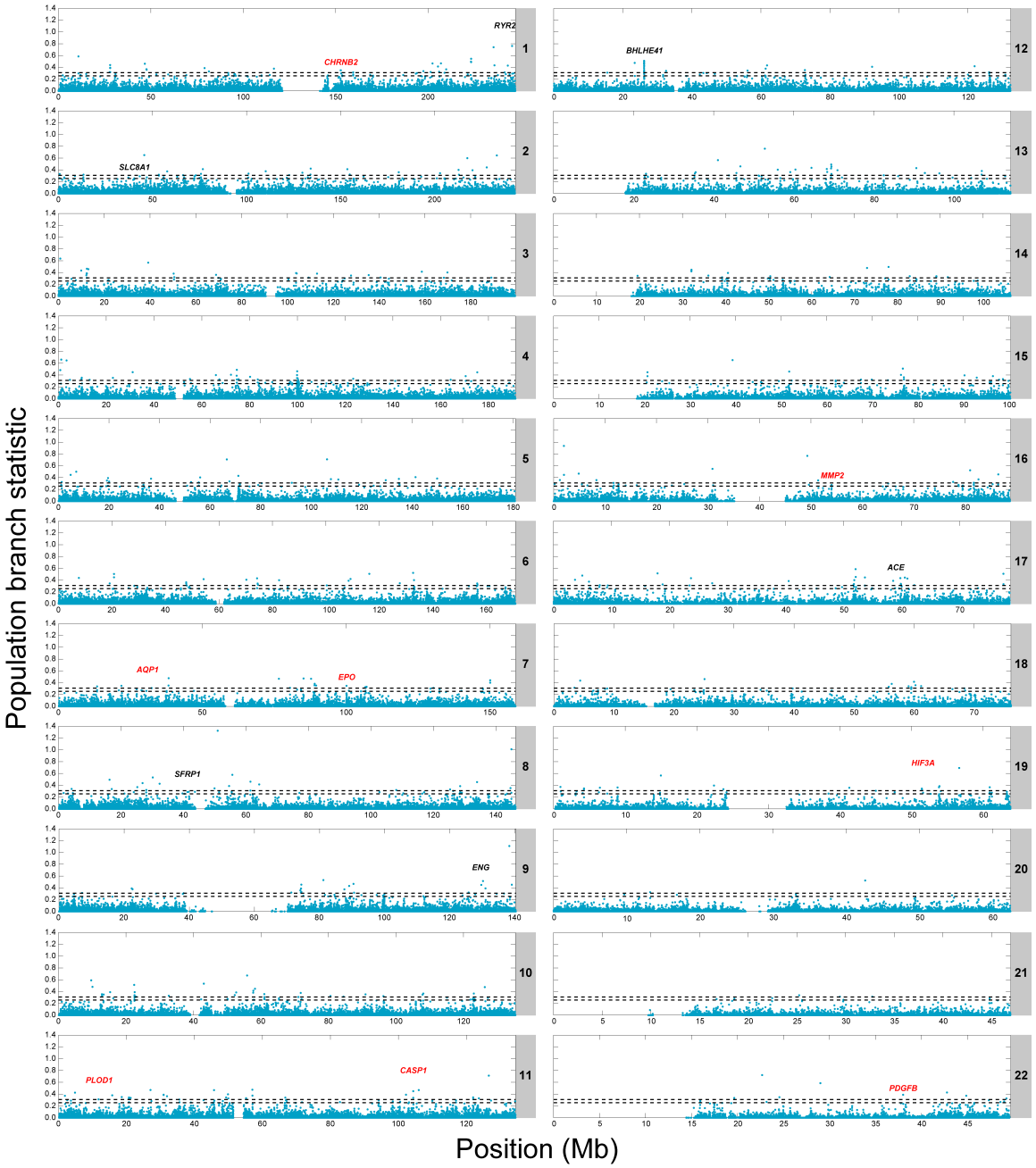

Figure S8: Per-SNP population branch statistic. Results for the analyses of the Tigray, Amhara and Oromo populations combined. Gaps represent regions of the genome that were not covered. Names of genes in black contain at least one SNP in the top 0.10% of SNPs that is located inside the gene. Genes in red contain at least one SNP in the top 0.10% that is located within 50kb of the gene but not within the gene. The numbers in the shaded gray area are the chromosome number. The top and bottom horizontal dotted lines are the 0.05% and 0.10% empirical cutoffs, respectively.

