www.sciencemag.org/cgi/content/full/science.1227721/DC1

Supplementary Materials for

Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History

Carina M. Schlebusch,* Pontus Skoglund, Per Sjödin, Lucie M. Gattepaille, Dena Hernandez, Flora Jay, Sen Li, Michael De Jongh, Andrew Singleton, Michael G. B. Blum, Himla Soodyall, Mattias Jakobsson*

*To whom correspondence should be addressed. E-mail: carina.schlebusch@ebc.uu.se (C.M.S.); mattias.jakobsson@ebc.uu.se (M.J.)

Published 20 September 2012 on *Science* Express
DOI: 10.1126/science.1227721

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S5 to S89
Tables S1 to S25
References (*31-173*)

Contents

List of Figures

List of Tables

Material and Methods

1 Preparation of the SNP data

1.1 Overview

In order to select unrelated individuals to be SNP genotyped, we had previously screened 340 individuals using 15 microsatellites (*31*). First and second degree related individuals were identified and 230 putatively unrelated individuals were selected for SNP genotyping from the initial set. The 230 individuals were genotyped using the Illumina Omni 2.5M SNP chip. Data filtering and merging with other datasets were performed using PLINK v1.07 (*32*) and custom scripts.

1.2 Sample collection

DNA samples from individuals were collected with the subjects' informed consent, and the project was approved by the Human Research Ethics Committee (Medical) at the University of the Witwatersrand, Johannesburg (Protocol Numbers: M980553, M050902, M090576, M10270), the Working Group of Indigenous Minorities in Southern Africa (WIMSA) and the South African San Council (SASC). A description of sample groups, group codes, group membership, number of individuals, place of sampling and origin are outlined in Table S1.

1.3 DNA extraction and WGA

DNA from EDTA-blood was extracted using the salting-out method (*33*) and the PureGene® Genomic DNA Purification Kit (Gentra Systems, Minneapolis, MN) was used to extract DNA from buccal swabs according to the manufacturer's instructions. DNA was quantified using the NanoDrop ND-1000 Spectrophotometer (Coleman Technologies Inc., Orlando, FL, LabVIEW®) and diluted to 10ng/μl with ddH_2O.

Whole Genome Amplification (WGA) was performed with the IllustraGenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Uppsala, Sweden) according to kit instructions, but with modifications to the protocol according to recommendations of Pinard *et al.* (*34*). The method involved adding 22.5μl GenomiPhiSample Buffer to each DNA sample (2.5μl of 10ng/μl DNA) after which the sample was denatured at 95°C for 5 minutes and cooled on ice. Thereafter 27μl GenomiPhi V2 Reaction Buffer and 3μl GenomiPhi Enzyme Mix was added and the sample was incubated for 14 hours at 30°C. After the incubation the enzyme was inactivated by heating the samples to 65°C for 10 minutes. Each 96 well plate of samples was co-amplified with a positive control (GenomiPhi Control DNA) and a negative control (ddH_2O). DNA samples were assayed by UV absorption at 260nm to determine the concentration after amplification.

1.4 Genotyping and initial quality control

In total, 240 samples (230 individuals – 10 individuals were genotyped twice as a technical control) were genotyped per manufacturers protocol on the Illumina (San Diego, CA) Omni 2.5M BeadChip. After initial comparison of sample intensities with the standard Illumina Omni 2.5M cluster file, all samples and loci were re-clustered within GenomeStudio v2011.1 to produce a custom cluster file. The re-defined cluster positions produced 240 samples with an average call rate of 96.31%. Ten individuals were genotyped both based on genomic DNA and whole genome amplified DNA, and showed very high concordance of genotype calls. Subsequent to Illumina QC procedures, 230 individuals and 2,443,179 markers were retained. For the X-chromosome, heterozygous SNPs were converted to missing in males. We subsequently removed SNPs with more than 10% missing data (82,597 SNPs), and all 8 'indels', leading to 2,360,574 SNPs being retained.

1.5 Individuals

After the SNP filtering, six individuals (one Khwe, one Ju/'hoansi and four /Gui and //Gana) with $> 15\%$ missing data were removed. (Note that after removing these six individuals, 455 SNPs had $> 10\%$ missing data, all these 455 SNPs had 10.27% missing data). Furthermore, four pairs of individuals (one pair each in !Xun, Ju/'hoansi, /Gui and //Gana and ≠Khomani (A) groups) were identified as potentially related from the fraction of SNP alleles that were Identical by state (IBS). These individuals had $\hat{\Pi} > 0.3$, $\hat{\Pi} = f(IBS_2) + 0.5f(IBS_1)$, where $f(IBS_2)$ denotes the fraction of SNPs that a particular pair of individuals share both alleles, and $f(IBS_1)$ denotes the fraction of SNPs that the pair of individuals share one allele. Four individuals (one !Xun, one Ju/'hoansi, one /Gui and //Gana and one ≠Khomani (A) individual) were selected to be removed at a later stage from the dataset, and they were retained for the phasing (see below) as the accuracy of phasing is likely to improve with larger sample size and related individuals (*35*). The four individuals to be removed were selected on the basis of the individual with greatest fraction missing data (for each pair).

1.6 SNPs

Of the 2,360,574 SNPs, 2,299,431 SNPs were located on the autosomes, 53,082 on the X-chromosome, 1,300 SNPs on the Y-chromosome, 489 SNPs on the pseudoautosomal region on the X- and Y-chromosomes, 93 SNPs on the mitochondrial chromosome, and 6,179 SNPs with unknown chromosome location.

In order to determine which autosomal SNPs deviate from Hardy-Weinberg equilibrium (HWE), the individuals were split into one Khoe-San (/Gui and //Gana, Ju/'hoansi, !Xun, Khwe, Karretjie, ≠Khomani (A), Nama) and one Bantu-speaking group (Bantu-speakers (South Africa) and Herero). SNPs that significantly deviate from HWE ($p < 0.05$ Fisher's exact test) in both groups were excluded from further analyses (5,377 SNPs).

Furthermore SNPs with a basepair position of 0 (unknown position) were removed (734 autosomal SNPs) and 2,293,320 autosomal SNPs were retained. For the X chromosome 44 SNPs with a basepair position of 0 was removed leaving 53,038 SNPs.

Further processing of the data focus on the autosomes and the X-chromosome, in total 2,346,358 SNPs.

1.7 Missing data

For the 2,346,358 autosomal and X-chromosomal SNPs, the total missing data was 1.67%. Most individuals and SNPs had a very low fraction missing genotypes, the median fraction of missing data across individuals was 0.59% and the median fraction of missing data across SNPs was 0.89%.

1.8 Including SNPs from the genome sequence of KB1

SNPs from the genome sequence of the reference San individual KB1 (*9*) were extracted and merged to our dataset. First, we used the software `liftOver` (`http://hgdownload.cse.ucsc.edu/admin/exe/`) to transfer hg18 mapping positions to hg19 positions (the 2.5M array has hg19 mapping positions). Second, we downloaded the `.bam` alignment files of 454 and Illumina sequence data from KB1 which has been aligned to the hg18 human reference genome (*9*) [`ftp://ftp.bx.psu.edu/data/bushman/hg18/bam/` downloaded February 2011]. Third, we extracted 2,067,433 SNPs (2,029,723 autosomal SNPs and 37,710 X-chromosomal SNPs) that overlapped with the SNPs on the 2.5M Illumina array from the KB1 individual. Diploid genotypes were called by requiring reads to have mapping quality > 10, positions to have coverage between 6 and 75, and `samtools` (*36*), pileup version, consensus quality > 30 (using the `samtools`' method for genotyping). Heterozygosity for KB1 (based on autosomal SNPs) was 0.154 (313,814 heterozygous SNPs), which was similar to the heterozygosity for the individuals genotyped using the 2.5M SNP array.

There are three strand-related issues that can arise when merging data from two (or more) different sources: i) unknown strand for GC- and AT-SNPs, ii) more than two variants are found for non GC- and AT-SNPs, and iii) monomorphic SNPs for different alleles in the two datasets. We solved these issues by *i*) setting all GC- AT-SNPs in KB1 to 'missing data', *ii*) flipping the stand of KB1 when needed, and *iii*) keeping track of the two variants in the strand orientation of our genotype data for each SNP (despite that all individuals may be monomorphic) and thereby knowing if the strand of KB1 needs to be flipped or not (in the case of GC- and AT SNPs, the KB1 variant was set to 'missing data'). After AT- and CG-SNPs were filtered out 2,005,720 SNPs remained for KB1 and these SNPs were merged to the primary dataset. All SNPs (723 cases) that did not match (after potentially flipping the strand) were set to 'missing data' in KB1, leaving 2,004,997 SNPs from KB1 (14.6% missing data).

1.9 Phasing

Haplotypes and missing genotypes were estimated for 225 samples and 2,348,263 SNPs (both autosomes and X-chromosome) using `fastPHASE` (*35*) version 1.4.5. The number of haplotype clusters was set to 25, and we used 25 runs of the EM algorithm. This analysis was used to generate a "best guess" estimate of the true underlying patterns of haplotype structure.

We included all 225 individuals during haplotype estimation (220 unrelated individuals, 4 individuals that were related to one individual each among the 220, and the reference individual KB1). Haplotype phase was estimated for all autosomes and for the X chromosome, the haplotype estimation procedure treated males as having a known haplotype. We removed relatives from the phased haplotype data to create a dataset of 221 unrelated individuals (220 individuals genotyped in this paper and the reference individual KB1).

1.10 Adding the chimpanzee variant and the 'Human Ancestor'

For the autosomes (2,293,320 SNPs), we extracted the variant from a chimpanzee provided by ref. (*9*) and the variant from the 'Human Ancestor' sequence (*37–39*). For the chimpanzee, 1,479,821 SNPs were overlapping, and for the 'Human Ancestor' 2,091,564 SNPs were overlapping. After removal of GC- and AT-SNPs, 1,435,094 SNPs (chimpanzee) and 2,025,519 SNPs ('Human Ancestor') remained and were merged with the primary data.

1.11 Summary of the dense SNP set typed for 220 Southern African individuals

In summary, the cleaned data contained 2,293,320 high-quality autosomal SNPs from 220 Southern African individuals. Additionally, for most SNPs, the data included the genotypes from the genome sequence of the San reference individual KB1, the variant of the chimpanzee, and the variant of the 'Human Ancestor' sequence. The cleaned X-chromosome data contained 53,038 high-quality SNPs from 220 Southern African individuals. Two versions of the data were prepared, the genotype data (with missing data) and phased data (missing genotypes were imputed except for the chimpanzee and the 'Human ancestor'). We will refer to this set of individuals and SNPs as the 'Southern Africa dataset'.

1.12 Merging with HGDP, HapMap, and Henn *et al.* (2011) (*10*)

We obtained phased genotype data from HapMap III (*40*) for 562 individuals from 9 populations: ASW, CEU, TSI, GIH, JPT, CHB, MKK, LWK, YRI, downloaded Nov 30, 2010 (`ftp://ftp.ncbi.nlm.nih.gov/hapmap/phasing/2009-02_phaseIII/HapMap3_r2/`). In total 45 individuals from the HapMap dataset were removed due to first and second degree relationships (*41*). For the HapMap III data, 1,437,973 SNPs were available (1,437,242 non AT-GC SNPs), and 509,899 (non AT-GC) SNPs overlapped with our set of SNPs.

We obtained phased genotype data from the Human Genome Diversity Project for 938 unrelated individuals (from 53 populations) [HGDP; `http://hgdp.uchicago.edu/Phased_data/`] (*4, 42–44*). For the HGDP data, 656,995 SNPs were available (all non AT-GC SNPs) and 354,223 SNPs overlapped with our set of SNPs.

We also obtained phased genotype data from Henn *et al.* (*10*) for 76 unrelated individuals from three sub-Saharan African populations (Hadza, Sandawe and ≠Khomani (B)). For the Henn *et al.* (*10*) data, 481,048 SNPs were available (480,943 non AT-GC SNPs), and 273,215 overlapped with our dataset.

These three datasets, HGDP, HapMap III, and Henn *et al.*, were merged with our Southern Africa dataset. Prior to merging the data, we removed AT- and CG-SNPs from the HapMap III data and the Henn *et al.* data (the HGDP data had no AT- and CG-SNPs). The three 'external' datasets were merged to our Southern Africa dataset, one after the other, with the same procedure as described previously for KB1 (note that only matching SNPs remained after flipping the strand of 54,851 SNPs). All SNPs that did not overlap between the datasets were removed yielding a dataset with no missing data (except for the 'Human Ancestor' and the chimpanzee). After the merger of the four datasets, we found evidence of related individuals between our Southern Africa dataset, and the HGDP, and between our Southern Africa dataset and the Henn *et al.* data. Nine pairs of individuals showed a high fraction of alleles being 'identical-by-state' (IBS) and when plotting the fraction of SNPs with two alleles shared versus the fraction of SNPs with zero alleles shared, the relationships for these 9 pairs were revealed: Between the HGDP set of individuals and our set of individuals, four individuals were identical (two in Ju/'hoansi (HGDP) vs. Ju/'hoansi in our data, Figure S1, and two individuals in Bantu-speakers (Southern Africa - HGDP) vs. Herero in our data, Figure S2). Furthermore, one Ju/'hoansi (HGDP) individual was related to two Ju/'hoansi individuals in our data. Between the Henn *et al.* data and our data, one identical individual and three related individuals were identified (between the ≠Khomani populations of the two studies, Figure S3). In total nine individuals were excluded from the HGDP and Henn *et al.* data; three Ju/'hoansi (HGDP), two Bantu-speakers (Southern Africa - HGDP), and four ≠Khomani (B) from the Henn *et al.* data. The final worldwide dataset contained 269,317 phased SNPs from 1,745 individuals, of which 536 were from sub-Saharan Africa. We will refer to this dataset as the 'global dataset'.

1.13 Identification and exclusion of individuals with putative recent admixed ancestry

For both the Southern Africa dataset and the global dataset we inferred and removed putatively recently admixed individuals to create versions of the datasets less affected by recent admixture. In these versions of the datasets, individuals that showed a clear signal of recent admixture were removed from our Southern Africa dataset as well as the Henn *et al.* dataset.

To identify individuals with putative recent admixture from European and Bantu-speaking groups in our Southern Africa data as well as in the Henn *et al.* data, we made a preliminary `ADMIXTURE` analysis for these two datasets together with HapMap Tuscans (TSI), HapMap CEU, and HapMap Luhya (LWK). We ran 20 replicate `ADMIXTURE` (*16*) analyses, assuming $K = 7$ clusters (we also investigated $K = 2$ to $K = 10$, and they all produced similar result for detecting recent admixed individuals). `ADMIXTURE` was run under default settings with random seed generated from the clock time. The replicate runs were combined using `CLUMPP` version 1.1.1 (*45*) using the *LargeKGreedy* algorithm with 10,000 random permutations.

We first identified the top 25% individuals in each predefined sample group that has the lowest combined membership coefficients of the EUR (Eurasian) and BS (Bantu-speaking) cluster. We then used a threshold that was 4.5 standard deviations (computed based on the top 25% individuals in the particular group) from the value of the individual with the lowest combined membership, add a "noise" factor of 0.05, and exclude any individual with a combined EUR+BS membership that exceeds this threshold. We validated the approach by comparing with a PCA where each group of interest was analyzed together with CEU, TSI and LWK. The rank order affinities to the Eurasian and Bantu-speaking clusters of the putatively admixed individuals was similar for the PCA-based and the `ADMIXTURE`-based approaches.

For the Bantu-speaking groups in our Southern African data set (Herero and Bantu-speakers (South Africa)), we also excluded individuals with putative San-related ancestry by performing an identical procedure as above with the combined membership coefficients of the clusters that were modal in Ju/'hoansi and Karretjie, respectively (SSAN+NSAN). One individual from each group was excluded in this step (KSP184 and KSP197). A summary of the excluded individuals are shown in Table S2. In total, 133 recently admixed individuals were removed from the global dataset, and 103 from the Southern Africa dataset. A summary of populations, sample sizes, and geographic location can be found in Table S3.

1.14 Summary of analyzed datasets

Our analyses focus on the phased autosomal data although several analyses do not use the phase information. There are five different versions (subsets of individuals or subsets of SNPs) of the data that we analyze: (a) ∼2.3M SNPs in 221 Southern African individuals (including KB1), (b) ∼2.3M SNPs in 118 Southern African individuals (including KB1 and admixed individuals removed), (c) ∼270k SNPs in 1,743 individuals from a worldwide set, (d) ∼270k SNPs in 1,610 individuals from a worldwide set (admixed individuals removed), and (e) ∼270k SNPs in 403 sub-Saharan individuals (admixed individuals removed). Table S4 provides additional information on the individuals in (a), (b), and (e).

Supplementary Text

2 Population and sample description

This section summarises the study-populations' background and is adapted from (*46*). The term "Khoe-San" has a collective meaning for two groups of people, the Khoi (old Nama word) or Khoe (modern Nama word), who were traditionally the pastoralist groups and the San, which included the hunter-gatherer groups (*47,48*). The groupings conventionally refer to generally different subsistence patterns. It remains a topic of debate as to whether this grouping presents a true reflection of subdivision (*48*). Different San and Khoe groups are distributed throughout Southern Africa where they live among and to some extent are admixed with the various Bantu-speaking populations surrounding them (*48–50*). To classify Khoe-San groups into their individual ethnic groups is, in many ways, problematic. Different words and spellings have been used to refer to the same groups of people over the years. Linguistic classification is the method most commonly used to identify different groups, but it is not clear if linguistic classification reflect genetic relationships.

Linguistic studies indicate three separate linguistic families for the Southern African Khoisan linguistic division, namely, Ju-≠Hõa (Ju was previously classified as Northern Khoisan), Tuu (previously classified as Southern Khoisan) and Khoe-Kwadi (Khoe was previously classified as Central Khoisan) (Table S5). Two eastern African Khoisan languages also exist, namely Hadza and Sandawe. These linguistic families are either unrelated or have genealogical relationships that extend further back than 10,000 years (*51*). Linguistic evidence supports the possibility that the Ju and Tuu branches may share a very deep common ancestor and were associated with the original San hunter-gatherers, while the Khoe branch was introduced to the area later, possibly in conjunction with pastoralism (*51*). The Khoe-Kwadi group includes Kwadi, the extinct language of Angola (which have been putatively linked to the East African Sandawe language) and the Khoe language branch. The Khoe language branch includes two very divergent subgroups, KhoeKhoe and Kalahari Khoe. The people who practiced pastoralism at the time of European contact, known commonly today as the Khoe, speak languages from the linguistic branch "KhoeKhoe" while certain San groups from Botswana speak languages belonging to the "Kalahari Khoe" branch, they are also known as the Khoe-speaking San groups.