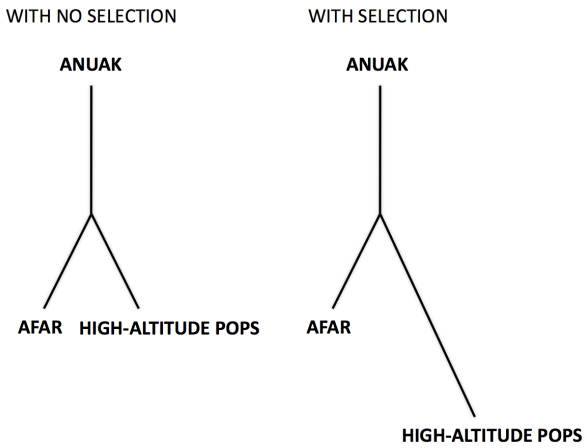

Figure S9: Graphical depiction of the type of genetic signal that is picked up by the population branch statistic as applied to a triple of populations that include the Anuak (the outgroup), Afar (the low-altitude population), and a high-altitude population.

Appendix 4 “Search for a set of 22 markers that would capture the main patterns of genetic diversity among Ethiopian populations”

Search for a set of 22 markers that would capture the main patterns of genetic diversity among Ethiopian populations

Luca Pagani

A4.1 Rationale

The results reported in Chapter 2 and Pagani et al. (2012) were based on the analysis of approximately one million SNPs that were genotyped with Illumina Omni1-Quad arrays in twelve Ethiopian populations. This strategy and the number of sampled populations was sufficient for the purpose of describing genetic diversity of Ethiopians in the context of similar data from other African and non-African populations. However, given the fact that more than eighty ethnic groups live in Ethiopia today, further attempts to describe the Ethiopian population structure at finer geographic and ethno-linguistic levels will require sampling and genotyping of thousands of individuals. In order to assess whether or not the general patterns of inter-population differences that were recovered from high-density SNP-chip data could be captured also by approaches that focus on genotyping small numbers of SNPs in large numbers of samples, a project was designed to summarize the observed Ethiopian genetic diversity using a small subset of the markers. The number of markers to be used was set to 22, one per autosome, in order to ensure full independence between them and to fit the capacity of PCR- or Sequenom-based genotyping facilities. If successful, this approach could be applied in local Ethiopian labs to genotype a broader set of samples in multiple populations.

A4.2 Study design

The objective was to find 22 markers (one per each autosome) so that:

1. The pairwise F_{ST} calculated using only those 22 markers on a set of Ethiopian populations has the best correlation (r) with the F_{ST} calculated with a much broader, genome-wide, set of markers.
2. The first two principal components obtained from the two above F_{ST} matrices are highly correlated.

The available Illumina Omni 1M data for Ethiopian populations were divided into two sets of individuals, "*Case-set*" to be used to search for the optimized SNP set and the "*Control-set*" to independently test the goodness of fit of each set of chosen SNPs. After performing a PCA on the full dataset, the two *Case* and *Control* sets were defined by taking an equal number of

populations from each formed cluster, to create two homogeneous groups. Since this study was carried out before the sample swap described in paragraph 2.3 was identified, the Ari Blacksmith and Gurage populations used were included with wrong ethnic labels. However, since the aim was to replicate patterns of observed diversity in the data, the sample swap does not affect the validity of the approach.

Case populations (20 individuals each): Agew; Amhara; Anuak; Ari Cultivators; Somali and Tygray.
Control populations (20 individuals each): AmharaHA; Ari Blacksmith; Ethiopian Somali; Gumuz; Gurage; Oromo.

The output of the F_{ST} values and PCA to be fitted by each of the optimized marker sets of 22 SNPs was generated using a filtered set of ~480K autosomal markers that had a minor allele frequency of at least 0.1 in each *Case* population and where the proportion of missing data was less than 50% in every population. These markers were then pruned down to ~80K (*Total*) removing those with pairwise LD (r^2) greater than 0.1.

The goodness of fit was measured as:

- Correlation (r) between pairwise F_{ST} measured on the *Total* set and on 22 markers for: *Case*, *Control* and *Case*-and-*Control* populations together (12 pops, "*All*").

-r between the first two pairwise principal components computed for the *Total* and each 22 markers set.

A4.3 Optimization approaches

Approach 1: For each pair of *Case* populations the average F_{ST} (avF_{ST}) per each chromosome was computed using the *Total* markers. For each chromosome, the SNPs within a defined range around the avF_{ST} (see below) were selected. The initial range (1% of the F_{ST} distribution on either side of the avF_{ST}) was then increased by 1% each time until, for each chromosome, at least one SNP met the inclusion criteria in each pair of *Case* populations.