Table S5 shows a simplified version of the linguistic classification of Khoisan (as outlined in (*51*)) and indicate groups represented in the present genetic study. In this study individual groups were referred to by their preferred community names. The application and spelling of the community names are in accordance with the usage in the book "Voices of the San" (*50*); a book compiled by young representatives from San communities. For the sake of clarity the Language Group as given in (*51*) is listed together with the group name used in this study (Table S5). The following section will generally discuss the ethnography of Khoe-San people belonging to the three Southern African Khoisan linguistic groupings and then focus on the specific populations represented in the current study, their sampling location and previous genetic studies on the particular group.

2.1 Ju linguistic division (Northern Khoisan)

The three main ethno-linguistic groups of Ju; the !Xun, the Ju/'hoansi (meaning "real people") and ≠X'ao//'ãesi (also called Auen or linguistic-branch úKx'au//'e) correspond to indigenously defined dialects that also parallel three different cultural units and geographic areas. The word !Xun (or the different spelling !Kung) has been widely used to describe all three of these groups, however, the only group that uses the term as self-identification are the !Xun groups of Angola and northern Namibia (!xu is a word indicating "person" in !Xun languages). The three groups together were estimated to comprise 25 000 to 30 000 individuals in the 1980s (*52, 53*). The largest group is the central Ju/'hoansi while the northern !Xun is distributed over a larger geographic area (*54, 55*).

The !Xun of northern Namibia and southern Angola

The northern !Xun do not live in the Kalahari like the other two Ju groups but rather in the forested areas of southern Angola and northern Namibia. Their self-designation is !o !xu which means "forest people" (*56*). Two groups found in Angola are known locally as Kwankala (Vakwankala) and Sekele (Vasekele) (*57*). In the local Bantu languages these names have derogatory connotations (meaning poor uncivilized wanderers) and are not used anymore. The !Xun lived in close association with the local Ambo (Ovambo), a southwest Bantu-speaking population, for centuries. It is through this association that the !Xun learned crop cultivation, herding and fishing with nets and spears. In the 1950s very few groups still followed a foraging lifestyle supplemented with assisting Bantu-speakers in the winter harvests in exchange for grain. In 1970-1980 Angola was a battleground between the government and guerrillas. Since then, no ethnographic studies have been conducted to assess the extent of damage the war has had on the !Xun way of life (*48, 57*).

Various genetic studies have been conducted on the !Xun, which include mitochondrial DNA studies (*13, 58–60*), Y-chromosome studies (*61–68*) and autosomal DNA studies (*8*). In some of these studies they were referred to as Vasekele, Sekele or Vasekele !Kung and being of South African origin, however, the San groups in these studies were the !Xun with an origin from Angola (as explained in the section below entitled "The !Xun and Khwe of Platfontein"). The !Xun samples from the present study were sampled at the Omega military base in the Caprivi Stip (Namibia) and in Schmidtsdrift (South Africa) (See full explanation of this groups' history in the sections to follow).

The Ju/'hoansi of Namibia and western Botswana

The central Ju/'hoansi groups occupies areas with a large supply of water and plant resources. Bands of people usually camp out near permanent waterholes and Mongogo nut groves. In the past they only camped out during the dry winter and moved away during the wet season to exploit other territories. In Botswana, however, over the past century, groups have increased their time camping out. Today most groups have settled at the waterholes, and depend on Herero and Tswana residents for their livelihood. Development projects, including schools and handicraft tourist shops, were implemented by the Botswana government and anthropologists. In Namibia a "homeland" reserve for the Ju/'hoansi (Bushmanland) was established and a school and administrative camp were built at Tsumkwe. In 1978 the South African Defense Force (SADF) built a military base at Tsumkwe and recruited Ju/'hoansi soldiers. Many families lived off the earnings from the military base. Traditional subsistence techniques started vanishing because of this and the fact that the reserve was too small to support the number of people. Anthropologists were partially successful in encouraging them to adopt cattle husbandry in the reserve but met with opposition from wildlife officials (*48, 69–71*).

The Ju/'hoansi of the present study was sampled in Tsumkwe reserve in Namibia. The Ju/'hoansi was studied in numerous genetic studies in which they are mostly referred to as San or !Kung (*3, 4, 10, 58, 72–74*). This Ju/'hoansi group also constitute the 7 San samples in the HGDP panel (*75*). The San samples in the HGDP panel came from the exact same sampling collection as the bigger sample from the current study (in fact two samples were identical and one related to the present study sample and were excluded from the HGDP sample incorporated in the present study).

2.2 Khoe linguistic division (Central Khoisan)

2.2.1 Kalahari Khoe (Khoe-speaking San groups)

The Khoe-speaking San groups (speaking the Kalahari Khoe language grouping) are the most numerous and culturally diverse of the San language groups (Table 1). They inhabit the central and northern parts of Botswana, including the central Kalahari Desert and Okavango swamps, the southern parts of Angola and the Caprivi Strip of Namibia. Groups included into this language group are the Naro of western Botswana, the /Gui, //Gana and Deti of central Botswana, the “river Bushmen” of northern Botswana and southern Angola (the different Khwe groups), the Tshua and Shua of eastern Botswana and the Tyua of western Zimbabwe (*48–50*).

The Khwe of northern Botswana and southern Angola

The Khoe-speaking San of northern Botswana, southern Angola and western Zimbabwe comprise the various Khwe (linguistic grouping – Kxoe) groups (including the Bugakhwe and //Anikhwe). They live in the Okavango swamp area and surrounding regions. This area is infested by tsetse flies and as a result livestock rearing is not viable. They sustain themselves through fishing as well as hunting and gathering. Linguistically, they are closer to the central Khoe-speaking San than the eastern groups. Phenotypically, however, they resemble Bantu-speakers and genetic evidence also suggests a genetic makeup similar to the Bantu-speaking populations that surround them (*76, 77*). They share their territory with various Bantu-speakers including the Mbukushu (cultivators), the Yei (fishermen) and to a lesser extent the Tswana, Kgalagari and Herero herders. Each group operates in a different ecological niche. The San groups are concentrated on the banks of the Okavango River and the delta area as their informal name ”river Bushmen” implies (*48*). It is not clear whether these northern Khoe speaking San groups are Khoe-San groups with extensive Bantu-speaking admixture, Bantu-speakers that lost their cattle, another pastoralist population closely related to Bantu-speakers who occupied the region before the Bantu expansions or maybe a mixture of various refugee groups driven from the grazing grounds into the Okavango swamps (*77*).

The Khwe and the !Xun from the present study originated from Angola, they (both) were sampled during two independent sampling trips at the Omega military base in the Caprivi strip of the then South West Africa and later in South Africa (Schmidtsdrift); and are currently living in Platfontein (South Africa). Their origin and history is discussed in the section below. They have been studied in various different genetic studies, including mitochondrial DNA studies (*13, 58–60*), Y-chromosome studies (*61–68*) and autosomal studies (*8*). In some studies combined results of !Xun and Khwe groups are presented, this however is not optimal since these groups have very different languages, cultures, and also as apparent from the current study, different genetics.

The !Xun and Khwe of Platfontein (South Africa)

Although not originally from South Africa, the !Xun and Khwe of Platfontein now made South Africa their permanent home. They originally came from Angola and were employed by the South African Defense Force (SADF) before they were relocated to SA. Five hundred veterans of the SADF together with 3500 dependants were relocated in 1990 from Namibia to South Africa (*78*). They currently live in Platfontein, near Kimberley (SA).

The people of Platfontein are two different San groups with separate identities. One third of the people are known as Khwe (also were called Barakwena) and two thirds are !Xun (also were known as Vasekele). They speak different languages and have a different phenotypic appearance. The groups have remained separate and have insisted to be settled in different parts of the camp. The !Xun group retained a much more cohesive nature and cling to their San identity. They have not mixed with outsiders beyond the camp and have retained a much more unified group than their Khwe counterparts. The Khwe have been more ambivalent about their group identity and have established relationships with surrounding South African groups (*78*).

Although the people of Platfontein have separated themselves into these two groups, members within these groups were not individuals that came from the same area or even knew one another. The !Xun came from a wide region in central Angola around Serpa Pinto (currently Menongue) where many of them lived as stock farmers or cultivators alongside Bantu-speaking groups. !Xun men from different regions were recruited into the Portuguese colonial military in the late 1960s. When the Portuguese moved out the !Xun affiliated with a liberation force, FNLA, in the Serpa Pinto region. FNLA had links with the SADF and when FNLA collapsed the !Xun were recruited by the SADF and brought to the Omega military base in the Caprivi strip of the then South West Africa (Namibia) (*71, 78*).

The Khwe on the other hand originally came from south-east Angola where they have lived along the river systems as cultivators and cattle keepers. They also originally came from a widespread region of southeast Angola and were recruited into a different unit by the Portuguese army. When the Portuguese moved out of Angola, the Khwe fled into neighboring countries like southwest Zambia, northwest Botswana and the Caprivi Strip of South West Africa where there were other Khwe people amongst whom many of the Khwe soldiers had kin. From there they were recruited into the SADF (*71, 78*).

Many of the !Xun were later (late 1970s) relocated to the second "Bushman battalion" in Tsumkwe. At Tsumkwe they were meant to join up with the Ju/'hoansi of Nyae Nyae but the Ju/'hoansi saw the !Xun as invaders and they had to be kept in isolated bases in western Bushmanland. Thus, in 1990 a large number of !Xun opted to come to South Africa while many of the Khwe stayed in the Caprivi where they had local contacts (*71, 78*).

Both these groups were relocated in 1990 to the Schmidtsdrift military base in South Africa. The South African government was reluctant to allocate land or commit funds to secure the future of the San groups. The SADF saw these two groups as "former mercenaries who have outlived their usefulness" (*71, 78*). The !Xun and Khwe trust where established in 1993 to look after the interests of the groups. They remained in tented camps near the Schmidtsdrift military base for several years until recently, the new South African government allocated land to them in Platfontein near Kimberley (SA), where they settled (*71, 78*). Both these groups have representatives on the South African San Council (SASC) today.

The /Gui and //Gana of the central Kalahari (Botswana)

The /Gui and //Gana groups lived in an area now occupied by the Central Kalahari Game Reserve (CKGR) in central Botswana. /Gui has no specific meaning other than the reference to the group while //Gana is derived from a word that means "people of the well". The /Gui and //Gana also shared the CKGR territory with the Kgalagari. The Kgalagari are the oldest existing Bantu-speaking tribe in Botswana. //Gana individuals all tend to speak Kgalagari as well as their own language and it is believed by the //Gana themselves that they originated from a intermixing of the /Gui and the Kgalagari. The /Gui occupied the region adjacent to the western CKGR as well as the western part of the CKGR and //Gana the central and eastern part as well as the region adjacent to the eastern CKGR. The CKGR was established in 1961 and extends over 52,600 square kilometers. Only the southern (wooded zone) and central (bushveld) parts have enough vegetation to support human occupation. The central part is good hunting territory. From the 1960s to the 1980s the population in the CKGR declined from 2,000 to approximately 1,000 individuals. The Ghanzi district commissioner George Silberbauer studied the /Gui and //Gana groups extensively and constructed a borehole in the south central parts of the CKGR near the ≠Xade pan. Subsequently ≠Xade became a settlement with permanent occupation which grew from ∼200 in the 1960s to ∼700 in the late 1970s. In the late 1970s the people of ≠Xade were taught subsistent farming practices but with little available water this was not a successful strategy. The introduction of farming led to an increased number of livestock such as horses, donkeys and goats, which put further pressure on water supplies. Hunting on horseback and donkeys also ensued which caused a decline in large game and attracted the attention of wildlife park officials (*48, 79*). A compromise was reached in which the San groups may stay as long as they only used traditional means of hunting.

In 1986 the government decided that the CKGR should strictly be a wildlife reserve and that residents should be relocated. San groups wished to stay in the reserve and proposed to work with park officials to sort out problems. This was declined and the resistance to resettlement was met with threats from the government and discontinuation of services. In 1997 the people of the CKGR were resettled from ≠Xade in the Central Kalahari Game Reserve to New ≠Xade, a large settlement in Ghanzi District, southwest of the reserve, and Kaudwane, a large settlement in Kweneng District not far from Khutse Game Reserve. Promises of large compensation to people who would move soon were made. In reality very little compensation was paid out and people struggled to keep their livelihoods. A San run non-governmental organization, 'First peoples of the Kalahari' (FPO), worked with CKGR residents and took the Botswana government to court. In 2005, the government ruled that the CKGR was off limits to people even though some residents still lived there. San people trying to access the CKGR were shot at by government officials with teargas and rubber bullets, some individuals were injured, arrested and detained. In 2006 the final decision of the court was that San groups were unlawfully removed. The government, however, was not required to restore services because it was not unlawful for them to have stopped these services. At the end of 2006 San groups were allowed to return but without any domestic stock. They are only allowed to live from hunting and gathering practices. Hunting licenses, however, are still not issued and people are living mainly of wild foods from the reserve and food they obtain from outside (*80*).

The /Gui and //Gana group of the present study were sampled at the Khutse Game reserve south of the CKGR. The group consisted of a mixed group of /Gui, //Gana (and possibly Kgalagari) individuals. Individuals that identified themselves as Basarwa (the Tswana name for the San) during sampling were selected for the present study. This group has not been represented in any previous genetic studies.

2.2.2 KhoeKhoe (Pastoralist Khoe groups)

The Khoe pastoralists (KhoeKhoe linguistic dividion) can be divided into three ethnic divisions, namely, the !Ora (or Korana), the Cape Khoe and the Nama. Early reports also made mention of a fourth division, the Einiqua (language "Eini") that lived along the Orange River to the east of the Korana, but very little is known about this group (*48, 81–83*). The only distinct Khoe group (speaking the KhoeKhoe language grouping) living today is the Nama of Namibia. The Korana (!Ora) and Cape Khoe (Cape KhoeKhoe) of South Africa represent extinct groupings of Khoe language and culture but their descendants live in the Coloured population of South Africa. The Hai//om, a San group of north Namibia also speak a KhoeKhoe language, however, this group is thought to have originated as result of contact between the Nama and the !Xun of northern Namibia (*48–50*).

The Nama of Namibia and north western South Africa

The Nama (or Namaqua) are the best-known Khoe group. Nama individuals live in south and central Namibia, and to a lesser extent in the northern Cape (SA) and eastern parts of Botswana. The Nama people most probably came from an area located in the current northern parts of the Cape province (SA) and divided into two large subdivisions of people, the Great and the Little Nama (*48, 84, 85*).

The Great Nama (Gai-Naman) settled in the great Namakwaland area of Namibia prior to European contact. Several tribes existed with certain associated territories. In recorded history the Great Nama were divided into seven tribes (the Gai-//haun or *Rooi Nasie*; the !Gami-≠nûn or *Bondelswarts*; the //Haboben or *Veldskoendraers*; the !Khara-khoen or *Kopers*; the //Khau-/gôan or *Swartboois*; the //O-gain or *Groot Doden*; the ≠Aonin or *Topnaars*) (*48, 84, 85*). The Nama presently use mainly the Afrikaans group delineations (*italic*).

The Little Nama (≠Kham-Naman) only migrated into Namibia in the 19th century in separate tribal groups. They were also known collectively as the "incoming groups" and the *"Oorlams"*. The Little Nama tribes were the /Hôa-/aran or *Afrikaners*; the /Khobesin or *Witboois*; the !Aman or *Bethaniers*; the /Hai-khauan or *Bersebaers* and the Gai-/khauan or *Lamberts* or *Amraals*. These Little Nama tribes came from the south in search for better grazing but met with the Great Nama and Herero that were already there and conflicts developed. The Nama, who remained south of the Orange River, became incorporated into the "Coloured" population of South Africa (*48, 84, 85*).

The Nama lived a nomadic life and were pastoralists. With the incursion of Bantu-speakers and Europeans into their territory, their tribal organization shifted from hereditary chiefs to military leaders and chiefs. Early forms of tribal organization and social structure quickly deteriorated with the German colonization of Namibia in 1890. Additional factors include a severe drought and a rinderpest epidemic. The Nama revolt and resultant wars (1904-7) finally broke up traditional tribal structure. Although the tribes are dispersed today there are still some chiefs that maintain control over their traditional locations (*48, 84, 85*).

The Nama group in the present study was sampled in the Namibian capital, Windhoek, and is expected to be a mixed sample of the groups discussed above. The Nama is the only Khoe (herder) group of the present study. The Nama have been represented in very few genetic studies (*63, 86*), and Nama results were not presented and discussed separately but rather grouped with other Khoe-San groups from these studies.

2.3 Tuu linguistic division (Southern San)

2.3.1 Remnants and descendants of Khoe and San groups living in South Africa

The people with Khoe-San ancestry in South Africa have to a large extent completely lost their identities and have integrated or transformed into other populations. What we presently know of Khoe and San peoples of South Africa are derived from studies on very few remnant populations that survived into the 1700-1800s.

Most of the South African San groups belonged to the !Ui family of the Tuu (Southern Khoisan) language division. In historical times a large diversity of !Ui languages were spoken throughout all parts of the interior of South Africa. Their geographic range stretched from Namaqualand in the west through the northern Cape, the Free State and Lesotho to KwaZulu-Natal and the south-eastern parts of Mpumalanga (old Transvaal). The best known of these languages is /Xam, a language used to be spoken mainly in the Karoo, south of the Orange River. There were, however, numerous other !Ui languages more or less related to /Xam throughout South Africa. A few of these languages were recorded and still had a few active speakers in recent history like //Xegwi in the southeastern Transvaal. Of the other !Ui languages very little other than a name is known, like //Kx'au of Kimberley, //Ku //e (≠Ungkue) of Theunissen in the Free State, Seroa (N//ng) of the Free State and Lesotho and !Gã !ne of the eastern Cape area (*87*).

Of the South African Khoe culture, language and traditions, very little remains. In 1652 the Khoe pastoralists of the Cape, or the Cape Khoe, rapidly converted their language to Afrikaans or Xhosa (on the eastern frontier). Waves of severe smallpox epidemics afflicted the Cape Khoe population and lead to the further demise of this population (*18*). In the western and northern Cape, the Khoe language survived until recently in the form of !Ora (Korana) and Xiri (Griqua). The descendants of the Korana and Griqua adopted Afrikaans as their mother tongue and today South African Khoe languages are virtually extinct outside a few scattered individuals who retained some knowledge of the languages. One such individual lived near Colesberg. He spoke a dialect of !Ora that was largely unintelligible to Nama speakers, illustrating the differences between these two Khoe languages (*87*).

The next few sections describe the little knowledge we have about the history of these South African Khoe-San descendant groups of whom the ancestry is expected to come from both Khoe (KhoeKhoe) and San (Tuu) ancestors.

/Xam descendants

The /Xam inhabited a region of the Cape Province known as the Great Karoo. The Great Karoo area of South Africa is an arid scrubland with dispersed hills that stretch over an area of 400 000 sq/km of the Northern, Eastern and Western Cape provinces. This area was inhabited by both San and Khoe groups up until the late 1800's. The San group was the /Xam and the pastoralist Khoe group was part of the Korana group. The /Xam had subgroups ("Ss'wa ka" or "Plain bushmen", "/nussa" or "Grass bushmen", "!Kaoken ss'o" or "Mountain bushmen" and "Brinkkop bushmen") but they all spoke the /Xam language with minor dialect differences (*87*).