When at least 1 SNP met the inclusion criteria in each chromosome, the maximum range that was obtained spanned as much as 26% of the F_{ST} distribution. This set of 22 SNPs was called *"Overlap".*

Approach 2: In each chromosome the SNP with the highest average F_{ST} across all the pairs was selected. This set of 22 SNPs was called *"Highest".*

Approach 3: In each chromosome the average between all the pairwise F_{ST} values was calculated (totF_{ST}). Then per each chromosome the SNP with the average pairwise F_{ST} closest to the totF_{ST} was selected. This set of 22 SNPs (in any chromosome no tie match was observed) was called *"22Global".*

Approach 4: a. In each chromosome all the SNPs with an average pairwise F_{ST} within 10^{-4} units from the totF_{ST} were selected. These SNPs, altogether, formed a group of 167 SNPs, hereafter referred as *"167".* Note that *167* by definition includes *22Global.*

b. A set of lists containing markers from the ones available in the *167* set were generated. These lists were compiled by keeping, for each chromosome, only one of the available SNPs and including all the SNPs for the other chromosomes. Therefore, if, for example, the *167* set contained 3 SNPs for chromosome 1, three lists were generated including each time only one of the three chr1 markers and keeping the rest intact. These three lists formed the "chr1" set, which was analysed further as described in point c. The correlation (r) between the results generated for the *Case* set from each of those lists and either the *Total* or the *167* sets were then taken into account.

c. The procedure described in point b yielded, for each chromosome, a set of lists each containing only one SNP for that given chromosome. Within each chromosome set, the list with the highest r values was chosen and the single SNP included for that chromosome selected. The 22 SNPs hence selected formed the *"22from167"* set.

d. The *22Global* and *22from167* sets were then merged and further optimized following the 4.b and 4.c steps to create the *"optiBob" (Bob=best of both)* set.

A4.4 Results

Each of the above sets of SNPs was used to calculate the pairwise F_{ST} between the *Case* populations, between the *Control* populations and among the 12 populations merged together (*All*). These results and their PCAs, were compared with those obtained from the same sets of populations, genotyped for the *Total* SNPs. A summary of the results is shown in Table A4.1.

Table A4.1. Correlation coefficients of the F_{ST} and PCA results obtained with each set and the Total set.

	r with *Total* in *Case*	r in *Control*	r in *Case+Control*	r with PC1 of *Total* in *Case+Control*	r with PC2s
Overlap	0.91	0.61	0.65	0.85	0.41
Highest	0.71	0.79	0.56	0.85	-0.27
22Global	0.85	0.86	0.87	0.99	0.70
167	0.90	0.87	N/A	N/A	N/A
22from167	0.75	0.53	0.67	N/A	N/A
optiBob	0.90	0.85	0.82	0.96	-0.76

A4.5 Discussion

Most of the sets performed well in at least one comparison with the *Total* set. However, of the sets including just 22 markers ("167" included more), only *22Global* and *optiBob* performed equally well on the *Case* and *Control* sets. As a further confirmation, these were the only two tests were both PC1 and PC2 r were greater than or equal to 0.7. It is important to note that, since the sign of the principal components is assigned independently across multiple experiments, negative r values should be considered as equally significant.

To test whether a set of markers that have been optimized to represent Ethiopian diversity can equally recapitulate worldwide diversity, the *22Global* and *optiBob* were deployed on a panel of 59 worldwide populations of the HGDP (Li et al. 2008) and current study. Furthermore, the same two optimization approaches were performed also on the worldwide panel hence creating the *22Global_world* and an *optiBob_world* sets. These two sets were in turn tested on the Ethiopian samples. In addition, 1000 sets of 22 random markers (taking one from each autosome) were generated to compare the performance of the sets of choice with the top 2.5% (the top side of a two tailed 5% significance threshold) of the r values obtained using the randomly chosen sets. Table A4.2 shows the correlation coefficients of the two pairs of sets on both worldwide and Ethiopian populations. All the sets listed in this table performed better than the top 2.5% of the random ones in at least one instance. Remarkably the sets optimized on the Ethiopians were overall better than the random sets when applied on the worldwide populations and *vice versa*. Particularly, the *optiBob_world* set performed better than the random sets in all the Ethiopian F_{ST} comparisons as well as on the first PC. However the sets optimized for the worldwide populations outperformed the ones designed for the Ethiopian populations when

applied to the world panel of populations, while the opposite is true when applied on the Ethiopian populations. Therefore it can be concluded that while the proposed approach managed to find two alternative sets capable of retrieving between 80% and 90% of the information contained in the total of 80K markers, the best results are only obtained when such sets are applied either to the local Ethiopian scale or to the broader worldwide scenario, but the same set cannot perform equally well in both cases. The SNPs included in the four sets tested in Table A4.2 are listed in Table A4.3 and were made available to Dr. Neil Bradman to be shared within the framework of his collaboration with other Ethiopian researchers.

Table A4.2 Correlation coefficients (r) between the *Total* set of SNPs and the two pairs of *22Global* and optiBob optimized on Ethiopian and worldwide populations. Values in bold show the best set for each column; values in *italics* show entries above the top 2.5% of the 1000 randomized sets.