The western world has learnt about the /Xam through the pioneering work of Wilhelm Bleek, a 17th century linguist who moved from Germany to the Cape Province. Bleek, his sister in law, Lucy Lloyd and his daughter, recorded the cultural practices, language and religion of the /Xam people while providing shelter to various /Xam individuals (`http://lloydbleekcollection.cs.uct.ac.za`) (*88*).

There are many reasons for the apparent disappearance of the /Xam; the principal factor probably being the advance of Bantu-speaking herders from the north and European colonists from the south, which led to the occupation and conquest of the great Karoo in the 18th century. Colonist hunters and farmers moved in and occupied all the remaining hunting ground previously used by the /Xam. The occupation of their resources was not the only reason for the disappearance of the /Xam, they were physically hunted by colonists and bounties were placed on their heads. Hunting parties were organized to hunt "Bushmen". Males that were not killed by hunters fled into the hilltops or were sent off to prisons. Females and children where relocated to farms to serve as farmhands, the so-called tame-bushmen. In the same way Khoe farmers living in the area were in competition with colonists for grazing ground. The Khoe people, however, claimed right to certain lands and had cattle to trade. They therefore generally received more respect from colonists than the San people (*48, 87, 89, 90*).

The descendants of the /Xam females and children who were relocated to farms, today still live on some of the farms but became admixed with the local Xhosa (Bantu-speaking) population. Older farm owners still call some of their labourers "Bushmen" or recall that parents or grandparents of their workers were "Bushmen". Many farmers, however, tell the tale of "Bushmen" that couldn't settle in one place and had "wanderlust". These people became the "Karretjie People" that had their donkey carts as mobile units and moved from place to place to do different periodic jobs (*91*).

The Karretjie People

The word "Karretjie" is an Afrikaans word for "donkey cart", alluding to the mobile lifestyle of the "Karretjie People" on donkey carts. Throughout the great Karoo there exist small bands of people living this mobile lifestyle but due to recent changes in economic factors, this way of living is quickly disappearing. The Karretjie People phenotypically resemble Khoe-San people. Oral and archaeological records also suggest San and Khoe ancestry but the group completely lost their original language and culture. They speak Afrikaans and are classified as "Coloured" in the South African census. Most of the Karretjie People live an itinerant lifestyle and are periodically employed as sheep shearers and fencers. Typically they have a home base or as they call it "uitspanning" or "staning" where they keep their cart in between jobs. These "stanings" are usually on a neutral piece of land such as the section of land between a road and a farm fence. They would stay in this space until their skills in shearing or fencing were required by a farmer. When this happens, they would pack their donkey cart and the whole family and living unit would move to the farm until the work was completed, after which they would move back to the same "staning" (*91, 92*).

The Karretjie People from the present study were sampled around Colesberg (SA) and mitochondrial DNA and Y-chromosome results from the group has been described in (*92*).

The ≠Khomani

The ≠Khomani together with the /'Auni tribe and several other now extinct groups lived in the far northern parts of the northern Cape (north of Upington), the southern part of Botswana and the southern parts of Namibia, roughly where the Kalahari Gemsbok Park is located today. They all spoke branches and dialects of the Taa-Lower Nossob branch of the Tuu family of Khoisan languages (Table S5). In 1980 there were only few individuals left who remembered a lifestyle of active hunting and gathering in this area. They self-identified as N/amani and !gabani but by then only spoke Nama (only one woman could speak the N/u language, but remembered only words). The individuals said that in the past the San of the Gemsbok park area used to live in small scattered groups in the summer and aggregated in the area of the Nossob River (southern Botswana) in the winter. There they traded goods (ostrich eggshell beads and animal skins) with Tswana groups. Their main food sources were gemsbok and small game as well as tsama melons and other wild food (*48, 93*).

The Khoe-San people presently living in this area, spanning the borders of northern South Africa, southeast Namibia and southern Botswana, are from several different tribes that lost their individual tribal identities and speak either Afrikaans or Nama. The southern parts of Namibia, before the Nama colonization, had many San groups from the Taa language family. Today, however, all their descendants speak Nama (*94*). The South African descendants of these San groups mostly classify themselves as Coloured today.

A group of South African descendants of these scattered southern Kalahari tribes now call themselves collectively ≠Khomani. They have had a recent rediscovery of their identity; they won a land claim and organized themselves into a community governed by a council. Only very few old individuals, from the Northern Cape (SA) and Botswana, however, still speak the N/u language. The term ≠Khomani was not known to the N/u speakers, it was introduced to San descendants of the northern Cape by representatives of the South African San Institute (SASI). Other than N/u, the only other extant Tuu language is !Xóõ, of southern Botswana. Unlike N/u, however, !Xóõ is still an active language and is being taught to children (*95, 96*).

The ≠Khomani group from the present study was sampled in the northern Cape, (200km north of Upington), in the area owned by the ≠Khomani community (located near Askham close to the Botswana-South Africa border). A genetic study involving autosomal SNPs was conducted on the same ≠Khomani community (*10*). Relative analysis reveled three relations between the current study ≠Khomani group and the ≠Khomani from (*10*), and these three individuals were excluded from the Henn *et al.* (*10*) set.

Coloured Groups – South African Khoe descendant groups

The Khoe groups of South Africa comprised the Cape Khoe of the southern parts of the Cape Province, the Korana who occupied large parts of central South Africa extending over the Northern Cape into the Free State and the Nama of the North Western Cape region in the Richtersveld area extending into Namibia. Although Cape Khoe and Korana do not exist anymore today as specific populations, their descendants were incorporated into "mixed culture" groups like the Griqua and Coloured groups with their own associated cultures. Certain aspects of Khoe culture can still be recognized in rural areas where livestock rearing is the prime economic goal. In a way the Khoe culture formed the base of the Griqua and Coloured cultures that developed (*48*).

The two Coloured groups included in this study were sampled at Colesberg (Northern Cape) and Wellington (Western Cape, near Cape Town). The Khoe groups occupying these areas at the time of European contact were the Cape Khoe (Wellington) and the Korana and Griqua (Colesberg area). Coloured groups have been studied in autosomal genetic studies before (*8, 97–99*). These four previously studied groups came from the region where the Coloured (Wellington) group from the present study comes from (Western Cape - near Cape Town). The mitochondrial DNA and Y-chromosome results from the Coloured group from the Northern Cape, Coloured (Colesberg), were published in (*92*).

3 Population structure

3.1 Principal Component Analyses

3.1.1 PCA of individuals

Principal Component Analyses (PCA) was performed using EIGENSOFT (*100*) for the Southern Africa dataset (with and without admixed individuals), global dataset (with and without admixed individuals) and a subset of all sub-Saharan individuals (non-admixed) from the global dataset. For each PCA, SNPs with a minor allele frequency (MAF) less than 5% across all individuals in the particular dataset were removed, and SNPs in Linkage Disequilibrium (LD, $r^2 > 0.2$ were also removed. Sample sizes were limited to a maximum of 20 (populations with more than 20 individuals were randomly sub-sampled to the size of 20, but note that the main text PCA figure (Fig. 1B and C) used all individuals).

Figures S4 (with admixed individuals) and S5 (without admixed individuals) shows the first 10 PCs for Southern African individuals based on ~2.3M SNPs. Figures S6 (with admixed individuals) and S7 (without admixed individuals) shows the first 20 PCs for the worldwide set of individuals based on ~270k SNPs. Figure S8 (without admixed individuals) shows the first 20 PCs for sub-Saharan individuals based on ~270k SNPs.

3.1.2 Correlation with geography

We computed Procrustes correlations (*101*) between PC1-PC2 space and geographical sampling locations of the individuals. We did this separately for all sub-Saharan individuals, and Southern African individuals, and compared the results when Bantu-speaking individuals were included and excluded, respectively (Figure S9).

3.1.3 Population PCA

Here we do population Principal Component Analyses on a matrix containing the SNP frequencies for each population. Unlike the PCA based on genetic data of individuals, where sample size within each population influences how well populations are represented in the PCA results, in the population PCA, each population appears once in the matrix and therefore all populations have the same weight in the analysis. Thus, the sample sizes only influence the accuracy of the SNP frequencies estimation from the data (although the number of populations from a given meta-population, e.g. Khoe-San vs. Americans, can influence the analysis).

The population PCA shown in Figure S10 gives similar results as the PCA on individuals in Figure S6 for the first 5 PCs. We may be interested the "quality of representation" of the populations on each PC. Intuitively, a population has a high quality of representation on a given PC if the population is close to the PC. Mathematically, this means that the scalar product between the multi-dimensional vector representing the population and its projection on to the PC is large. Conversely, if the population and the PC are orthogonal, the quality of representation of that particular population on the PC is basically zero. African populations have the best quality of representation on PC1, followed by East Asian and American populations, and finally European populations and Middle Eastern populations, which have a poor representation score. This allows us to roughly "assign" PCs to populations. PC2 is driven by European and Middle Eastern populations, PC3 by South American populations, PC4 by the African Khoe-San populations. PC5 together with PC12 represents Oceanian populations and PC6 is completely driven by the Hadza population.

Each PC is a linear combination of SNPs, so that $\mathrm{PC}_j = \sum(\alpha_i \times \mathrm{SNP}_i)$. The α values differ from SNP to SNP so that some SNPs "participate" more to the general direction of the PC than other SNPs that have α close to 0. Figure S11 displays the distribution of the absolute values of the αs, for the most population-informative PCs. PC1 and PC2 have a very different profiles compared to the other PCs. For PC1, quite many SNPs participate in the definition of PC1's direction. This is also true for PC2, but to a lesser extent. For the other PCs, many SNPs do not have a strong effect on the PC and only a small proportion of SNPs have a reasonable influence on the PC. At those SNPs, we may observe some particularities in the populations driving the direction of the PC.

3.1.4 Project populations onto predefined axes using population PCA

Here we project particular populations onto axes that are predefined by some reference populations. A similar approach has previously been used to e.g. project modern humans onto axes of archaic humans (Neanderthal and Denisova) and chimpanzee (*102*), although in this case we use allele frequencies of populations rather than haploid sets of gene copies from each individual. The approach can be used to test particular hypotheses. For example, due to their intermediate geographic position, it might be expected that Northern San groups have genetic contributions from both Southern San groups and central African hunter-gatherers (today represented by e.g. Pygmy groups), and potentially also some admixture with west Africans. To query such a hypothesis, we could set up reference populations of Southern Khoe-San, Pygmy groups, and Bantu-speaking groups.

In short, the procedure is the following:

- For each window (20 or 50 SNPs/window in order to capture multiple parts of the genome separately) do population PCA with three anchor groups. For example (for the 270k SNPs):
  - Southern Khoe-San: ≠Khomani (A), Nama, and Karretjie,
  - Bantu: Bantu-speakers (South Africa), Herero, Bantu-speakers (Southern Africa – HGDP), Bantu-speakers (Kenya – HGDP),
  - Pygmies: Mbuti and Biaka.
- You end up with a PC1-PC2 plot like Figure S12.
- Project other populations along the PC1-PC2 axes (Figure S13).
- Assign the projected population to the closest anchor (grey color) population.
- Count the proportion of windows assigned to each anchor group and for each group, we have a vector (p_{SKS}, p_{BS}, p_P, where SKS, BS, and P stands for 'Southern Khoe-San', 'Bantu-speakers', and 'Pygmies'.) .

If we follow the procedure above (there are some 15,000 non-overlapping windows of size 20 SNPs) and project the set of global populations onto the two axes defined by Southern Khoe-San, Pygmies, and Bantu-speakers, we find the following pattern, Figure S14. Figure S15 shows only African and Middle Eastern populations. We note that the results for 50 SNP windows are very similar, the correlation r between the 20 and the 50 SNP windows are for Southern Khoe-San-component $r = 0.9999998$, for the Bantu-speaking-component $r = 0.999926$, and for the Pygmy-component $r = 0.9999966$. Using single SNPs (instead of a window-based approach) produce qualitatively similar results.

We can consider the "component" of the three anchoring groups in the investigated populations as ancestry from the particular anchoring group. We find for:

- 'Southern San ancestry' (decreasing order): Ju/'hoansi, !Xun, /Gui and //Gana, and KB1.
- Bantu-speaking ancestry (decreasing order): Luhya, Yoruba (HGDP), Mandenka, Maasai, Yoruba (HapMap), African American, Sandawe.
- Pygmy ancestry are (by decreasing order) Hadza, Surui, Karitiana, Colombian, Pima, Papuan, Melanesian, Maya.

We find that the northern Khoe-San populations do not appear to be a admixed group of Southern Khoe-San and Pygmy groups (or Bantu-speaking groups). Furthermore, the Sandawe, a click-speaking East African group, does not have much shared affinity with the Southern Khoe-San. We also find that not many populations have substantial Pygmy affinity, the populations with largest ancestry are Native American and Oceanian populations. This may appear a bit strange, but it is probably due to the fact that all other populations have a little affinity with some other anchoring group and these Native American and Oceanian populations are the ones with basically no preference. The Hadza is the group that shows the greatest connection to Southern Khoe-San groups and Pygmies, seen as the Hadza are located close to the center in the two Figures S14 and S15.

3.2 ADMIXTURE analyses

Individuals were clustered based on SNP genotypes using an unsupervised clustering algorithm implemented in ADMIXTURE (*16*). Default settings were used (except for generating a seed from the system clock). For a given value of K (the number of clusters considered) 100 replicate analyses were performed for each dataset and subset of individuals. The 100 replicate runs of ADMIXTURE (for each subset of individuals and each K) were analyzed with CLUMPP version 1.1.1 (*45*) to identify common modes among replicates. The CLUMPP analysis used the *LargeKGreedy* algorithm with 10,000 or 2,000 random permutations. Common solutions were identified by looking at the CLUMPP pairwise G' values. All pairs with a symmetric similarity coefficient $G' > 0.9$ were selected to be representative of a single mode. For each K we used the most frequenly occurring mode identified and ran CLUMPP a second time (using the LargeKGreedy algorithm and 10,000 random permutations), using only the replicates belonging to the mode. From this second analysis, we obtained the mean across replicates of the cluster membership coefficients of each individual, for each mode at each value of K. The clustering results were visualized with Distruct (*103*).

3.2.1 SNPs

We ran ADMIXTURE for both the full set of individuals and the set of 'non-admixed' individuals for both the global dataset and the Southern Africa dataset. We also investigated a subset of all sub-Saharan individuals (non-admixed) from the global dataset. For the Southern Africa dataset, the results for the most common modes are presented in Figures S16 and S17. For the global dataset, the results from $K = 2$ to $K = 15$ are presented in Figures S18 and S19. For the subset of sub-Saharan individuals, the results for the most common mode from $K = 2$ to $K = 12$ are presented in Figure S20. Number of assignments (out of 100 cluster assignments for each K value) to the major mode and the most common minor mode are given in Table S6. Only major modes for each K value and dataset were plotted in DISTRUCT.

3.2.2 Haplotypes

We constructed haplotypes by dividing the phased data in Dataset B into contiguous windows of 20,000 bp. Since variable number of SNPs between windows can affect analyses of diversity and population relationships, we excluded windows that had less than 5 SNPs, and windows with more than 5 SNPs were randomly downsampled to 5 SNPs. We only considered SNPs that had a minor allele frequency of at least 10% in the whole data set. We then counted each haplotype allele at each locus in each individual, resulting in counts of 0, 1 or 2 haplotype-alleles.

3.2.3 Discussion of population structure in sub-Saharan Africa

Focusing on sub-Saharan Africa (Figure S20), and starting by assuming two clusters (K2), the genomes of the individuals are split into a component (red) present mostly in African hunter-gatherer groups (San, Pygmy, Hadza and to a certain extent Sandawe and the Southern African Bantu-speaking groups) and a component (green) present mostly in pastoralists that speak either Niger-Kordofanian languages (Mandenka, Yoruba and Bantu-speaking groups) or Nilo-Saharan (Maasai). Neither the Bantu-speaking groups from Kenya (Bantu-speakers (Kenya - HGDP) and Luhya) nor the two Niger Kordofanian groups (Yoruba (HGDP) + Yoruba (HapMap) and Mandenka) contain the red component. This can be perceived as evidence that the red component in the Southern African Bantu-speaking groups was the result of admixture from the resident Khoe-San groups when the Bantu-speakers moved down into Southern Africa. The southeastern Bantu-speakers (Bantu-speakers (South Africa)) seem to have had a greater amount of geneflow with Khoe-San groups than southwestern Bantu-speakers (Herero).

At K3, a component (blue) that seem to be predominant in the East African Nilo-Saharan speakers (Maasai) as well as the two East African hunter-gatherer groups (Sandawe and Hadza), become visible. It is also visible at low frequencies in the two Kenyan Bantu-speaking groups (Bantu-speakers (Kenya - HGDP) and Luhya). The blue East African component is also present in low frequencies in certain Khoe-San groups. The frequencies of the blue component vary across Khoe-San groups but are the highest in the pastoralist Khoe group. The highest presence of the East African component in the pastoralist Khoe group (Nama) gives support to the theory that pastoralism was introduced to Southern Africa by an East African pastoralist group ancestral to the Khwe (*104*). The Khwe indeed also contain the blue component in low frequencies. Although the Khwe are not pastoralists today, they might have been in the past. The other two groups that contain the blue component are the two southern groups ≠Khomani (A and B) and Karretjie. It is very likely that these two groups are composed of a mixture of both Khoe and San ancestry. It is therefore interesting to observe that the blue component has a more even representation across Nama individuals (which are all Khoe pastoralist) while in the two groups that has been proposed to be a mixture (very recently) of Southern San and Khoe, the representation is not as even. The remaining San groups, /Gui and //Gana, !Xun and Ju/'hoansi do not harbor the blue component.

At K4 a component (sea-green) present in the two Pygmy groups can be observed. The Biaka Pygmy group has evidence of more gene flow with Niger-Kordofanian groups than the Mbuti Pygmy group. The Hadza also contain this component to a certain extent. Very low but observable representation of the sea-green component is in the northern San groups (Khwe and !Xun and less in Ju/'hoansi and /Gui and //Gana) but not in the Southern Khoe and San groups (≠Khomani (A and B), Nama, Karretjie).

At K5 a Hadza component (light-blue) is observed and at K6 the two Pygmy groups form separate clusters (sea-green and blue-green). At K7 a component that separates the northern San groups (purple) from the Southern Khoe-San groups (red) become visible. The Ju/'hoansi, Ju/'hoansi (HGDP) and !Xun contain the northern San component. The northern San component is also present in the Khwe, although the Khwe seems to have a large green (Niger-Kordofanian) component as well. The !Xun also contain a visible amount of the green component. The Southern Khoe-San groups, ≠Khomani (A and B) and Karretjie contain predominantly the red component, while the Khoe group (Nama) also principally contain the red component. The /Gui and //Gana, seem to be a mixture of the northern and Southern Khoe-San component with a visible amount of the green Niger-Kordofanian component.

At K7 all Niger-Kordofanian populations (including Bantu-speakers form across Africa) still maintain a homogenous cluster. Thus the population structure analyses of sub-Saharan African populations shows a clear signal of the range expansion of West Africans who today live in most areas of Africa. South-eastern Bantu-speaking groups arrived in Southern Africa around 1,200 years BP as part of the 'Bantu expansion' driven by the development and spread of agriculture (*105*). At the time of European colonization, the eastern parts of Southern Africa were occupied by Bantu-speaking groups but rock-paintings and archaeological remains indicate that Khoe-San groups had previously occupied these regions (*106*). On the contrary, in the western parts of Southern Africa, many indigenous Khoe-San groups still exist today and Bantu-speakers arrived only a few hundred years ago. In the south-eastern Bantu-speakers (South Africa), the distinct fraction of Southern Khoe-San ancestry indicates assimilation of local Khoe-San populations during the “Bantu expansion”, whereas the south-western Bantu-speakers (Herero) display a much smaller fraction of Khoe-San ancestry. There is also ample evidence of gene flow into the Khoe-San groups from Bantu-speaking groups in essentially all sampled Khoe-San groups.

At K8, a component that are more prominent in the Bantu-speakers (green) than in the non-Bantu-speaking Niger Kordofanians (Mandenka and Yoruba) emerges. The Mandenka and Yoruba contain a component (dark green) that is at lower frequencies in the Bantu-speaking groups. The southwestern Bantu-speaking Herero, seem to contain a higher proportion of the dark green component than the southeastern Bantu-speakers, while it is almost absent in the east-African Bantu-speakers.

At K9 the Sandawe form a separate component (light blue) and at K10 the Khwe form a separate component (orange). The genetic makeup of the Khwe is distinct from other San and Khoe groups, for example, assuming few clusters (K2-K9), the largest fraction of the genomes of the Khwe cluster with Bantu-speaking groups and the second largest ancestry component clusters with Khoe-San groups, but allowing additional clusters (K10 onwards), the Khwe forms a distinct group (except three individuals, possibly recent migrants). The complex ancestry of the Khwe could be explained by high levels of (non-recent) admixture – possibly related to the fact that the Khwe live in a natural migration corridor between central and Southern Africa formed by the Zambezi and Kunene rivers and the Okovango delta. Alternatively, the genetic makeup of the Khwe could represent a distinct group that shares ancestry with many sub-Saharan groups, including Bantu-speaking and Khoe-San groups.

At K11 the east-African and Southern African Bantu-speakers form separate clusters (green and light-green). At K12 it seems that the Nama is assigned to a cluster, this cluster is present in the two ≠Khomani (A and B) groups but not substantially in the Karretjie group. K13 introduced an internal split in the Sandawe, which disappears at K14 when instead two other pairs are separated namely the Mandenka and Yoruba and also the southeast and southwest Bantu-speakers (Herero). At K15 the internal split in the Sandawe re-appear.

The discussions above concerns the analyses involve datasets where recently admixed individuals had been removed. It should be remembered that even more geneflow has occurred between various populations (Figures S16 and S18) than is apparent in Figure S20. However, the basic Southern African population structure illustrated by Figure S17 is still apparent in Figure S16, despite the inclusion of admixed individuals. Furthermore, large groups of people in Southern Africa today forms part of mixed culture groups such as the Griqua and the various Coloured groups in Southern Africa. These groups originated from numerous Khoe and San groups who lost their individual identities and have various levels of admixture with Bantu-speakers and non-Africans. Thus, although the San and Khoe groups are relatively small populations today, their genetic contribution to the group that self-identify as ”Coloured” in South Africa may be substantial, e.g. (*97*). The Coloured groups from South Africa showed varying amounts of admixture from various different populations including Khoe-San, Bantu-speakers, Europeans, East Asians, and South Asians (Figure S21, Figure S16 and S18). The contribution varied depending on the region where the Coloured groups were sampled (Table S1) (*92*), with individuals sampled in Wellington having a larger fraction of non-African ancestry compared to individuals sampled in Colesberg. Note also that South African individuals with Khoe-San ancestries frequently self-identify as Coloured in official censuses although their personal account of ancestry will affirm their Khoe-San heritage. For instance, ≠Khomani, Karretjie People and South African Nama (not the Namibian Nama from this study) individuals will often formally self-identify as “Coloured”. The complex ancestry of the Coloured group highlights the importance of genetic profiling for medical studies in South Africa.

3.3 F_{ST}

Population differentiation was estimated based on the SNP data and using F_{ST}, using eqn. 5.3 in (*107*).

3.3.1 Global set of populations

Pairwise F_{ST} was calculated for for the global dataset (∼270k SNPs, admixed individuals removed), displayed in Figure S23, and visualized as a population-tree using the Neighbor-Joining (NJ) algorithm (Figure S24).

3.3.2 Sub-Saharan African populations

Figure S25 displays the F_{ST}-values as a distogram. We plot F_{ST} as a function of geographic distance for all sub-Saharan African populations (Figure S26) and for all sub-Saharan African populations excluding Bantu-speaking populations (Figure S27) and find significant correlations.

3.3.3 Southern African populations

For the larger set of SNPs (∼2.3M), pairwise F_{ST} was computed in the same way as for the smaller set, and Table S7 gives the pairwise F_{ST}-values for the Southern African populations. Figure S28 displays the F_{ST}-values as a distogram and Figure S29 shows a population-tree (obtained using the NJ algorithm) for Southern African populations.

4 Geography, language, subsistence and genetic relationships

4.1 Predictive approach

To investigate the relationships between genetic and non genetic information, one approach is to assess how well a non-genetic variable can predict the observed genetic patterns. This approach has been used to study the role of geography and languages in shaping human population genetic structure (*108*). Here, in order to understand the relationships of geography, language and subsistence with patterns of genetic variation in Africa, we regressed the principal component scores – computed from the SNP data – with geographic, linguistic and subsistence co-variates. The predictive capacity of different co-variates and groups of co-variates was assessed using cross-validation scores. Cross-validation is useful for comparing several models with different complexity levels.

For each principal component we estimated the predictive capacity of seven models containing different type of information: “Geography”, “Subsistence”, “Language”, “Geography + Subsistence”, “Geography + Language”, “Subsistence + Language”, “Geography + Subsistence + Language”. Geographic information is represented by a spatial quadratic trend ($latitude$, $longitude$, $latitude^2$, $longitude^2$, $latitude \times longitude$), linguistic information by the linguistic subgroups and subsistence information by the categories given in Table S4.

For each model, each of the 20 first PCs is regressed on the corresponding set of co-variates. The residual sum of squares is not a good measure of the predictive error of a model, as it will always decrease when the complexity of a model increases. To prevent overfitting we thus estimated a 5-fold cross-validation score. In this procedure an individual can not be part of the training set and the validation set at the same time. For each model and each PC, we split the dataset in 5 groups and calculated the predictive errors for individuals in a group using the regression parameters estimated from the four other groups. The average over all groups provides the 5-fold cross-validation score, which is a good estimate of the predictive capacity of the model (the smaller the score is, the better the model).

First we analyzed the sub-Saharan dataset. Figure S30 shows the predictive error for each model for the 20 first PCs. Interestingly, the predictive errors highly vary between models for PC1 to PC10 (mean sd = 0.0009), but are quite stable for PC11 to PC20 (mean sd = 0.0003). The first principal components are also easier to predict than the remaining components, as the errors are generally smaller for PC1-10 than for PC11-20 (average minimum error: 0.0004 for PC1-10, 0.0018 for PC11-20). Both results and the fact that the first PCs are the most informative in terms of genetic variation imply that the relationships between the different co-variates and population genetic structure should be investigated by looking further at the regressions of the 10 first PCs. Therefore, we compute the predictive errors averaged over the 10 first PCs for each of the 7 models (Figure S31A). These errors are relative to the model with geography only: values smaller than one indicate improved predictive capacity compared to geography. The models including subsistence alone or language alone have both higher predictive errors than the model including geography alone. However, the models “Geography + Language” and “Subsistence + Language” have both predictive errors much smaller than 1, which indicates that languages provide additional information for predicting PCs, compared to geography alone. Moreover, the model including geography, and both linguistic and subsistence information performed better than the model without languages (“Geography + Subsistence”). Thus, even after accounting for geography and subsistence, languages provide information for the prediction of the PCs.

We also analyzed the Southern African dataset of 2.3M SNPs (Figure S31B). Again, models including language alone or subsistence alone have predictive errors larger than 1, and models including several co-variates have errors smaller than 1, indicating that geography alone is not the best predictor of genetic variation. For this geographically restricted dataset, models including linguistic information and models including subsistence information (in addition to geography) have a similar predictive capacity. Therefore we can not pinpoint whether language or subsistence is the best predictor.

Finding that additional information improves the predictive capacity of the geographic model is not surprising since the recent 'Bantu expansion' is expected to reduce the correlation between geography and genetics. Indeed, when excluding the Bantu-speakers from the Southern African dataset, models with geography and additional information are only slightly smaller than 1, indicating that language and/or subsistence do not improve the predictive capacity of a geographic model for this dataset as much as for datasets that include Bantu-speakers (Figure S31C).

In conclusion, languages improve the predictive capacity of a model that includes only geography in both the sub-Saharan and the Southern African dataset. However, once accounting for both geography and subsistence, linguistic information improves the prediction of the PCs for sub-Saharan dataset only. A possible explanation is that the linguistic classification is not fine-scale enough when focusing on the Southern African populations. Another possibility is that the hypothesis of the language shift coupled to the introduction of pastoralism (*51*) in certain Khoe-San groups is correct. In this hypothesis, all the groups speaking the central Khoe-Kwadi branch (Nama, Khwe, /Gui and //Gana from this study), adopted their languages recently (last 2000 years) and their linguistic affiliation is not expected to predict their genetic identity. Indeed in the genetic structure analysis they do not form one homogeneous cluster (unlike the two Northern San and Southern San groups). In addition, when excluding the Bantu-speakers from the Southern dataset, neither the subsistence nor the language improves the prediction of the PCs.

4.2 Mantel tests

To confirm our findings, we calculated the significance of the correlations between genetic distances, and geographic, linguistic or subsistence distances, using Mantel tests and partial Mantel tests. Mantel tests are designed to reject the absence of correlation between two matrices of distances, and partial Mantel tests allow to control for other variables while doing this test (*109, 110*). We used pairwise F_{ST} between populations (computed as explained in section 3.3) as genetic distance. The linguistic distances, $dLang$, are computed as follows (see also Table S4):

- $dLang(pop_1, pop_2) = 0$ if the languages of the two populations are in the same linguistic subgroup,
- $dLang(pop_1, pop_2) = 1$ if the languages are in the same linguistic main group,
- $dLang(pop_1, pop_2) = 2$ otherwise.

The subsistence distances, $dSub$, are computed as follows:

- $dSub(pop_1, pop_2) = 0$ if the subsistence categories of the 2 populations are the same,
- $dSub(pop_1, pop_2) = 1$ otherwise.

For the sub-Saharan dataset, and for each considered distances (geographic, linguistic or subsistence) the correlation to the genetic distance was significant (respective p-values: 0.04 ($dGeo$), 0.001 ($dLang$), 0.03 ($dSub$)). Using a partial Mantel test, we found that the correlation with linguistic distance remained significant even after correcting for both geographic and subsistence distances (r = 0.43, p-value = 0.001). Similarly, the correlation with geographic distance remained significant after correcting for both linguistic and subsistence distances (r = 0.28, p-value = 0.04). Finally, the correlation between genetic and subsistence distances did not remain significant after correcting for linguistic distance using a partial mantel test (p-value: 0.7). For the Southern African dataset, none of the distances were significantly correlated with the genetic distance (respective p-values: 0.7 ($dGeo$), 0.13 ($dLang$), 0.06 ($dSub$)).

These results confirm that, once accounting for geography and subsistence, linguistic information is significantly correlated with genetic distances for the sub-Saharan dataset. Moreover, languages are more correlated than subsistence to genetic distances in this dataset (regardless of accounting for geography). More surprisingly, we did not find any variable correlated to genetic distance for the Southern African dataset. Once again we may lack the fine-scale linguistic information that would be correlated with F_{ST}. Note that using another method for calculating linguistic distances might change the results somewhat (*111*). However, a recent language shift in Khoe-Kwadi speakers (as discussed above), and a predominant cultural diffusion of pastoralist practices to the Khoe people (Nama) probably contribute to the lack of coloration between linguistic distance and subsistence to genetic distance. The reason why geographic distance do not correlate could be due to the inclusion of Bantu-speakers and the fact that the Nama have a recent (last 500 years) place of origin far south (northern Cape, SA) from where they were sampled (Windhoek, Namibia).

5 Reconstructing population history using genealogical concordance

Here we take an approach for investigating the population history of sub-Saharan African populations that is based on randomly sampling a single gene copy from each of four populations. This has the advantage of avoiding the effect of genetic drift within populations that could potentially affect other inference methods based on population samples, e.g. neighbor-joining trees based on pairwise genetic distance between populations. In the history of four genetic lineages, there are only three possible genealogical topologies, of which the one that mirrors the population topology (i.e. the 'concordant' topology) is expected to be most frequent. By testing possible topologies, we can build a tree that is supported by each performed four-wise hypothesis test (*17*). While the history of sub-Saharan Africa does not conform to a simple tree-like model, our approach is designed to capture the major flow of ancestry in the past. Since the excess of the concordant topology increases with the amount of genetic drift separating populations, we also use a basic coalescent model to estimate population divergence times within the same framework below (*112, 113*).

5.1 Population topology inference using concordance tests

For each combination of four human populations and/or chimpanzee (A, B, C and D), we sample one random copy from each population and score which genealogical topology is supported by the configuration of alleles at each site (assuming an infinite sites model of mutation). We then test if a particular topology, *e.g.* (A,B),(C,D), is significantly more frequent than the other two topologies, (A,C),(B,D) and (A,D),(B,C). Our test statistic C (*17*) computes the excess of the putatively 'concordant' category [(A,B),(C,D)] over the second most frequent category (the most frequent 'discordant' category, denoted 'discordant 1') as

$$C = \frac{N_{conc} - N_{disc1}}{N_{conc} + N_{disc1}}. \tag{1}$$

Here, N_{conc} and N_{disc1} are the counts of the concordanct category and the most frequent discordant category, respectively. To obtain standard errors, we use a block jackknife procedure (*114*) [Box 18.6] where we divide the genome in 50-200 contiguous blocks with the same number of informative SNPs in each, and obtain Z-scores based on the number of standard errors the statistic deviates from zero. If the C statistic is significantly positive ($>$ 2 standard errors from 0), we consider the topology supported. Note the difference from E defined by (*102*) as $E = \frac{N_{conc} - 0.5 \times (N_{disc1} + N_{disc2})}{N_{conc} + 0.5 \times (N_{disc1} + N_{disc2})}$ (notation as above), which represents a similar approach, but includes the possibility of two topologies both being significantly supported, where as for C there can only be one most frequent category (*e.g.* even if a single category is much less frequent than the other two).

To look for evidence of gene flow, we also computed the D-statistic of ref. (*7*), which in our notation can be defined as

$$D = \frac{N_{(A,C)(B,D)} - N_{(A,D)(B,C)}}{N_{(A,C)(B,D)} + N_{(A,D)(B,C)}}. \tag{2}$$

5.1.1 Application to sub-Saharan Africa

We tested all possible four-wise combinations of sub-Saharan African populations in our data set (admixed individuals excluded), randomly sampling a single gene-copy from each population of interest at each site. To avoid the influence of ascertainment bias, we only considered the intersection of SNPs that had minor allele frequency (MAF) $>$ 10% in each of the 4 tested populations (except for the chimpanzee). We then used the set of 8,768 significant four-population topologies (Z-score $>$ 2) to create a supertree using the neighbour-joining algorithm in the software `clann` (*115*). To assess node support, we randomly sampled 100 bootstrap pseudo-replicates of the accepted topologies, and collapsed nodes that were seen less than 85% of the time.

5.1.2 Results

While there is substantial evidence of gene flow in many comparisons (Table S8), Figure S37 and Table S8 provides evidence that all Khoe-San groups except the Khwe share a common population history that is separated from all other populations, supporting the view that these populations indeed represent a basal population lineage in the human species. Also, we find a common history between the East African hunter-gatherer groups Hadza and Sandawe that was not apparent in previous studies e.g. (*8, 10*). We also find support for a division of the Khoe-San groups into a Northern and a Southern group. Together with previous studies finding that non-Africans are more closely related to West Africans and/or East Africans than the Khoe-San e.g. (*7, 8, 102, 116*), our results suggest that the historical roots of much of today's population structure in Africa dates back prior to the out-of-Africa expansion.

5.1.3 Robustness of concordance tests in the presence of gene flow

We investigated the robustness of the C-statistic in the presence of admixture using coalescent simulations (*117*). We considered a model where 4 populations A, B, C and D shared a history where populations A and B diverged T years ago from an ancestral population, which in turn had diverged $T \times 2$ years ago from C, and before that, the ancestors of these populations diverged $T \times 3$ years ago from D. We considered two different models of admixture (Figure S33): (1) an admixture fraction c is contributed by population B to population C (2) an admixture fraction c is contributed from population C to population B. In both cases, the admixture occurred 500 years ago. We assumed an effective population size of 10,000 individuals and a generation time of 25 years. We generated 10 gene-copies from each population and sampled a random gene-copy at each SNP for each population. We simulated 1 million independent SNPs and computed C from the observed frequencies of each topology. We compared the results of either using a randomly sampled copy from population D to estimate the topology between (A,B,C), or alternatively using the ancestral allele instead of the allele found in population D.

We find that for model 1, the method is robust to admixture fractions up to $c = 0.5$. That the approach is not robust for values of $c > 0.5$ is not surprising since in that case, population C is expected to share $> 50\%$ of it's genome with population B, more recently than the split of population A and population B (Figure S34A and S34B). For model 2, the method is more sensitive to admixture, with the wrong topology ($C < 0$) being inferred for $c \geq 0.4$ (Figure S34C and S34D), but such high levels of admixture might not be very common. For the present study, most admixture is expected to be from Bantu-speaking groups and their relatives into hunter-gatherer groups, which under the hypothesis that hunter-gatherer groups diverged from other populations relatively early is most similar to model 1. Reassuringly, we also find similar results using both the ancestral allele and an allele from a basal population lineage (Figure S34), suggesting that analyses using 4 human populations or alternatively 3 human populations and chimpanzee are both useful.

5.1.4 Non-Africans share the most recent history with East Africans

Due to computational constraints, non-Africans were not included in the supertree topology inference above. Here, we show evidence that they share the most recent history with East Africans, and thus many population divergences between extant African populations predate the divergence and expansion of modern humans outside of Africa.