	World *Case* F_{ST} (r)	**World *Control* F_{ST} (r)**	**World *Case+Control* (r)**	**PC1 *Case+Control* (r)**	**PC2 (r)**
***22Global*_world**	*0.75*	*0.83*	*0.80*	*0.88*	0.55
22Global	0.72	0.55	0.61	0.68	0.71
optiBob_world	**0.86**	**0.83**	**0.84**	**0.92**	-0.70
optiBob	*0.77*	0.68	0.71	0.76	**0.74**
1000 Random Sets (top 2.5%)	0.73	0.80	0.79	0.85	0.84
	Ethiopia *Case* F_{ST} (r)	**Ethiopia *Control* F_{ST} (r)**	**Ethiopia *Case+Control* (r)**	**PC1 *Case+Control* (r)**	**PC2_corr (r)**
***22Global*_world**	0.34	*0.77*	*0.66*	0.78	-0.13
22Global	*0.85*	**0.86**	**0.87**	**0.99**	*0.70*
optiBob_world	*0.70*	*0.86*	*0.74*	*0.86*	-0.28
optiBob	**0.90**	*0.85*	*0.82*	*0.96*	**-0.76**
1000 Random Sets (top 2.5%)	0.48	0.65	0.63	0.80	0.46

Table A4.3. 22 Autosomal SNPs that make up the sets described in Table A4.2. Note that although the optimization processes are independent, there is some overlap between the two Ethiopian and the two World sets. The SNPs that are shared by each pair of sets are reported in *italics.*

Chromosome	***22Global*_world**	**optiBob_world**	***22Global***	**optiBob**
1	rs12090480	rs2777840	*rs912772*	*rs912772*
2	*rs2894450*	*rs2894450*	rs4399763	rs13023260
3	*rs262215*	*rs262215*	rs6440788	rs9289788
4	*rs7691920*	*rs7691920*	*rs902665*	*rs902665*
5	*rs13185269*	*rs13185269*	rs3111632	rs10515738
6	*rs10498870*	*rs10498870*	rs9391920	rs1359227
7	*rs6959814*	*rs6959814*	*rs2113849*	*rs2113849*
8	rs7823221	rs4385459	*rs1995957*	*rs1995957*
9	rs2039184	rs4877761	*rs3904577*	*rs3904577*
10	*rs11251021*	*rs11251021*	*rs545874*	*rs545874*
11	*rs676761*	*rs676761*	*rs6590735*	*rs6590735*
12	rs618859	rs1123425	rs7314138	rs12305014
13	*rs693039*	*rs693039*	rs1359480	rs2181502
14	*rs328239*	*rs328239*	rs3783785	rs9788478
15	rs1872826	rs1865997	rs2958268	rs12591326
16	*rs2170828*	*rs2170828*	rs8058168	rs2334207
17	*rs4796573*	*rs4796573*	rs2033719	rs8067370
18	*rs1346233*	*rs1346233*	rs1372194	rs8087538
19	rs1058205	rs4807016	*rs1560092*	*rs1560092*
20	*rs12624843*	*rs12624843*	*rs4813044*	*rs4813044*
21	*rs2837269*	*rs2837269*	*rs381716*	*rs381716*
22	*rs1297376*	*rs1297376*	rs4436	rs2006866

Appendix 5 “Fieldwork documents and blank forms”

A5.1 Template of form used to collect phenotypes and donor's sociological information.

STATION 1	
WHAT	**RESULTS**
Collect SIGNED paper	Yes/No
Location of test	...
Gender	Male/Female
Occupation	...

If donor has problems with the sociological questionnaire, please read it at this stage

Tests to be performed at Rest

Skin Colorimeter

Left Arm	E: ...	M: ...
Right Arm	E: ...	M: ...
Cheek	E: ...	M: ...

Blood Pressure

Systolic	...
Diastolic	...
Pulse (Heartbeat)	...

Spirobank!!!*CREATE A NEW USER***!!!**

SpO2	...
Pulse (Heartbeat)	...
FVC	Just scroll Output to check for test quality

ASK DONOR A SAMPLE OF URINE, DELIVER IT TO STATION 3

GIVE THE DONOR A STOP WATCH AND ASK TO COME BACK AFTER 6 MINS

ASK NEXT DONOR TO COME AND WATCH

Tests to be performed soon after the 6 min walk

Meanwhile setup the "post" option on

Spirobank

Spirobank

SpO2	...
Pulse (Heartbeat)	...

Blood Pressure

Systolic	...
Diastolic	...
Pulse (Heartbeat)	...
FVC	check quality

STATION 2	
WHAT	**RESULTS**
Waist circumference (use tape meter)	...
Hips circumference (use tape meter)	...
Height (use rigid meter)	...
Weight (use scale)	...

From Bodystat 1500MDD Bio-Impedance

While Donor put his/her shoes back on, the writer can annotate the following.	**Measure and (Range)**
FAT (%)	..... (..... - .....)
FAT (Kg)	..... (..... - .....)
LEAN	..... (..... - .....)
TOTAL WEIGHT	..... (..... - .....)
DRY LEAN WEIGHT	..... (..... - .....)
WATER (%)	..... (..... - .....)
WATER (Litres)	..... (..... - .....)
BASAL MET. RATE	..... (..... - .....)