We performed 1,029 concordance tests of the form *(chimpanzee,(African,(Hadza,non-African))* (Figure S35) with 21 sub-Saharan African populations (all except the Hadza) and 49 non-African populations (all except HapMap African Americans). We find that all 980 tests not involving the Maasai as the other African group were significantly supported ($Z > 3$).

In the tests involving Maasai, the C-statistic was positive in 40 cases (21 of which were significant at $Z > 2$) and negative in 9 cases (1 of which was significant at $Z < -2$). The single topology which was rejected was *(chimpanzee,(Maasai,(Hadza,Bedouin))* for which the Maasai was inferred to share a more recent history with Bedouins. The remaining 8 tests that were not positive were Brahui, Druze, HapMap GIH (Gujarati Indians), Hazara, Makrani, Palestinian, Pathan, and HapMap Tuscans. We suggest that this is due to the recent admixture in the Maasai from a Middle Eastern source (*118*).

We conclude that East Africans, represented by Hadza and Maasai groups in our data, are the closest sub-Saharan African relatives to non-Africans in our data.

5.2 Population divergence time estimation

The framework used for inferring population topology above also allows estimation of genetic drift in specific branches of the population tree. Since the expected fraction of concordant topologies increases proportional to the time between population divergence events, we can estimate the length T (measured in units of $2N_e$ generations where N_e is the effective number of diploid individuals) of this population divergence time using simple coalescent model *e.g.* (*112*). We have

$$P_{concordant} = \frac{N_{conc}}{N_{conc} + N_{disc1} + N_{disc2}}, \tag{3}$$

where N_{conc} is the number of concordant genealogies, N_{disc1} is the number of discordant genealogies (one of two types), and N_{disc2} is the number of discordant genealogies (of the other type). We denote $N_{disc} = N_{disc1} + N_{disc2}$. Furthermore,

$$\mathrm{E}(P_{concordant}) = \frac{2e^{-T}}{3}, \tag{4}$$

and an estimator of $\hat{T}$ is thus,

$$\hat{T} = -\log\Big(\frac{3 - 3P_{concordant}}{2}\Big). \tag{5}$$

To implement this method, for each locus (total ∼270k), we randomly sample two gene copies from a particular population (**pop1**), randomly sample one gene copy from a second population (**pop2**), and extract the chimpanzee gene copy. We infer a genealogical topology only in those cases where two copies are identical, but differ from the other two (one of which is the chimpanzee – representing the ancestral state), which are also identical, *i.e.*, the sample configuration {2}{2} (the remaining configurations {1}{3}, and {0}{4} are not informative). We score sites where the two copies from **pop1** share a derived allele not found in the **pop2** sample as 'concordant' and the remaining two configurations that allow a topology to be inferred 'discordant' (that one of the samples in **pop1** share the same allele as the sample in **pop2**). When two copies from the same population are sampled in this way, the internode time T in the framework outlined above becomes equal to the total divergence time between **pop1** and **pop2**. We use the equation above to estimate T and obtain confidence intervals in a maximum likelihood framework (*112*) using

$$Log(L(T)) = N_{conc} \times \left(1 - \frac{2e^{-T}}{3}\right) - N_{disc} \times \frac{2e^{-T}}{3}. \tag{6}$$

For a more detailed description of the approach, see (*112, 113*).

5.2.1 Comparisons between San and Mbuti show no discrepancy with resequencing data

Since the estimation of demographic parameters using SNP array data can be heavily influenced by ascertainment bias, we focused our analyses on relative divergences between pairs of populations that are expected to be equally affected by ascertainment bias. This is likely to be the case for comparisons between different Khoe-San groups that have little evidence of gene flow from other populations, as the results shown above suggest that they are approximately symmetrically related to populations used in the ascertainment panel (populations of European, East Asian, and West African [Yoruban] descent). We also hypothesized that this could be the case for comparisons between these Khoe-San groups and Mbuti, since the Mbuti appear to share the least recent population history with the groups mainly used for ascertainment, second only to Khoe-San (*12*).

To test this hypothesis, we compared estimates of F_{ST} computed on our ∼270k SNP data set (with SEs computed using a block-jackknife over contiguous blocks of 2,000 SNPs) for Ju/’hoansi, Mbuti, Biaka, Mandenka & Luhya with previous estimates based on 40 resequenced regions of 2 kb each in samples from these populations (*12*). While SEs are not available for the previously published estimates, we find them largely concordant with our estimates (Figure S36). However, F_{ST} estimates between Ju/’hoansi (San) and one of Luhya, Mandenka and Biaka are notable exceptions, with F_{ST} based on our SNP-data appearing to be overestimated in each of these cases (Figure S36). This could likely be due to ascertainment bias under-represnting variation in the San. However, $\hat{F_{ST}}$ between Mbuti and Ju/’hoansi computed on our SNP-data (0.09952 ± 0.00079 (± 1 SE)) and the estimate based resequencing data (0.09705) are highly similar. This suggests that ascertainment bias has a similar effect on both these populations, and that estimates of genetic differentiation between San and Mbuti are relatively robust to the complex ascertainment scheme behind the ∼270k SNPs common to the various SNP arrays.

To explore this further, we simulated divergence models between San and Mbuti, and San and Mandenka based on their estimated demographic history (*12*) using the software `ms` (*117*), applying different thresholds for Minor Allele Frequency (MAF). MAF thresholds are motivated by the simplified assumption that ascertainment bias can be viewed as an overrepresentation of common alleles in populations related to the ascertainment panel. By applying a MAF threshold to all populations (including those more distantly related to the ascertainment panel), we hope to aquire a set of SNPs that are common in all populations.

Briefly, the simulated model (*12*) included a San effective population size of 20,000 individuals, an Mbuti effective size of 5,000 individuals, and a divergence between the two 110 kya (∼0.1 coalescent time units). We assumed an ancestral effective population size of 10,000 individuals. We also included a change in the Mbuti effective size at 32 kya to 15,000 individuals. For Mandenka, we assumed an N_e of 17,000 that was reduced to 15,000 at 48 kya. We sampled 100,000 independent SNPs from the same number of chromosomes as in our SNP data. The `ms` command line for the Ju/’hoansi-Mbuti divergence model was:

```
ms 64 100000 -s 1 -I 2 38 26 -n 2 0.25 -en 0.0145 2 3 -ej 0.05 1 2 -eN 0.05 0.5
```

And for Ju/’hoansi-Mandenka:

```
ms 82 100000 -s 1 -I 2 38 44 -n 2 0.85 -en 0.0145 2 0.88 -ej 0.02225 1 2 -eN 0.002225
0.5
```

We find that the results for the Ju/’hoansi-Mbuti are largely consistent between our empirical SNP data and data simulated under the models over the range of MAF thresholds applied (Figure S36 B and C). In contrast, the Ju/’hoansi-Mandenka comparisons show more discrepancy, with a similar overestimation of Ju/’hoansi-Mandenka differentiation as suggested by the empirical data above.

5.2.2 Choice of populations

Since the topology inferred above suggest that the ancestral population of Khoe-San groups share a history separated from that of all other human populations, the chronological divergence time between Khoe-San and Mbuti is expected to be the same as between Khoe-San and all other human populations (ignoring later gene flow). Motivated by the observation above that differentiation between San and Mbuti is relatively unaffected by ascertainment bias, we use the Mbuti in place of other non-Khoe-San populations. To estimate the divergence time between the two major Khoe-San groups as indicated by our analyses, we chose the Ju/'hoansi to represent the Northern San and the ≠Khomani and Karretjie to represent Southern Khoe-San). To investigate the divergence time of Mbuti to other non-Khoe-San groups, we chose East African Hadza, which are also a hunter-gatherer group and show less evidence of admixture with pastoralist groups compared to other hunter-gatherer groups in the data set (Biaka and Sandawe), but we provide evidence below of possible gene flow between the Hadza and San that could influence this estimate. In the version of our dataset where admixed individuals had been excluded, the Ju/'hoansi, Karretjie, ≠Khomani, Mbuti and Hadza have the least evidence of admixture from neighbouring pastoralists of the hunter-gatherer groups in our data set (Figures S17, S19, and S20) and thus these populations were used for all following analyses.

5.2.3 Divergence between San and Mbuti

We estimate a divergence time between Ju/'hoansi and Mbuti of $\hat{T} = 0.083$ (95% ML CI: 0.075-0.091) (Table S9). This quantity is measured in units of time appropriate to the coalescent effective population size over the relevant time period in Ju/'hoansi. Assuming a generation time of 28 years and an average N_e of 10,000-30,000 individuals in the Khoe-San, this would correspond to a chronological date between 46,000 and 140,000 years ago ($T \times 2N_e \times g$, where g is the generation time), or for an N_e of 21,000 individuals (*11*) 97,600 years ago (95% ML CI: 88,200-107,000). While based on ascertained SNPs, this range is compatible with a recent study using shotgun sequencing data where the Yoruba-San population divergence time was estimated to 67,000-164,000 years ago (*7*). It is also compatible with the results of two other studies based on sequence data that also allowed for gene flow in their models (*11, 12*), which estimated the divergence time between San and other populations to 53,000-187,000 years ago (*12*), and 108,000-157,000 years ago (*11*).

5.2.4 Divergence within the Khoe-San relative to their divergence to other groups

Interestingly, we find that the divergence between Northern San (represented by Ju/'hoansi) and Southern San (represented by ≠Khomani (A) and Karretjie) is $\hat{T}$=0.029 (95% CI: 0.021-0.035), approximately 1/3 as far back in time as the divergence of Khoe-San from other populations, suggesting population separation within the Khoe-San that dates back approximately 35,300 years (95% CI: 25,600-42,600) under the same assumptions on N_e and generation time as above. While gene flow with Bantu-speakers could possibly influence this estimate, we found a similar estimate when using only the Southern San individual most distant from Bantu-speakers in the PCA (≠Khomani (B):19; $\hat{T}$=0.027, (95% CI: 0.020-0.035; Table S9) and only Karretjie (Table S9).

While ascertainment bias makes it difficult to accurately estimate effective population size using the SNP data, we can scale divergence times within the Khoe-San with divergence to the Mbuti, for which chronological estimates are available (see above). Since no significant evidence for changes in the effective population size in the Northern Khoe-San lineage has been found (*12*), these coalescent time estimates are expected to scale approximately linearly with chronological time. To investigate the possible robustness to ascertainment bias in these relative estimates, we also applied a range of MAF thresholds to each population (see above), and took the average estimate of $\hat{T}$ over 10 repetitions to circumvent the increased sampling bias when more SNPs were filtered out. We find that population divergence times between the Northern and Southern San are between 1/3 and 2/5 of the time of the first divergence from other populations. This relative estimate is robust to MAF threshold and choice of populations for scaling (Figure S38), and is in the range of the divergence between Mbuti and Hadza measured in the Mbuti time scale. However, we caution that a recent reduction in effective population size of the Mbuti (*12*) might inflate this estimate, together with more complex gene flow between the populations (see below).

Furthermore, within the Southern San (≠Khomani (A) and Karretjie) we find evidence suggesting a divergence approximately 1/10 of the time back to the Khoe-San divergence from other populations (Figure S38), suggesting that divergence even among relatively closely related Khoe-San populations could reach back on the order of 10 thousand years.

5.2.5 Comparison with F_{ST}

The genealogical approach has the advantage of allowing the estimation to focus on one branch of a divergence model, ignoring genetic drift that has occurred in parallel lineages. However, to facilitate comparisons and validate our approach we also estimated the relative divergence time between Southern and Northern San (Ju/'hoansi and Karretjie) using an F_{ST}-based approach. F_{ST} can be re-scaled to be proportional to the total genetic drift that separates two populations T_{a-b} (measured in units of $2N_e$ generations) as follows

$$T_{a-b} = -\log(1 - F_{ST}). \tag{7}$$

We can thus use F_{ST} to estimate a relative divergence time as

$$\frac{T_{Ju/'hoansi-Karretjie}/2}{T_{Ju/'hoansi-Mbuti}/2}. \tag{8}$$

We find that the intra-Khoe-San divergence also in this case is approximately 1/3 of the time back to the ancestral population of all modern humans. While this is likely an underestimate due to the recent decrease in N_e in the Mbuti (and could also be affected be gene flow), it broadly confirms our findings from above.

5.2.6 Discussion

To our knowledge, three previous studies have used genetic data to estimate a population divergence time between the San and other human populations (*7, 11, 12*), two of which allowed for gene flow after isolation (*11, 12*). Here, despite being limited by our use of ascertained SNP array data, we were interested in the scale of divergence between the newly studied groups and other more well-studied groups. We were reaffirmed in that we had found an approach that is relatively robust to ascertainment bias in that our estimates of genetic divergence were similar to previous studies (see above), but can not exclude some remaining influence.

Population divergence time between Khoe-San and other populations: Assuming a generation time of 28 years, that the ancestors of the Khoe-San were isolated at a single point in time from the ancestors of other groups (including the Mbuti), and that the harmonic mean of the effective population size in the Ju/'hoansi is 21,000 individuals, we estimate a divergence of about 100,000 years (>88,000 years). We identify two main factors that could affect this estimate: (*i*) gene flow between the Khoe-San and other human groups more closely related to the Mbuti, and (*ii*) calibration of mutation rates affecting the effective population size inferred by previous studies (*11, 12*). Both these factors are expected to lead to an underestimation of the divergence time. Previous studies have indeed inferred low rates of gene flow between the Khoe-San and other African populations such as the Mbuti (*11, 12*), and the slightly lower estimate on the coalescent scale obtained here

(0.075-0.091 compared to approximately 0.1 in previous studies (*11, 12*)) could be a reflection of not taking gene flow into account. Moreover, recent studies have cast doubt on mutation rates assumed of most studies to date e.g. (*119*). If mutation rates are overestimated, the effective population size assumed here could be an underestimate, leading us to also underestimate the chronological population divergence time. A third factor that, in contrast to the other two, lead to an overestimation of the population divergence time of the Khoe-San to other groups is admixture from more basal ('archaic') human population lineages (*120*). An important direction of future research is thus to investigate the extent of ancestral structure in these populations, but assuming that such contributions are relatively minor (*120*), we suggest that the majority of the ancestry present Khoe-San groups today is from a population that diverged from a population lineage leading to most of the ancestry of all other modern humans at least 100,000 years ago. This is compatible with *Helicobacter pylori* (HP) strain variation suggesting a coalescence time of *hpAfrica2* (the strain associated with Khoe-San) with all other HP strains at 88,000-116,000 years ago (*121*). Furthermore, the Khoe-San associated mtDNA haplogroup, L0d, coalesces around 100,795 (±10,317) years ago (the divergence time of L0d, however, is deeper, 152,384 (±12,698) years ago) (*5*).

Population divergence time between Northern and Southern Khoe-San: In contrast to the divergence time between the Khoe-San and other human groups, no previous genetic studies have attempted to estimate a population divergence time between different Khoe-San groups. Scaling time according to the magnitude of genetic drift in the Ju/'hoansi, the most well-studied San group, we estimate that the divergence between Northern- and Southern San groups is approximately 35,000 years ago (>25,000 years ago), under the same assumptions as above. We furthermore find that the ratio between this date and the earlier divergence from other populations is about 1/3 for a range of thresholds on MAF. A similar time frame (between 32,000 and 47,000 years ago) was obtained for the coalescence of the northern and southern isolates of *hpAfrica2* (the Khoe-San associated *Helicobacter pylori* strain) – the southern isolates were associated in this case with the ≠Khomani group and the northern isolates with the !Xun and Khwe groups (*121*). Moreover, archaeological evidence of San material culture has been found to date back approximately 44,000 years (*122*).

This estimate is also confounded by the two main factors mentioned above, both expected to result in underestimation of the divergence time, the magnitude of which depending on calibration and/or the rate of gene flow between Northern San and Southern San. However, an additional factor that could potentially lead to an overestimation of divergence time is Bantu-related admixture in the Southern San or the Northern San. For instance, parts of the genome of admixed Southern San individuals (introgression of Bantu-speaker ancestry) will encompass the oldest population divergence among modern humans (≥100,000 years ago) and lead to overestimation of the population divergence. The extent of this effect is unknown, but our analysis suggest that the individuals with the least evidence of admixture gives consistent results and that our inference here is relatively robust to such admixture. Since many details of the demographic history and contact between Khoe-San groups are not well-known, the estimated chronological dates should be seen in this context, however, the data clearly points to deep structure between Southern and Northern Khoe-San groups.

5.3 Evidence for gene flow between sub-Saharan hunter-gatherer groups

The concordance test-based framework employed above is designed to capture the major historical flow of ancestry of sub-Saharan African populations, but we also note inconsistencies compared to a basic tree model, indicating gene-flow. Table S8 and Table S10 report the results of several D-tests between the two discordant categories which, under the hypothesis of no gene-flow, is expected to be consistent with 0. In addition, we applied the allele frequency based estimator of D (*7, 123*), which has more power but is potentially affected by changes in genetic drift through population history.

We used equation S15.2 of (*7*) with the chimpanzee allele denoting the ancestral state to perform a number of tests of topologies including sub-Saharan hunter-gatherer groups, excluding loci at which the minor allele frequency was less than 10% in any of the three populations other than chimpanzee.

The sub-Saharan African hunter-gatherer groups in our data could be divided into three main groups: Southern African Khoe-San, Central African Pygmy groups, and East African hunter-gatherers (Hadza and Sandawe). From each of these three groups, we focused on the Ju/'hoansi, the Mbuti and the Hadza as the populations with the least evidence of recent admixture from neighbouring pastoralists. Broadly, we find evidence of gene flow between the East Africans and both the Khoe-San and Central African Pygmy groups. There was evidence of gene flow between the Ju/'hoansi and non-Pygmies (Hadza, Sandawe, Maasai, and Yoruba), but the strongest connection is with the Hadza, both in terms of the magnitude of D and in direct tests of the topology *(chimpanzee,(Ju/'hoansi,(Hadza,[Sandawe, Maasai, or Yoruba])))* (Table S10). Prehistoric connections between East, Central and Southern African hunter-gatherers has also been proposed based on Y-chromosome haplogroup A and B data (*66, 124*).

Similarly, we find evidence for gene flow between the East African Hadza and Central African Mbuti Pygmies, for example in the topology *(chimpanzee,(Mbuti,(Hadza,Yoruba)))*. Thus, assuming that the broad topology inferred above reflects the major population history of the groups, our analyses provide evidence of gene-flow between East African hunter-gatherers and both Southern African Khoe-San and Central African Pygmy groups. Since the Khoe-San show more affinity to the Hadza compared to the Mbuti in a direct test, there is presently no evidence of gene flow between the Mbuti and the Khoe-San, and any such gene flow must have had less impact than the gene flow between East Africans and the Khoe-San.