BMR/Body Weight	..... (..... - .....)
EST. AVERAGE REQ	..... (..... - .....)
BMI	..... (..... - .....)
BFMI	..... (..... - .....)
FFMI	..... (..... - .....)
Waist/Hip and (High Risk)	..... (.....)
WELLNESS MARKER	.....
IMPEDANCE 5 kHz	.....
IMPEDANCE 50 kHz	.....
Resistance 50kHz	.....
Reactance 50 kHz	.....
PHASE ANGLE 50K	.....

Mid-upper arm circumf. (use tape meter, mark the middle)

LEFT arm	...

Skin Folds and circumferences

RIGHT arm **CIRCUMFERENCE**	...
RIGHT Bicep	...
RIGHT Tricep	...
RIGHT Subscapular	...
RIGHT Suprailiac	...
LEFT arm **CIRCUMFERENCE**	...
LEFT Bicep	...
LEFT Tricep	...
LEFT Subscapular	...
LEFT Suprailiac	...

Cranial measurements

Head length (glabella-inion)	...
Head width (distance between parietal eminences)	...
Face length (nasion-gnathion)	...

Face width (distance between zygomatic arches)	...

SOCIOLOGICAL QUESTIONNAIRE

QUESTION	ANSWER
Important diseases from which the donor is suffering or has suffered:	
Malaria	Yes / No
Cancer (type)	Yes / No
Cardio-vascular / Heart problems (including hypertension)	Yes / No
Diabetes (state Type 1 or Type 2)	Yes / No
Parasitic diseases (state which)	Yes / No
TB (Tuberculosis)	Yes / No
HIV/Aids	Yes / No
Kidney disease	Yes / No
Other major diseases	...

Current medication

State medication (if possible donor to bring packages and interviewer to record details of drug and dose)	...

Lifestyle

Smoke (if yes, how many)	No / Yes: n...
Alcoholic Drink (every day, sometimes, never)	...
Chew gatt (every day, sometimes, never)	...

Personal information

DONOR

Age / Year of Birth	...
Place of Birth	...
Ethnic identity	...

Occupational group (example: Pottery, Tannery, IronSmith, Goldsmith, Weavers, Agriculturalist, Agropastoralist, Pastoralist, Other)	...
Religion	...
First language	...
Second language	...

PARENTS	**Father**	**Mother**
Age / Year of Birth	...	...
Place of Birth	...	...
Ethnic identity	...	...
Occupational group (example: Pottery, Tannery, IronSmith, Goldsmith, Weavers, Agriculturalist, Agropastoralist, Pastoralist, Other)	...	...
Religion	...	...
First language	...	...
Second language	...	...

PATERNAL Grandparents	**Grandfather**	**Grandmother**
Age / Year of Birth	...	...
Place of Birth	...	...
Ethnic identity	...	...
Occupational group (example: Pottery, Tannery, IronSmith, Goldsmith, Weavers, Agriculturalist, Agropastoralist, Pastoralist, Other)	...	...
Religion	...	...
First language	...	...
Second language	...	...

MATERNAL Grandparents	**Grandfather**	**Grandmother**

Age / Year of Birth	...	...
Place of Birth	...	...
Ethnic identity	...	...
Occupational group (example: Pottery, Tannery, IronSmith, Goldsmith, Weavers, Agriculturalist, Agropastoralist, Pastoralist, Other)	...	...
Religion	...	...
First language	...	...
Second language	...	...

A5.2 Protocol to extract DNA from blood samples

DNA EXTRACTION “To Do” LIST

1. Take 10 blood samples from the freezer. If there are samples in the fridge, give them priority.
2. Allow some time for the samples to thaw or to stabilize at room temperature
3. Place the samples in the centrifuge and make sure they are BALANCED!
4. Run the centrifuge with the following settings:
    a. Temperature: 25 C;
    b. Time: 20 minutes;
    c. Rotor: 12157 (if working on Addis Ababa big centrifuge);
    d. Speed: 1200xg (do not set min^{-1}!);
    e. Program: --

WHILE THE CENTRIFUGE IS RUNNING:

1. Switch on the water bath at 65 C;
2. Start a new page on the black lab-book: write down the Extraction date and a progressive number (extraction code). Write down the sample codes one per each line.
3. Label 10 criovials and lids (or eppendorf 2ml tubes) as follows:
   Collection place (example: Jimma);
   Sample code (example: AV);
   Plasma
4. Label 10 15ml falcon tubes and lids as follows:
   Collection place (example: Jimma);
   Sample code (example: AV);
   DNA
5. Clean the hood using Ethanol and kitchen paper;
6. Clean the P1000 pipette as above;
7. Make sure you have at least 40ml of Ethanol 70% and 40ml of Isopropanol 100%.
8. Otherwise use the automatic pipette to refill your stock. Be careful with the Ethanol calculations!
9. If you are going to extract 10 samples, make sure of having at least:
   100 ml of FG1;
   20 ml of FG2;

4 ml of FG3;

120ul of Protease.