We see evidence of inconsistencies in four different tests of the form *(chimpanzee,(Ju/'hoansi, (Mbuti,[Hadza, Sandawe, Maasai, or Yoruba])))*. Since the D-statistic deviates the greatest from 0 in the case of the Hadza, followed by the Maasai, Sandawe and finally Yoruba, this could possibly be due gene flow between the Hadza and the Ju/'hoansi that produces indirect similarity to East and West Africans over the Mbuti (correlated with the distance from *e.g.* Hadza). An alternative explanation could be that the Mbuti have ancestry from a human group that is more basal (in terms of divergence) than the Khoe-San, which could produce an apparent excess of ancestral alleles. This has also been suggested by recent studies (*6, 120*).

6 Summary statistics and genetic diversity

6.1 Haplotype diversity

Heterozygosity at individual SNPs on standard genotyping arrays is highly subject to ascertainment bias, and does not accurately capture genetic diversity in Africans (*9*). To circumvent this and exploit the ability of recombination to generate genetic diversity, we analyzed haplotype heterozygosity and haplotype richness (number of alleles), at haplotype windows from 1kb to 100kb in size (with 1kb increments) in each population using the phased data (global set of individuals, 270k SNPs, no admixed individuals).

We divided the genome into non-overlapping windows of size S, and for each window:

1. We downsampled the total number of chromosomes to n=14 (note: a new random draw from the population was performed in each window).

2. We excluded SNPs with minor allele frequency ∼10%.

3. Windows with more than 5 SNPs were randomly down-sampled to 5 SNPs. Windows with less than 5 SNPs were discarded.

4. We computed expected heterozygosity (*'Haplotype heterozygosity'*) H using each unique haplotype i as a separate allele (total number of alleles k) as

$$H = 1 - \sum_{i=1}^{k} p_i^2 \tag{9}$$

5. We counted the number of distinct haplotypes (*'Haplotype richness'*). Since this value is for a draw of 14 randomly sampled chromosomes no sample size correction is needed for comparison between populations.

This procedure was repeated 10 times for each population and the average over replicates was taken for the two summary statistics.

Haplotype diversity in human populations have been investigated previously (*3, 10, 125–128*) but not simultaneously on a genome-wide and worldwide scale. We find that haplotype richness and heterozygosity (Figures S40 and S53, Table S11, and Table S12) mirror previous analysis of genetic varation showing gradients of microsatellite variation (*19*), haplotype variation (*125*), and linkage disequilibrium (*3*), which all decrease as a function of distance from Africa. In addition, we are able to replicate previously observed (*126, 127, 129*) regional gradients of variation over Asia (Figure S53) as well as Europe and the Levant (Figure S57). However, no clear gradient or pattern of variation appears in Africa (see also Figure 3 in the main text), and our results suggest that the greatest diversity is often found in populations that have a great deal of ancestry related to the Bantu-expansion. We also find that our estimates for pairs of populations known to be from the same ethnic groups are highly similar (Figure S41).

6.2 Private haplotypes

Using the same strategy for creating haplotypes, we also computed the number of private haplotypes (*3, 125, 128, 130*). We note that the sampling strategy has a great impact of levels of private haplotypes. For instance, an extreme example here is the two ≠Khomani samples that represent one population (which may be easy enough to handle by merging the two samples), but there are more difficult cases, such as ≠Khomani and Nama. However, a few things can be noted from the distributions of private haplotypes. The greatest levels of unique haplotypes are found in the two Pygmy groups Mbuti and Biaka, which is not surprising given the isolation of these two groups (both from other populations and from each other). Second, despite the dense sampling of Khoe-San groups in this study, the /Gui and //Gana show an extraordinary level of private haplotypes (Figure S42 and S43). Third, the patterns seen here are similar to the results of (*128*), who investigated a much smaller set of SNPs (a few thousand SNPs) for a large sub-Saharan African panel of populations.

6.3 Linkage Disequilibrium

We measured linkage disequilibrium (LD) by the correlation coefficient, r^2, between all pairs of SNPs. All populations were sub-sampled to 14 chromosomes (randomly without replacement) for each pair of SNPs. Pairs of SNPs containing at least one SNP with a minor allele frequency (MAF) less than 10% (corresponding to 1 of 14 chromosomes) were removed. For each population, we computed the mean r^2 and the mean distance between pairs of SNPs for all SNP pairs within bins of size b; a bin centered on distance x contains all pairs of distinct SNPs in the interval $(x - b/2, x + b/2]$. The effect of the choices of MAF-cutoff and bin size has previously been investigated and found to have basically no impact on observed patterns of LD and relative levels of LD (*128*).

6.3.1 Southern Africa dataset (2.3M SNPs)

LD, as measured by mean r^2 values for SNP pairs in physical distance bins, declines with increasing physical distance between SNPs for all 9 populations (Figures S44 and S45). The Bantu-speakers (South Africa) has the lowest level of LD, followed by the Khwe, !Xun and ≠Khomani (A). The Nama had the greatest level of LD, followed by the /Gui and //Gana, Herero, Ju/'hoansi, and Karretjie.

6.3.2 Global dataset (270k SNPs)

We computed r^2 based on the ∼270k SNP set, see Figures S46 and 3 (main text).

6.3.3 Inferring effective population size from LD

There are a few approaches to infer N_e from LD information, including (*131, 132*), and we chose an alternative direct and transparent approach in order to capture information of N_e from LD. To summarize information on N_e from the pattern of LD decay over increasing physical distance, we use an approach to simulate data from a constant size model (for various choices of population size) and fit the simulated LD decay curves to empirical LD decay curves. We define here "effective population size" as a single parameter represented by the average genetic drift effect during the history of the population. That is to say, we infer the constant population size that best reproduce some genetic pattern observed in the data. We used the r^2-decay curves shown in Figure 3C (main text) for comparison to data simulated under standard neutral model with constant population sizes. We simulate SNP data using the software `ms` (*117*). The simulations were conducted to mimic the empirical data (populations with fewer than 7 individuals were excluded from this analysis) as close as possible by:

- sampling 7 diploid individuals from each population,
- removing SNPs with Minor Allele Frequency (MAF) < 0.1,
- compute fit of curves for physical distances between pairs of SNPs up to 100kb,
- use the same binning strategy for computing r^2 as described above.

We simulated data on a grid of θ (40, 60, 80, . . . , 1000) and ρ values (20, 40, 60, . . . , 1000) and we used 10 million pairs of SNPs to obtain smooth r^2 curves for each combination of θ and ρ. the fit between population curves and simulated curves was measured by computing the $-\log_{10}$ of the relative Mean Square Error (MSE). Six examples of such fits are showed as heat-maps in Figure S47. The fit is independent of θ. By averaging the ρ values leading to the best fit over each θ value, and using a recombination rate of 1.5×10^{-8} (*15*), we compute the population sizes of the constant standard neutral model that best approximate each population (Figure S48). The "effective population sizes" estimated by this method are greater than the effective population sizes derived from methods based on genetic diversity. In particular, the result suggests that African effective population sizes span between 11,000 and 17,000 (except for Hadza which has a much smaller effective population size). The black curve shows the values of $-\log_{10}$(MSE). Populations have different degrees of agreement to the constant model. Hadza, for example, has a low $-\log_{10}$(MSE), suggesting a strong departure from the constant population size model, while the Colombian population seem to be well approximated by such a model.

The advantage of using LD is mainly to provide us with effective population size estimates that are independent of the mutation rate. When more whole-genome sequences are available, a very similar similar approach can be performed, together with statistics such as genetic diversity, and it will be interesting to compare the effective population size estimates using the two different types of statistics.

6.4 Runs of homozygosity

Long runs of homozygous haplotypes (*RoH*) are a frequent characteristic of human genomes. Such runs of homozygosity are created when identical parental haplotypes are inherited. The length of the haplotype is dependent on the number of generations since the last common ancestor of the haplotype. Therefore very long runs are caused by recent inbreeding and shorter runs are due to shared ancestry a longer time back in the past. When *RoH* stretches are subset into classes of increasing lengths (i.e. 0.5-1Mb, 1-2 Mb, 2-4Mb, 4-8Mb, etc.), the shortest run classes could be used as a representation of past population history. The shorter the "run class" the longer time back in the past it represents. It has been observed previously, that if only the shorter classes are considered (ruling out recent inbreeding), African populations have the shortest average "cumulative *RoH* (*cRoH*) per individual" and the average *cRoH* (of the short class) of populations increase with distance from Africa (*133*). Previous studies have shown the short class *RoH* to be a good predictor of N_e in populations (small average *cRoH* in populations indicates large N_e), while the longer classes represents past consanguinity in populations (*10, 133*).

To calculate the prevalence of long runs of homozygosity (*RoH*) for population groups, the different groups were first down-sampled to seven individuals before a 10% minimum allele frequency cutoff was applied. An exception is the Bantu-speakers (Southern Africa - HGDP), who had a smaller sample size and were thus represented by only six individuals. *RoH* was calculated with `PLINK` v1.07 (*32, 134*) under the following parameters: sliding window of 5Mb, one heterozygote per window allowed and a minimum *RoH* length of 500kb and 25 SNPs per window. To ensure that the length of an *RoH* is not artificially increased due to local low SNP density, the minimum density was set to 50 kb/SNP and the maximum gap between two consecutive SNPs was set to 100 kb. These parameters were similar to parameters used previously in *RoH* studies (*10, 133*). `PLINK` was run for each population separately. Since phased/imputed datasets were used and each population was screened in a separate *RoH* run (no merging of datasets), the dataset contained no missing data, which was previously noted to be a problem for *RoH* scans (*10*). The *RoH* runs was repeated 50 times, each time sampling (without replacement) a random set of 7 individuals from each population.

Table S13 show a sorted list of the average *cRoH* per population (0.5-1Mb class) averaged over the 50 repeats of sub-sampling for African groups (results depicted in Table S13 are represented as a heat map in main text Figure 2H). The result of one of the 50 repeats is also represented by violin plots in the main text Figure 2D. Figure S49 shows the *cRoH* of 50 repeats of sub-sampling

7 individuals from each population for the global dataset. Figure S50 show the average *cRoH* of classes of different lengths of *RoH* as well as an all-inclusive class (for one of the 50 sub-samples) for African populations. Figure S51 show the number of *RoH* as a function of *cRoH* of populations and Figure S52 the number of *RoH* as a function of *cRoH* for individuals separately (for one of the 50 sub-samples) for African populations.

Two west African groups had the shortest average *cRoH* followed by the East African Maasai (Table S13). The Hadza group had a much higher *cRoH* indicative of a population bottleneck, which was noted before (*10*). The bottleneck in the Hadza is probably not due to recent inbreeding as the Hadza had high *cRoH* in most of the *cRoH* length classes (Figure S50). It was discussed before that the African farmers had much shorter average *cRoH* than the African hunter gatherers (*133*). However, here we see that, while the northern San groups, /Gui and //Gana, Nama, and the two Pygmy groups had generally higher average *cROH*, the Southern San, Karretjie group and ≠Khomani (A) group, are towards the lower end of average population *cRoH*. The Karretjie group had a lower average *cRoH* than several farming groups, for instance all of the the Southern African Bantu-speakers. The study by Kirin *et al.* (*133*), however only included the HGDP Pygmy and Ju'/hoansi (northern San) groups. Populations that experienced recent inbreeding are represented in the larger classes of the barplot (Figure S50) and can be seen to diverge to the right of the diagonal in the plot of number of *RoH* vs. *cRoH* (Figure S51) for example the Khwe, /Gui and //Gana and the Sandawe. However, sometimes only single inbred individuals in a population can skew population *cRoH* values. In the individual plot of *cRoH* against number of *RoH* (Figure S52), one can observe that there are single Khwe, Sandawe and /Gui and //Gana individuals that probably have an effect on their populations' representation in the longer *cRoH* class.

6.5 Summary and comparison of summary statistics and genetic diversity

We assessed the geographic distribution of the four summary statistics (Haplotype Heterozygosity, Haplotype Richness, ρ ($= 4N_e$) estimated from LD, and Runs of Homozygosity (*RoH*)) as well as correlations between them. We find that all four statistics are significantly correlated with distance from East Africa in non-African populations, with Pearson correlation coefficients ranging from 0.72 for *RoH* to -0.94 for Haplotype Heterozygosity (Figures S53-S56). Further, in most statistics we also see inter-regional patterns in Asia and Europe that match those previously observed (*3, 125–128*) (Figures S53-S57).

Finally, we find that the two haplotype statistics and the statistic based on LD are all highly intercorrelated ($R > 0.85$; Figure S58). The *RoH* statistic also displays a non-significant negative relationship with the other statistics, as expected, and the discrepancy probably stems from *RoH* being better at capturing consanguinity in some populations, as well as possibly some effect of ascertainment bias in West Africans (when applying the 10% MAF filter to west African groups around 40,000 more SNPs are retained compared to hunter-gatherer groups, which lead to a higher SNP density that in the end affects the calls of *RoH* – with more *RoH* being called in the less dense population datasets). We do not find compelling evidence pinpointing East Africa (*135, 136*), Southern Africa (*10*), or any other region as having clearly higher diversity that could indicate the origin of an expansion of modern humans.

6.6 Frequency spectrum

We characterize the frequency spectra based on SNP chip data (both the set of ∼270k SNPs and the set of ∼2.3M SNPs described above) and compared to the frequency spectra from sequence data ('The HOMINID' data) (*137, 138*). Both the SNP chip data and the sequence data contain individuals from the same populations and the same sample collection, in some cases even the same individuals. These comparisons can give important information about the effect of ascertainment bias on SNP chip data.

The investigated populations that have a sample size greater than 7 individuals were randomly sub-sampled to 7 individuals (14 gene copies) to allow direct comparisons among populations. The frequency spectra from the sequence data is computed as an average of sub-sampling 7 individuals 100 times. The presented frequency spectra are all based on the autosomes and frequency spectra are relative. Thus, among a total of x SNPs (variable positions for the sequence data), there are x_1 SNPs have 1 derived allele among the 14 sampled chromosomes, ..., x_{13} SNPs have 13 derived alleles among the 14 sampled chromosomes. The relative frequency of category "1" (1 derived allele) is x_1/x, and so on.

6.6.1 Frequency spectra for ∼270k SNPs and ∼2.3M SNPs

Frequency spectra computed from the ∼270k SNPs (Figures S59, S60, and S61). Figure S62 show the frequency spectra for the dense (2.3M) set of SNPs.

6.6.2 Comparison with 'HOMINID' sequence data

The 'HOMINID' data consists of French Basque, Biaka, Han, Mandenka, Melanisian, Papuan, and Ju/'hoansi (denoted 'San' in HGDP and in the HOMINID data. The same populations have been genotyped with the 650k Illumina SNP-chip (*4*) (270k SNPs remaining in our data as described above). We have also genotyped the Ju/'hoansi using the 2.5M Illumina SNP chip, and the Ju/'hoansi can therefore be compared across two SNP-genotyping panels and sequence data. Figures S64 and S63 show the frequency spectra for the HOMINID sequence data and our two versions of SNP-chip data. Figure S65 shows the sum of squared differences between the Hammer sequence data and our SNP-chip sequence data.

6.6.3 Some remarks on the frequency spectra

Populations that are geographically and/or genetically distinct from the ascertained populations (which are known to be European, East Asian and West African populations) display frequency spectra that are less flat, presumably due to the weaker effect of ascertainment bias in these populations. We also note, in accordance with previous observations, that African frequency spectra are less flat (but not for Hadza, likely due to heavy recent bottleneck) compared to non-African spectra (Figure S60). Two main African groups crystallize – roughly corresponding to hunter-gatherers and pastoralists. A third group contains Biaka and Khwe, two groups that show shared ancestry with Bantu-speaking groups (Figure S61). Finally, we note that the effect (measured as the difference between the frequency spectra of a particular SNP-chip set and the frequency spectra from sequence data) of ascertainment bias is substantially lower for the ∼2.3M SNP set compared to the ∼270k SNP set (Figures S64 and S65). Furthermore, by filtering out singletons and doubletons (removing *e.g.* alleles with a minor allele frequency less than some threshold as we do for most analyses in this paper) most of the well-known effect of ascertainment bias (*e.g.* greater diversity in ascertained populations) are equalized and will have little effect for relative population comparisons.

7 Selection

7.1 F_{ST}-based scans

To search for evidence of genomic regions that are particularly differentiated between groups we performed F_{ST} scans across the genome. We excluded individuals with putative evidence of admixture. All analyses were based on 2,293,320 autosomal SNPs. We contrasted the top observed F_{ST}-value in each pairwise comparison with the genome-wide estimate of F_{ST} (see main text). In addition, we made two comparisons between groups of populations, see below.

7.1.1 Khoe-San vs Bantu-speakers

We computed F_{ST} between Khoe-San groups (!Xun, Karretjie, ≠Khomani (A), Ju/'hoansi, Nama and /Gui and //Gana) and Bantu-speakers from Southern Africa (Bantu-speakers (S. Africa) and Herero) across the genome. Figure S66 shows a box plot with the distribution of the F_{ST} values for single SNPs. Tables S15 summarizes the results for the top 20 greatest F_{ST} values. The SNP with the greatest F_{ST} in the entire genome was rs1322784 ($F_{ST} = 0.65$), a SNP that has previously been associated with Asperger syndrome (`http://www.snpedia.com/index.php/Rs1322784`). In addition, the top 20 SNPs contained 3 SNPs from the same region on chromosome 11 (rs4372467 $F_{ST} = 0.62$; kgp31391 $F_{ST} = 0.60$; kgp12532905 $F_{ST} = 0.60$), that harbors a gene, DISC1, involved in neuron activity (`http://www.ncbi.nlm.nih.gov/nuccore/NM_014715`). The ETS1 gene on chromosome 1 also had three SNPs that were among the top 20 greatest F_{ST} values. This gene is a transcription factor that regulate numerous genes.

7.1.2 Nama vs San

The Nama (a Khoe group) is believed to have transitioned from a hunter-gatherer lifestyle to a pastoralist mode of subsistence prior to the arrival of Bantu-speaking pastoralist to the southern parts of Africa, while the other groups (San) kept to a hunter-gatherer lifestyle. We therefore conducted a separate scan looking for regions with unusually high differentiation between the Nama (Khoe) and San groups (!Xun, Karretjie, ≠Khomani (A), Ju/'hoansi and /Gui and //Gana).

Figure S67 shows a box plot with the distribution of F_{ST} values. Table S15 summarizes the results for the top 20 greatest F_{ST} results. Here, we found that 10 of the 13 top SNPs were from a region on chromosome 16 (position 13658506 to 13813289). Figures S68, S69, S70, S71, and 4C-D (main text) show clearly the high F_{ST} values for this region on chromosome 16. No gene was identified within this region. However, the region is located 200kb upstream of the *ERCC4* gene in an active enhancer region. Mutations in the *ERCC4* gene have been linked to Xeroderma Pigmentosum Complement group F (XPF). People suffering from XPF have a 3-fold increased sensitivity to the lethal effects of ultraviolet light. However XPF has mild symptoms comparing to other Xeroderma Pigmentosum subclasses. Cells from XP F patients are slightly UV-sensitive and exhibit low levels of repair initially after UV-irradiation (`http://omim.org/entry/278760`). Accordingly, mutation carriers show slightly high sun sensitivity of the skin. The mild clinical symptoms consist of numerous pigmented freckles and mild skin lesions. Skin cancers in mutation carriers are elevated (*23, 139*). Since the SNPs with elevated F_{ST} values do not directly overlap with the *ERCC4* gene region, a direct link cannot be inferred. The SNPs however are located 200kb upstream of the gene in a region that has been marked as an active binding site to transcription enhancers (H3K27Ac histone mark), through CHiP seq assays in various cell lines.