Otherwise open a new kit.

MEANWHILE, IF THE CENTRIFUGE HAS FINISHED ARE YOU ARE NOT READY YET, CENTRIFUGE 5 MINUTES MORE BEFORE PERFORMING EXTRACTION.

10. Perform the Plasma and DNA extraction (see specific protocol)
11. Prepare the gel (see specific protocol).

GEL PREPARATION PROTOCOL

1. Make a 1x TBE solution from the 10x mother stock:
2. Take 15 ml of 10x TBE and put into a Becker;
3. Add 135 ml of dH_2O;
4. Total volume is 150ml.
5. Weight 1,5 g of Agarose and add to the Becker;
6. Add spinning magnet to the Becker;
7. Place Becker to the hob and make it boiling until the solution is transparent and homogenous.
8. Bring the HOT Becker to the gel area.
9. Let it cool down to a temperature that allows you to hold the Becker in your hands
10. Meanwhile prepare the gel frame and comb
11. Add 7,5ul of EtBr (2,5ml per each 50ul of solution) and stir the gel moving the Becker.
12. Carefully pour the get into the frame and let it polymerize for at least 1 hour.
13. Once the gel is ready, make tiny blue-dye drops on a piece of parafilm
14. Take 5ul of each sample and add to each drop
15. Load the samples into the gel. DNA will migrate toward the positive pole, so make sure you are placing the gel in the correct orientation
16. Run for 20 minutes at 130 Volts.
17. Place the gel on the gel reader, switch on the appropriate software and take a picture of the transilluminated gel.
18. Save the picture according to the code of the extraction.

DNA Extraction Protocol for 10 samples of 4ml each using Qiagene DNA extraction

1. Prepare: a liquid waste bottle, a propanol waste Falcon, a contaminated solid waste Becker, a clean waste bin (on the floor).
2. Take twice 1000 ul of plasma and transfer it to the criovial.
3. With the same pipette tip take 4 times 1000 ul of buffy coat and annexed fluids and transfer it to the 15ml tube.
4. Close the blood tube, close the criovial, and close the 15ml tube.
5. Perform the above for all the 10 samples.
6. Transfer the criovials and the leftover blood tubes to the freezer.
7. Prepare the FG2/Protease stock solution by mixing 20 ml (with the automatic pipette) of FG2 and 200 ul (with P1000) of Protease. Don't prepare this solution until you are sure that you are processing the samples shortly (example: don't prepare it if there is power failure!). Put the protease back into the fridge.
8. With the automatic pipette transfer 10ml of FG1 into each 15ml tube
9. Invert the tubes 5 times
10. Centrifuge 5 minutes at 2000xg (25 C, rotor 12157, program --)
11. Discard the supernatant (into liquid waste) but keep the pellet and the loose things into the tube!
12. Add 2ml of FG2/Protease to the first tube and vortex the tube until you see no precipitate in it. Be careful and don't cheat, since the precipitate might be barely visible!
13. Perform the above for each sample, adding FG2/Protease and vortexing straight after.
14. Place the 10 samples into a glove or some sort of bag and put the bag into the water bath for 10 minutes.
15. Add twice 1000ul of isopropanol (total = 2 ml) to each tube and invert many times until you see "the DNA" as threads or clumps. If you see little DNA in one of the tubes, make a note on the lab-book.
16. Centrifuge 5 minutes at 2000xg (25 C, rotor 12157, program --)
17. Make sure the pellets are all formed. If one pellet is still floating attached to some bubble, try to make it sink and centrifuge again.
18. Discard the supernatant into the "Propanol waste" 50ml tube
19. Let the sample dry face-down onto a piece of kitchen paper.
20. Add 2ml of 70% Ethanol and vortex making sure that the pellets detached from the bottom of the tube.
21. Centrifuge 5 minutes at 2000xg (25 C, rotor 12157, program --)

22. Discard the supernatant (liquid waste) and face down the tubes onto a piece of kitchen paper
23. Leave the tubes to dry for at least 5 min then put them back in the rack.
24. Make sure that the pellets are completely dry and no Ethanol drops are visible on the tube walls. Don’t cheat! It’s important!
25. Add 400ul of FG3 to each sample and put all the samples in the water bath for 1 hour.
26. Load 5 ul of each sample on the gel and store the rest into the fridge.

A5.3 Donors information sheet and consent form

Evaluation of possible differences in genetically determined causation of disease and response to treatment in the peoples of Ethiopia

We would like to invite you to participate in a study carried out at Addis Ababa University and with overseas collaborators. Before you decide to give your permission it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

Genetics, causes of disease and treatment

An individual's genetic make-up (the genes inherited from the person's mother and father) combined with outside effects (for example parasites and unhealthy activities such as smoking) can cause disease. They can also affect how individuals respond to treatment.