The SNP with the greatest F_{ST} value that was located inside a gene was contained within the CSNK2A1P gene. Polymorphisms within this gene has been implicated to play an oncogenic role in lung cancer (*140*).

7.2 Inferring genome-local ancestry for regions of interest

We further investigated the region on chromosome 16 in the Nama that showed high F_{ST} values in the Nama vs. San comparisons. To see if this region was possibly introduced by a neighboring population group into the Nama we used a genome-local clustering approach (*24*), to infer the ancestry of local genome regions in the Nama. We compared the Nama individuals to various populations, including Khoe-San populations and Bantu-speaking populations in the dense genotype dataset (∼2.3M SNPs). We ran `CHROMOPAINTER` (*24*) for each of the 14 Nama recipient chromosome 16 haplotypes for 10 iterations. Each of the 10 iterations had 50 prior EM iterations for optimizing the following three parameters; recombination scaling constant (N_e), copying proportions, mutation (emission) probabilities. Each of the eight other (excluding 'Coloured' populations) populations in the dense genotype dataset (∼2.3M SNPs) were considered as a possible donor population (Bantu-speakers (South Africa), Herero, Khwe, Ju/'hoansi, !Xun, /Gui and //Gana, ≠Khomani, Karretjie). Thus each SNP position in the 14 chromosome 16 Nama haplotypes was assigned probabilistically to one of these eight donor populations. The result of assignment from 10 iterations in the 14 Nama haplotypes were summarized by plotting the population assignment fraction (140 possible assignments per SNP for the Nama: 10 iterations x 14 haplotypes). The chromosome-wide assignment of Nama to each of the eight donor populations are shown in Figure S68 and in Figure S69 the assignments to the 6 San donor population is summarized. The main text Figure 4D is similar to Figure S69, except that 200-SNP windows have been used in the main text Figure to smooth lines. Figure S70 and S71 shows enlarged versions of the region of interest for San donor populations separate and summarized into one respectively.

Population assignments via `CHROMOPAINTER` indicated that the region of interest might be due to admixture from the Bantu-speakers (South Africa) group. Population assignment results from 23 populations in the global dataset were less conclusive due to low SNP coverage over the region, however, these analyses also indicated the Bantu-speakers (South Africa) were the most likely donor.

7.3 Searching for selective sweeps based on extended regions of haplotype homozygosity

We scanned the genome for regions of extended haplotype homozygosity using the integrated haplotype statistic iHS (*20*) which can detect selective sweeps that have not yet reached fixation in a population. Following the approach described in (*43*), we chose the top 10 windows from 7 populations (Nama, Herero, and /Gui and //Gana were excluded due to small sample size). Among these 70 windows, those that had at least one $|iHS|$ value among the genome wide top 10 constituted our candidate windows (Table S16).

Peak on chromosome 10 for Ju/'hoansi

A particularly interesting region is found for Ju/'hoansi, it spanned 69.8 to 70.0Mb on chromosome 10 (Figure S72 and S73, and main text Figure 4A) and overlaps the *MYPN* (myopalladin) gene, which is associated with muscle mass growth and the function of striated muscles (*21, 141*), and a polymorphism in the gene has been found to change musculature in mammals (*142*). The iHS peak in this window coincides with a 7SK snRNA, which controls the transcription of many genes, probably also *MYPN*, and suppression of these genes has been shown to cause enlarged muscle cells (*143*). Although the signal for a selective sweep is strongest in the Ju/'hoansi, we also find significant signals in Bantu-speakers (South Africa) and ≠Khomani. The signal can also be detected in other groups, including non African populations, suggesting that the selective sweep is either old or reoccurring. All genes that overlap the candidate region are listed in Table S17.

Peak on chromosome 1 for Karretjie

A strong signal of selection is also found on chromosome 1, in the Karretjie (Figure S74), probably connected to the *HHAT* (hedgehog acyltransferase) gene, which encodes a signaling molecule that governs SHH (Sonic hedgehog) signaling (*144*). The SHH and the hedgehog signaling pathway is one of the most studied developmental pathways and plays a key role in the development of a wide range of organs, morphology and the nervous system. The *HHAT* gene is furthermore involved in the organization and morphology of the developing embryo and has also been implicated in melanoma (*145*) and Crimean-Congo hemorrhagic fever (*146*). All genes that overlap the candidate region are listed in Table S18.

Peak on chromosome 6 for Karretjie

The *iHS* signal peaks inbetween the genes *PRL* and *HDGFL1* and are upstream of both (*PRL* on minus strand and *HDGFL1* on positive strand). Of these two genes, *PRL* seems the more likely candidate for a selective sweep. *PRL* encodes the anterior pituitary hormone prolactin. This secreted hormone is a growth regulator for many tissues, including cells of the immune system. It may also play a role in cell survival by suppressing apoptosis, and it is essential for lactation. Moreover, (*147*) suggested an important role for *PRL* in regulating adipose tissue metabolism during lactation. All genes that overlap the candidate region are listed in Table S19.

Peak on chromosome 6 for ≠Khomani (A)

A strong signal of selective sweeps is found on chromosome 6 (around 27.4Mb) near, but not overlapping, the major histocompatibility complex (MHC, located between 29.9 and 33.1 Mb). This peak is particularly pronounced in ≠Khomani and Karretjie (Figure S76 and S77 and main text Figure 4B), and constitutes the most prominent peak across the entire genome and among all investigated populations. All genes that overlap the candidate region are listed in Table S20. There are several genes associated with the immune system surrounding the peak, including *PRSS16* and *POM121L2*, which have been suggested to protect against infectious diseases. The strong signal is unique to Southern Khoe-San (Figure S76 S77 and main text Figure 4B), who had early and extensive contact with European colonists (in contrast to northern Khoe-San). The contact led to exposure to new and introduced infectious diseases and historic reports note, for example, several waves of smallpox epidemics in the years 1713, 1755 and 1760 during which a large proportion (up to 90%) of the Cape Khoe group died, while others fled to the interior spreading the infection (*18*). Such a drastic population reduction caused by disease could leave footprints in the genome, potentially as the strong signature of selection around some immune genes that we observe in the Southern Khoe-San, but the signature could also be caused by adaptation to general disease exposure that is known to increase with the transition to a sedentary lifestyle.

Peak on chromosome 1 for Khwe

This peak spans a 2.6 Mb region and appears to consist of more than one peak perhaps reflecting several selective sweeps (Figure S78). All genes that overlap the candidate region are listed in Table S21. Picking out a single gene among the 20 genes is difficult but if only protein coding genes with known function are considered, there are 11 genes remaining. Among these, genes that stand out include 3 genes associated with mental phenotypes (*SLC30A10*, *BPNT1*, *RAB3GAP2*), 3 genes encoding mitochondrial proteins (*IARS2*, *MARC1*, *MARC2*), and one gene involved in the immune system (*HLX*). Mutations in *RAB3GAP2* are associated with Martsolf syndrome and Warburg Micro syndrome (*148*). The diagnostic symptoms of this syndrome include mental retardation, cataracts, short stature, primary hypogonadism, and minor digital and cephalic abnormalities. Other features included brachycephaly, short philtrum, low posterior hairline, scoliosis, talipes valgus, flat feet, and lax finger joints (*149*).

Peak on chromosome 4 for Khwe

The location of the peak suggests that the selected gene should be one of the genes numbered 3 to 9 in Figure S79. Among these, number 6 is a psesudo gene while number 5 is of unknown function. The remaining 5 genes are two immune system genes (*IL2* and *IL21*), a gene whose nonfunctionality is the cause of Bardet-Biedl syndrome type 12 (*BBS12*), and a fibroblast growth factor gene (*FGF2*) implicated in a multitude of physiologic and pathologic processes, including limb development, angiogenesis, wound healing, and tumor growth. *NUDT6* is the antisense gene of *FGF2* and may regulate *FGF2* expression. The region harboring *IL2* and *IL21* has also been implicated in celiac diseases (*150*) Bardet-Biedl syndrome is a genetically heterogeneous, autosomal recessive ciliopathy characterized by progressive retinal degeneration, obesity, cognitive impairment, polydactyly, and kidney anomalies (*151*) and the link to obesity seems especially strong for *BBS12* (*152*). All genes that overlap the candidate region are listed in Table S22.

Peak on chromosome 10 for Bantu-speakers (South Africa)

All genes that overlap the candidate region are listed in Table S23. The location of the peak suggests that the selected gene should be one of the genes with number 2 to 7 in Figure S80. Among these, 3 genes are particularly interesting. Mutations in *RAB18* are associated with Warburg micro syndrome type 3 (*153*). This provides a link to *RAB3GAP* which has also been implicated for Warburg micro syndrome (see peak on chromosome 1 for Khwe above). The phenotype of Warburg micro syndrome type 3 (similar to the Martsolf syndrome) include microphthalmia, microcephaly, pericardial edema, delayed jaw formation, a reduced overall body size, and a general developmental delay. *MKX* codes for a homeobox protein and studies in mice suggest that this protein may be a regulator of tendon development (*154*). *BAMBI* negatively regulates TGF-beta signaling and is potentially related to pigmentation (*155*) as well as adipogenesis (*156*).

7.4 Other regions of phenotypic interest

Other regions which do not show up in selection scans but have been linked to specific phenotypes in previously studied populations are discussed in this section. There are two such regions of interest discussed in more detail here, which are also mentioned or discussed in the main text in context of lifestyle factors.

7.4.1 Lactase Persistense

The first region is a region on chromosome 2, which contains the *LCT* gene and has been associated with Lactase Persistence (LP) in adults (*157*). A specific East African LP SNP (*158*) and putative underlying haplotype (*159*) have been identified previously and high frequencies of the SNP and haplotype have been reported in the East African Maasai group (*158, 159*). Here we report the frequencies of the putative LP haplotype in all African populations in our 270 K dataset (Table S24). We observe that the Maasai group has the highest frequency of the putative LP haplotype (40.4%), followed by the Nama (35.7%). Indeed previous studies reported that Lactase Persistence is common in Nama (50% in adults) compared to San groups (e.g. <10% in Ju/'hoansi (*18*)). The pastoralist Bantu-speaking populations, on the other hand, have lower frequencies and the West African Niger-Kordofanian farmers lack the putative LP haplotype. This reinforces the hypothesis that the first introduction of pastoralism into Southern Africa (through the Khoe associated culture) was from a group with East African ancestry. The second wave of pastoralists moving into Southern Africa (Bantu-speakers that expanded from West Africa) did in fact not carry this LP haplotype at high frequencies. Low frequencies in Eastern and Southern African Bantu-speakers might be due to admixture with local pastoralist East Africans and Southern African Khoe groups.

7.4.2 Muscle Performance

The second phenotype of interest has to do with muscle performance. A selection signal (particularly strong in the Ju/'hoansi) associated with the *MYPN* gene, that play a role in muscle mass growth (*21*), is discussed in the main text. Another interesting gene (also discussed in the main text) associated with muscle performance is the *ACTN3* gene (*22*). A specific SNP (rs1815739 C) in this gene has been directly associated with fast twitching muscles and elite athletic performance. It has been noted that Africans carry this SNP in high frequencies (*160*). For our African dataset we found that Khoe-San populations, in fact, carry the highest frequencies of this functional SNP (Figure S81).

8 Pastoralism in Southern Africa

Taking together various lines of evidence presented in the main article as well as preceding supplementary sections, we propose that pastoralism was first introduced to Southern Africa via cultural diffusion and later via demic diffusion. The Nama – a Khoe group that traditionally had a pastoralist lifestyle in contrast to the hunter-gatherer lifestyle of the San groups – shows great genetic similarity to the descendants of Southern San groups, such as ≠Khomani and Karretjie (main text Figs. 1 and 2). The Nama individuals contain a small distinct genetic ancestry component that is shared with East African groups, in particular the pastoralist Maasai (16% averaged across individuals assuming 11 clusters, main text Fig. 2B) and not the Sandawe or Hadza. This "East African" component is also present at lower levels in the two ≠Khomani groups (9.3% and 4.5%) and the Karretjie (4.5%) – although less evenly distributed among individuals compared to the Nama –, but basically absent (<1%) from the !Xun, the Ju/'hoansi and the /Gui and //Gana.

Furthermore, formal tests of "treeness" (*7, 123*) were also highly significant in favor of gene flow in the topology ([chimpanzee or Mandenka], Maasai),(Ju/hoansi, Nama) ($D > 3\%$, Z-score > 8). We also investigated the evidence of gene-flow between East African groups (Hadza and Maasai) and the Khoe (Nama) and San (Ju/'hoansi) by applying `TREEMIX v.1.03` (*116*) to these populations (Figure S82). We rooted the ancestry graph using Ju/'hoansi and Nama, allowed for two migration edges and computed SEs based on blocks of 2,000 SNPs. Allowing more migration edges resulted in a migration from the base of the Maasai branch to its tip, and thus did not change the inference. The resulting tree suggests gene flow from the Maasai to the Nama (point estimate of 27.9%), and separate and lower level of gene flow from the Ju/'hoansi to the Hadza (point estimate of 10.7%). While these precise quantities may be uncertain, this illustrates that although there is evidence of low amounts of gene flow from East African hunter-gatherers and San groups, the affinity between the Nama and pastoralist Maasai is a separate and more distinct signal.

Further supporting this association; lactase persistence is common in Nama (50% in adults) compared to San groups (e.g. <10% in Ju/'hoansi (*18*)) and, interestingly, we found that the Nama carries high frequencies of a haplotype putatively associated with lactase persistence in the Maasai (*159*). This haplotype is rare in pastoralist Southern African Bantu-speakers, suggesting that the genetic basis of lactase persistence in the Nama has an East African origin (Table S24).

Thus – consistent with some linguistic hypotheses on the origin of the Khoe-Kwadi language-group (*51*) and with Y-chromosome haplogroup similarities (*104*) between East African and Khoe-San groups – our genome-wide population-genetic data suggest that the pastoralist Nama originate from a Southern San group that adopted pastoralism, with limited introgression from a group that migrated to Southern Africa from East Africa, potentially bringing pastoralist practices, but who were later assimilated by the resident populations.

The population structure analyses of sub-Saharan African populations showed a clear signal of the range expansion of West Africans who today live in most areas of Africa (main text Figs. 1 and 2). South-eastern Bantu-speaking groups arrived in Southern Africa around 1,000 years BP as part of the "Bantu expansion" driven by the development and spread of agriculture (*105*). At the time of European colonization, the eastern parts of Southern Africa were occupied by Bantu-speaking groups but rock-paintings and archaeological remains indicate that Khoe-San groups had previously occupied these regions (*106*). On the contrary, in the western parts of Southern Africa, many indigenous Khoe-San groups still exist today and Bantu-speakers arrived only a few hundred years ago. In south-eastern Bantu-speakers (South Africa), the distinct fraction of Southern Khoe-San ancestry (main text Fig. 2) indicates assimilation of local Khoe-San populations during the "Bantu expansion", whereas the south-western Bantu-speakers (Herero) display a much smaller fraction of Khoe-San ancestry. There is also ample evidence of gene flow into the Khoe-San groups from Bantu-speaking groups in essentially all sampled Khoe-San groups (main text Fig. 2).

To summarize, the different patterns of admixture in Southern African groups have implications on whether new cultures and technologies, such as farming practices, spread in Southern Africa via cultural or demic diffusion (*2, 161, 162*). Pastoralism, first appearing in Southern Africa about 2,000 years ago (*161–163*), was potentially introduced by a small group of migrants from East Africa that were assimilated by local hunter-gatherers, who changed to a pastoralist lifestyle. The second introduction, via the "Bantu expansion", led to a much more dramatic impact on the indigenous peoples of Southern Africa, which can be seen in the genomes of current Southern Africans.

9 Scan for evidence of selective sweeps in the ancestors of modern humans

We devised an approach for searching for selective sweeps in the ancestral population of two extant populations by detecting patterns of high derived allele frequencies in both populations (Figure S83). In contrast to targeting the ancestral population of two populations, we could look for unusually high derived allele frequencies in a single population, but the top outliers would many times be more recent selective events. If a selective sweep occurred in the ancestral population, derived allele frequencies at linked SNPs will be swept to high frequency or possibly fixed. When the two descendant populations become isolated and selection at the locus subsides, allele frequencies are once again expected to be subject to random genetic drift, but the effect of the sweep may be detected for quite some time. However, two alternatives to ancestral selection for producing high $aPBS$ values would be either convergent selection or more recent selection in both populations mediated by gene flow.

To search for evidence of selective sweeps in early modern humans, we use the F_{ST}-based framework of (*164*) to estimate the ancestral population branch statistic ($aPBS$) as the 'ancestral branch' in the topology (chimpanzee,(Khoe-San, Bantu-speakers)). We estimated pairwise F_{ST} between each population using eqn. 5.3 in (*107*) and the transformed F_{ST} to the coalescent time scale as

$$T_{a-b} = -\log(1 - F_{ST}) \tag{10}$$

and define the $aPBS$ statistic (*164*) as:

$$aPBS = \frac{T_{KhoeSan-chimpanzee} + T_{BantuSpeakers-chimpanzee} - T_{KhoeSan-BantuSpeakers}}{2}. \tag{11}$$

Bantu-speakers were represented by Herero and Bantu-speakers (South Africa) (n=27 individuals), whereas Khoe-San where represented by Ju/'hoansi, !Xun, /Gui and //Gana, Karretjie, ≠Khomani (A), and Nama (n=73). Admixed individuals were excluded. We computed an average aPBS statistic for windows extending 10 SNPs to each side of each focal SNP, obtaining estimates for a total of 956,579 SNPs (SNPs for which F_{ST} or $aPBS$ were undefined due to monomorphy were excluded). The mean aPBS over all loci was -0.0212 and the standard deviation was 0.27. We define 'peak regions' as consecutive runs of SNPs with $aPBS$ values over 0. We excluded regions with an average of less than 1 SNP per 50 kb, obtaining a total of 25,851 peak regions ranging from 1 to 900kb (shown in Figure 4F in the main text).