The aim of the study:

Our aim is to investigate how variable are the genes in the peoples of Ethiopia that might contribute to the causes and efficacy of treatment of diseases and the light genetic variation may throw on course and processes involved in human evolution..

Why is the study being done?

There is very little information currently available about genetic variation in the peoples of Ethiopia. Since knowledge of variation can help to understand the causes of disease and the most effective way to treat it we are undertaking this study to increase knowledge of variation and to see how the extent and type of variation compares with other groups throughout the world. In addition since variability in drug response means that all medicines do not work equally well for all patients we want to understand how knowledge of genetic variation amongst the peoples of Ethiopia might improve healthcare. Many previous studies elsewhere have observed that disease susceptibility and environmental responses vary by ethnicity. Investigating the occurrence of these genetic changes in different ethnic groups will allow researchers to identify those groups in which a given medication may work best or have unintended adverse consequences. A further objective of the research is to understand better how humans as a whole vary and how over long periods of time humans may have evolved.

Why have I been chosen?

To identify how common a genetic change is within an ethnic group many individuals must be tested. We have therefore selected individuals from ethnic groups living in the north and south of Ethiopia. We have asked you to take part because you have identified yourself as a member of an ethnic group and we wish to include representatives of both large and small groups.

What is involved in the study?

If you agree to take part in this study we will ask you to fill in a consent form and answer a short (10 minute) questionnaire about the birth places, residences, ethnic identities, occupations, spoken languages and religions of yourself, your parents and your grandparents; any diseases from which you have suffered in the past or currently have, any medication you are currently taking and information on whether you drink alcohol, smoke or chew get (and if so in what quantity and with what frequency). Then a doctor,

nurse or phlebotomist will take two blood samples of about 10mls each (about two teaspoons full). These blood samples will be processed in a laboratory and genetic material (DNA) derived from the sample will be stored for future studies. Furthermore the DNA extracted by these blood samples may be exported outside Ethiopia and the sequencing and genotyping data anonymised and stored indefinitely in an electronic archive (European Genome Phenome Archive) at EMBL-EBI, Hinxton, UK and potentially shared with other *bona fide* researchers. The liquid will also be preserved for future study. You will also be asked to provide a sample of urine (10ml) for future analysis and saliva as an additional source from which DNA will be extracted. Provided you consent the following data will be collected from you by non-invasive methods using a number of different instruments designed for the purpose (in all cases this will only involve you taking the actions described and being touched): height, waist, hips, arm, cranial and skin fold measurements, estimates of fat and water, skin colour, lung capacity and blood pressure at rest and following a six minute walk.

If you agree to take part in this study, the donated DNA samples, blood fluid (plasma) and urine may be used for future ethically approved research into genetic variation of relevance to human health.

Possible risks:

There is only a small possibility that you may experience bruising around the area where the blood sample has been taken.

Potential benefits:

There is no direct clinical benefit to you from taking part in this study. However, the information we obtain from this research will provide valuable knowledge to clinicians,

enabling them to understand better genetic contributions to disease and to make more informed decisions when prescribing medications to future patients.

Do I have to take part?

You do not have to take part in this study if you do not want to. If you decide to take part you may withdraw at any time without having to give a reason. Your decision whether to take part or not will not affect the standard of care you receive. If you decide to take part you will be given a copy of this information sheet to keep and be asked to sign a consent form, a copy of which you will also be given to keep.

The information held about the research subject

No information that is collected will be able to be connected to you. The only information that will be permanently linked to your DNA sample and other stored material will be the information you provide on the questionnaire and the data collected by the researchers described above which will not have either your name or details that would enable you to be contacted in the future..

Professor Endashaw Bekele, based at Microbial Cellular and Molecular Biology Program Unit, Addis Ababa University, PO Box 1176, Addis Ababa Telephone 01 23 94 71 will be responsible for the safety and security of any data collected, which will be processed and handled at or as authorised by him or Addis Ababa University. Only persons authorised by Professor Bekele or Addis Ababa University will have access to the data. By signing the consent form you give permission for these persons to access the necessary information.

Gene Studies

Analysis of the DNA and other material may involve any and all methods including sequencing of the entire genome and the information may be used for commercial purposes.

What will happen if the findings affect you personally?

We will not be able to disclose any findings of genetic variation or medical relevance to individual volunteers because the samples will be anonymous.

Who do I speak to if problems arise?

Any complaint about the way you have been dealt with during the study or any possible harm you might suffer will be addressed.

Complaints:

If you have a concern about any aspect of this study, you should ask to speak to the researchers who will do their best to answer your questions. Contact details are provided at the end of this information sheet. If you remain unhappy and wish to complain formally, you can do this through Professor Endashaw Bekele, Microbial Cellular and Molecular Biology Program Unit, Addis Ababa University, PO Box 1176, Addis Ababa Telephone 01 23 94 71All communication will be dealt with in strict confidence.

Harm:

Every care will be taken to ensure your well-being or safety is not compromised during the course of the study. In the event that something goes wrong and you are harmed and this is due to someone's negligence then you may have grounds for a legal action

for compensation against Addis Ababa University but you may have to pay your legal costs.

What will happen to the results of the research study?

We hope to report the results of this study at international meetings and in published, peer reviewed journals. You will not be identified in any report or publication. If you would like to obtain copies of publications resulting from the research, these can be obtained by contacting Professor Endashaw Bekele,

Who is organising and funding the research?

This research is funded by the Melford Charitable Trust. It is being organised by Professor Endashaw Bekele.

Who has reviewed the study?

Contact details

Thank you for taking the time to read this information sheet.

Please note that you will be given a copy of this information sheet and a signed consent form to keep.

A5.4 Donor Consent form

Chief Investigator: Professor Endashaw Bekele

1. I confirm that I have read and understood the information sheet dated XX/XX/XXXX for the above study and have had the opportunity to ask questions. ☐

2. I confirm that I have had sufficient time to consider whether or not I want to be included in the study ☐

3. I understand that my participation is voluntary and that I am free to withdraw at any time before my samples are taken, without giving any reason and without my medical care or legal rights being affected. ☐

4. I understand that the DNA and other samples I donate will be not be linked to me and may be stored for use in future ethically approved research into genetic variation of relevance to human health and the study of human evolution. ☐

5. I understand that samples and data collected during the study may analysed by individuals and organisations outside Ethiopia and may be used for commercial purposes. ☐

6. I agree to take part in the above study. ☐

Signature of donor	**Date**	
Name of person taking consent	**Date**	**Signature**
(if different from researcher)		
Professor Endashaw Bekele	01 23 94 71	
Researcher	**Telephone**	
(to be contacted in case of problems)	**number**	

Comments or concerns during the study

If you have any comments or concerns you may discuss these with the investigator. If you remain unhappy and wish to complain formally, you can do this by writing to Professor Endashaw Bekele, Microbial Cellular and Molecular Biology Program Unit, Addis Ababa University, PO Box 1176, Addis Ababa Telephone 01 23 94 71. All communication will be dealt in strict confidence.

Appendix 6 “SNP calling pipeline (vr-pipe)”

As specified elsewhere in this thesis, the SNP calling process was performed by Dr. John Maslen and Dr. Petr Danecek, starting from the bam files generated by John Maslen at the WTSI. The SNP calling was performed within the framework of a pipeline called “vr-pipe”, developed at Sanger and with source code and detailed instructions available here: https://github.com/VertebrateResequencing/vr-pipe

For the pipeline to work, a configuration file needs to be supplied. Such configuration file contains all the path and information required by the pipeline and, a template of the one used to call the Ethiopian low coverage and high coverage samples is reported below. The information provided in such configuration file, together with the detailed instructions reported on Github is sufficient to re-create the dataset generated for this thesis.

Vr-pipe configuration file:

```
mpileup  => '/software/vertres/bin-external/samtools-0.1.18 mpileup -EDS -C50 -m2 -F0.0005 -d 2000',
bcftools => '/software/vertres/bin-external/bcftools-0.1.18 view -p 0.99 -vcgN',
bams    => 'PATH_TO_BAMS_FOFN',
fa_ref  => '/PATH_TO_REFERENCE/hs37d5.fa',
limits   => { runtime=>24*60 },
do_clean => 0,         # Remove runner's temporary files
ploidy  =>
{
default => 2,
X =>
[
 # These are pseudoautosomal: 60001-2699520, 154931044-155270560, call with ploidy 2
{ region=>'1-60000', M=>1 },
{ region=>'2699521-154931043', M=>1 },
],
Y =>
[
# No chrY in females and one copy in males
```

```
{ region=>'1-59373566', M=>1, F=>0 },
],
MT =>
[
# Haploid MT in males and females
{ region=>'1-16569', M=>1, F=>1 },
],
},

chroms => [ qw(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y) ],
pops   =>
{
# With no populations, at least 'pooled' must be given
pooled => [ '.' ],
},
filter => 'vcf-annotate -f +',
chunk_size        => 1_000_000,
debug_chunks      => 0,
shuffle_chunks    => 0,
keep_bcfs         => 1,
chunks_overlap    => 0,
whole_genome_bams => 1,         # Set to 1 if BAMs are not splitted by chromosome
assumed_sex       => 'F',     # Set to 'F' for females, 'M' males and undef mysql key above if the DB shouldn't be used.
sample_list       =>'PATH TO SAMPLE LIST',     # Provide list of samples with sex.
chunk_options => {
},
```

The End