9.1 The top candidates for selection in early modern humans

Figure S84 shows the $aPBS$ values across the human genome and Table S25 indicates the top 20 identified candidate regions for selection in the ancestors of modern humans. The top candidate for selection in early modern humans is located in a region immediately upstream of the *ROR2* gene involved in regulating bone and cartilage development, and a gene (*SPTLC1*) involved in hereditary sensory neuropathy (main text Figure 4F, Table S25; Figure S85). Mutations within *ROR2* are known to cause autosomal recessive brachydactyly type B (shortening of digits) and Robinow syndrome, a skeletal dysplasia with a set of symptoms that include short stature and limbs, characteristic facial morphology and vertebral segmentation. Furthermore, *ROR2* is up-regulated by *FOXP2* (*165*), the only gene presently known to be involved in speech and language disorders with Mendelian inheritance (*166*), and which appears to have undergone accelerated evolution in the human lineage (*167*). The second greatest $aPBS$ value (main text Figure 4F, Table S25; Figure S86) is observed immediately upstream of *SULF2* that encodes a signaling enzyme that regulates cartilage development (*168*), and phenotypes associated with mutations in *SULF2* include skeletal malformations (*169*) and distorted brain development (*25*). Notably, three of the top five candidate regions for selection in early modern humans contain genes that have been found to be involved in aspects of skeletal development (*ROR2*, *SULF2* and *RUNX2*). The largest of all 25,851 regions (~900kb) (main text Figure 4F, Table S25; Figure S88) comprises the gene *RUNX2*, which is implicated in a syndrome known as craniocladial dysplasia. Interestingly, *RUNX2* variation has been associated with phenotypic differences between anatomically modern and archaic humans (*7*), such as frontal bossing, clavical morphology and a bell-shaped rib cage (*26*) as well as regulating the closure of the fontanel which is crucial for brain expansion (*27*). The region spanning *RUNX2* was also identified as the 13th best candidate for selection in early modern humans in an analysis using the Neandertal draft genome (*7*). Furthermore, some morphological features, such as short stature and increased space between the eyes (hypertelorism), are observed in both craniocladial dysplasia and Robinow syndrome cases.

The two other genes in the top 5 candidate list for putative selection in early modern humans comprise *SDCCAG8* (Figure S87), involved in retinal-renal ciliopathy (*170*) Bardet-Beidl syndrome (*171*), microcephaly (*28*), and obesity/weight loss (*172*). Finally, the fifth candidate, *LRAT* (Figure S89), is putatively associated with Alzheimer's disease (*29*). Including *SULF2*, three of the top five candidate regions are thus associated with neuronal development.

Finally, we note that the 2.3M SNPs used in this analysis do not capture all genetic variation in African populations (in particular the Khoe-San). An even better view of possible selection in the ancestors of modern humans will thus be available from multiple complete genomes.

9.2 TMRCA of candidate regions for selection in two complete genomes

To investigate whether the candidate regions for selection had atypical time to most recent common ancestor (TMRCA) compared to the rest of the genome, we estimated TMRCA of the two complete diploid genomes from a San individual and a Bantu-speaking individual from Southern Africa sequenced by ref. (*9*). We used an alignment with the chimpanzee genome provided by the authors (*9*) and a simple estimator of TMRCA (*173*) and applied it to 5,057 regions of 20 kb each which were separated by at least 5,000,000 bp to attain a view of the genome-wide distribution of TRMCAs between chromosomes of the two genomes. Additionally, we estimated TMRCA in regions extending 10 kb to either side of the top SNP in each aPBS peak (also total 20 kb). We find that while two genomes only offer a limited view, the analysis does not provide evidence for TMRCAs around aPBS peaks to be atypical compared to the rest of the genome (Figure S90).

Figures

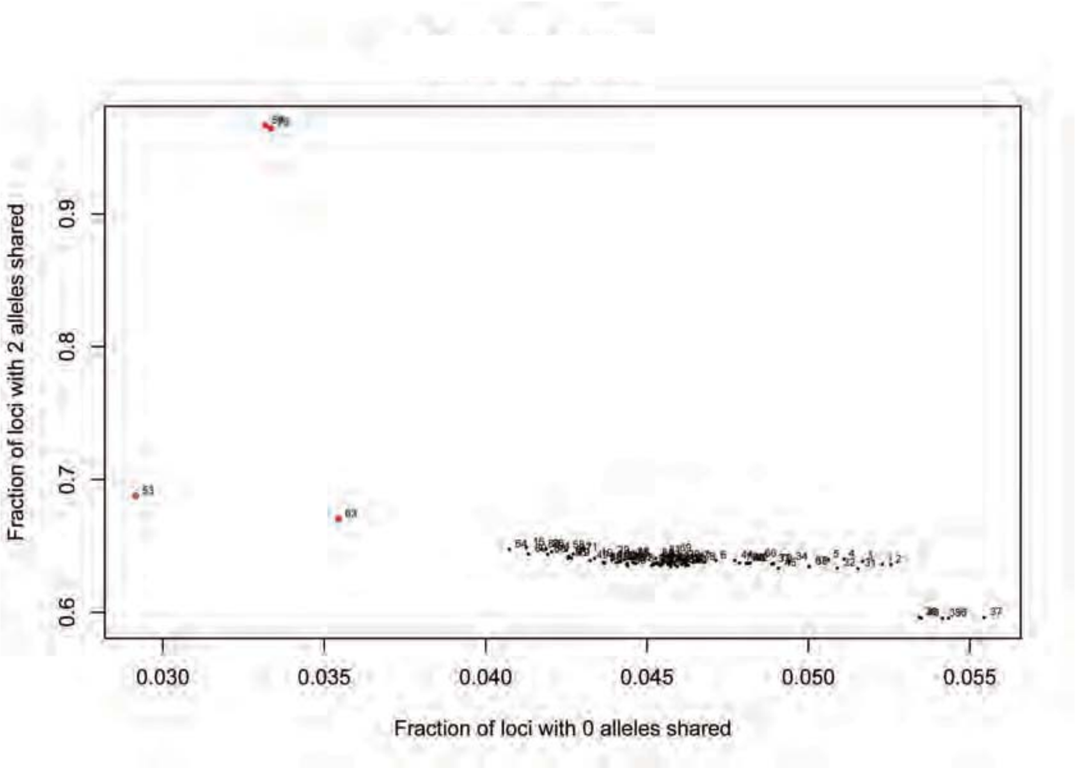

Figure S1: Fraction of SNPs that share no allele (X-axis) and fraction of SNPs that share two alleles (Y-axis) for pairs of individuals from the Ju/'hoansi and the HGDP San. One individual from each pair marked in red were removed from further analysis.

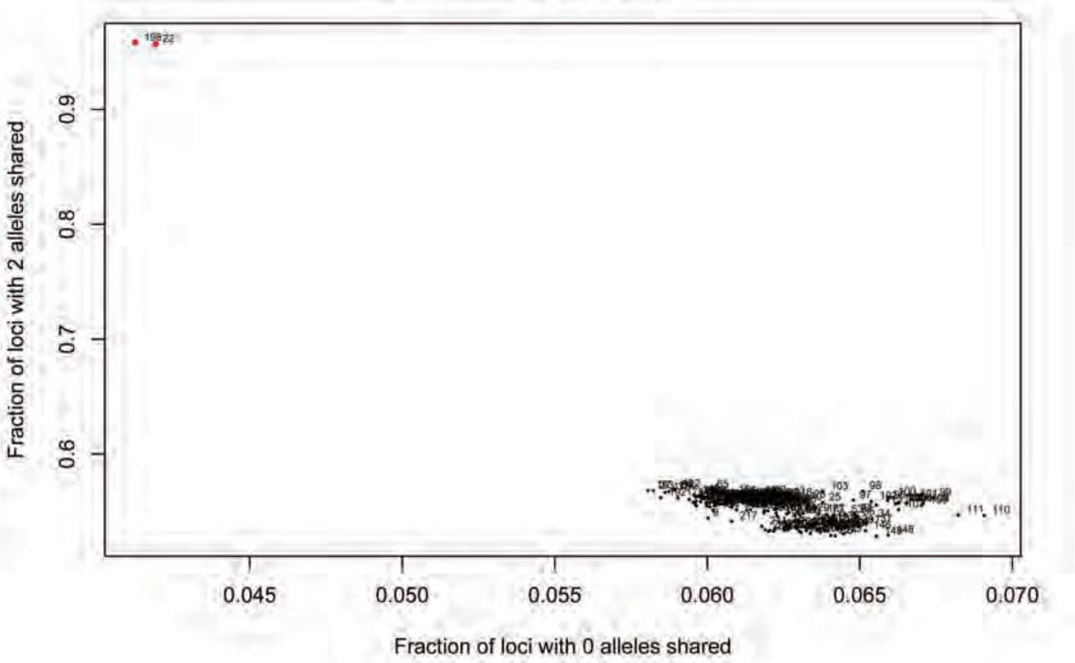

Figure S2: Fraction of SNPs that share no allele (X-axis) and fraction of SNPs that share two alleles (Y-axis) for pairs of individuals from Herero and the Bantu-speakers (Southern Africa + Kenya – HGDP. One individual from each pair marked in red were removed from further analysis.)

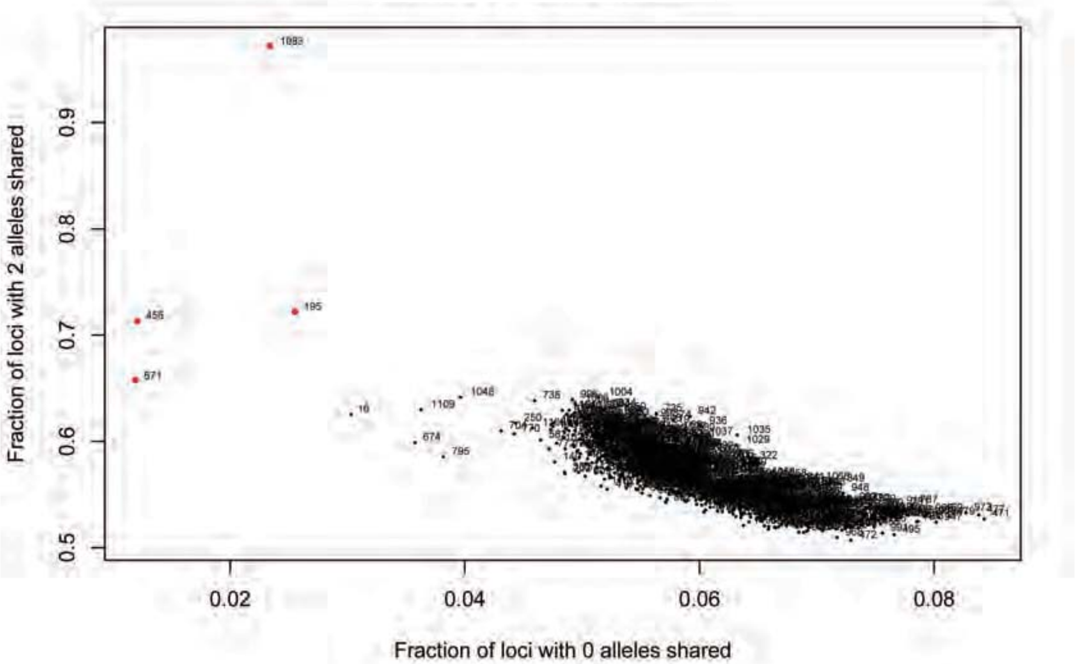

Figure S3: Fraction of SNPs that share no allele (X-axis) and fraction of SNPs that share two alleles (Y-axis) for pairs of individuals from the ≠Khomani (A) and the ≠Khomani (B). One individual from each pair marked in red were removed from further analysis.

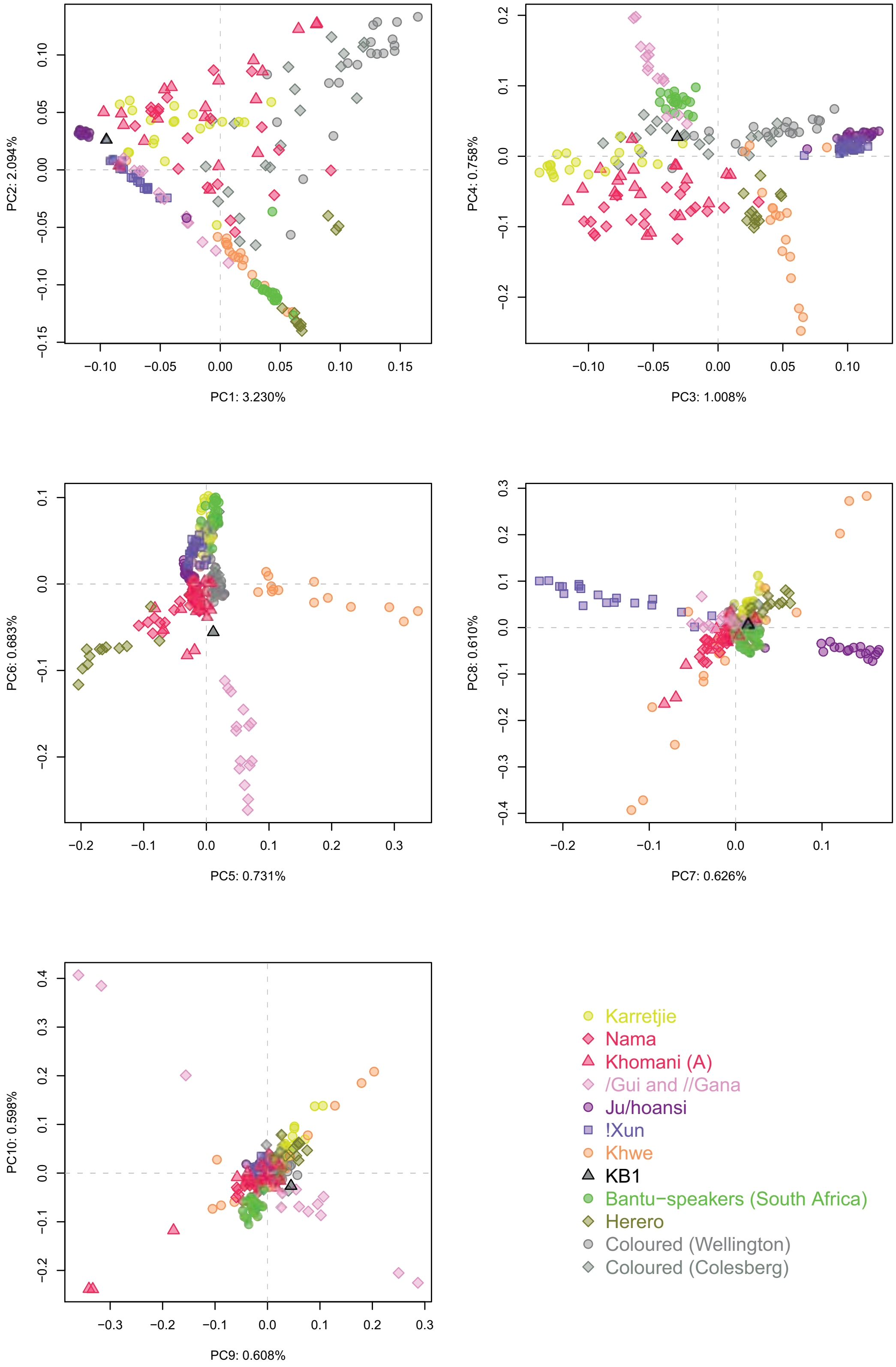

Figure S4: PCA for Southern African individuals based on ∼ 2.3M SNPs.

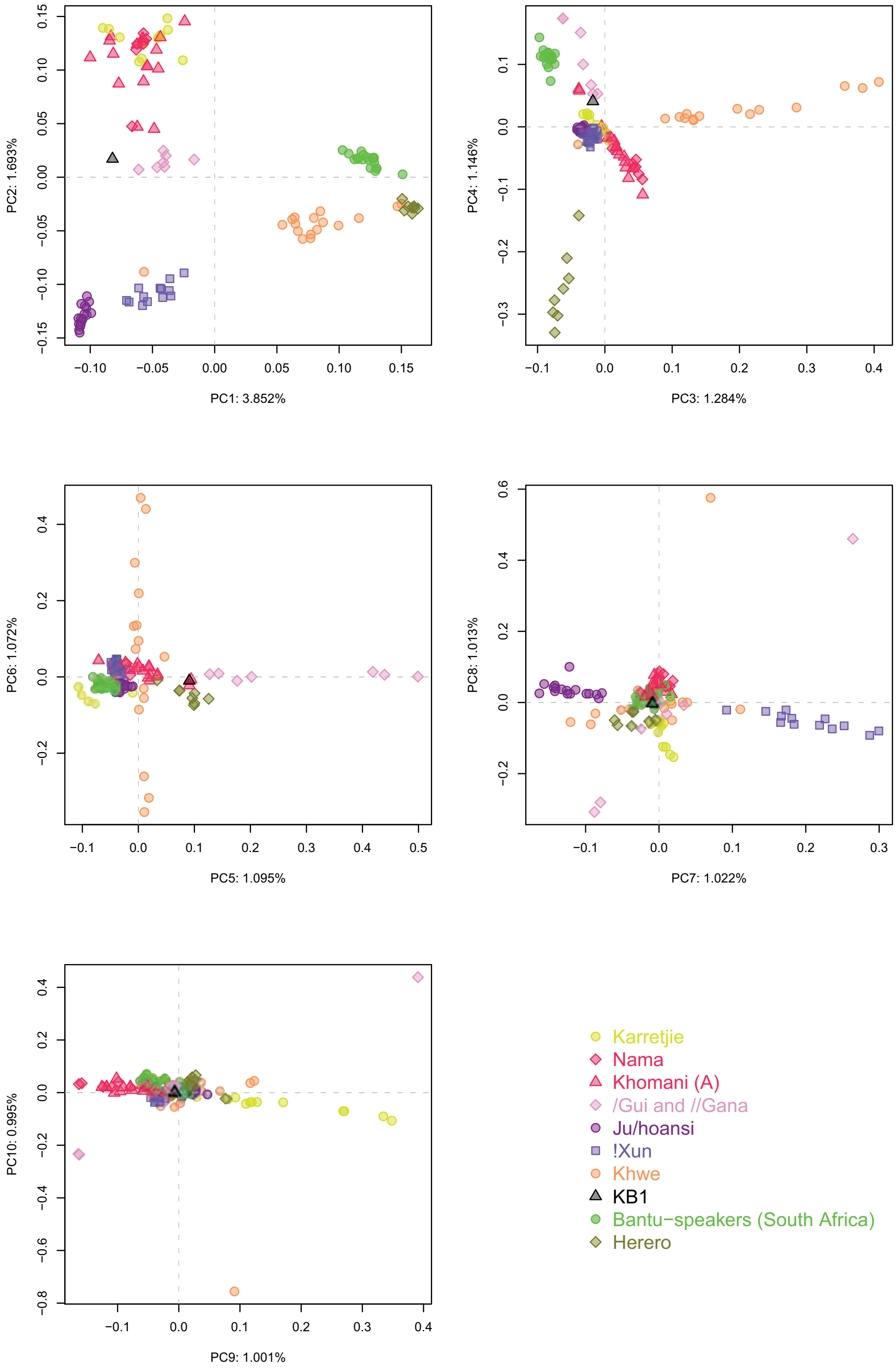

Figure S5: PCA for Southern African individuals (admixed individuals removed) based on ∼ 2.3M SNPs.

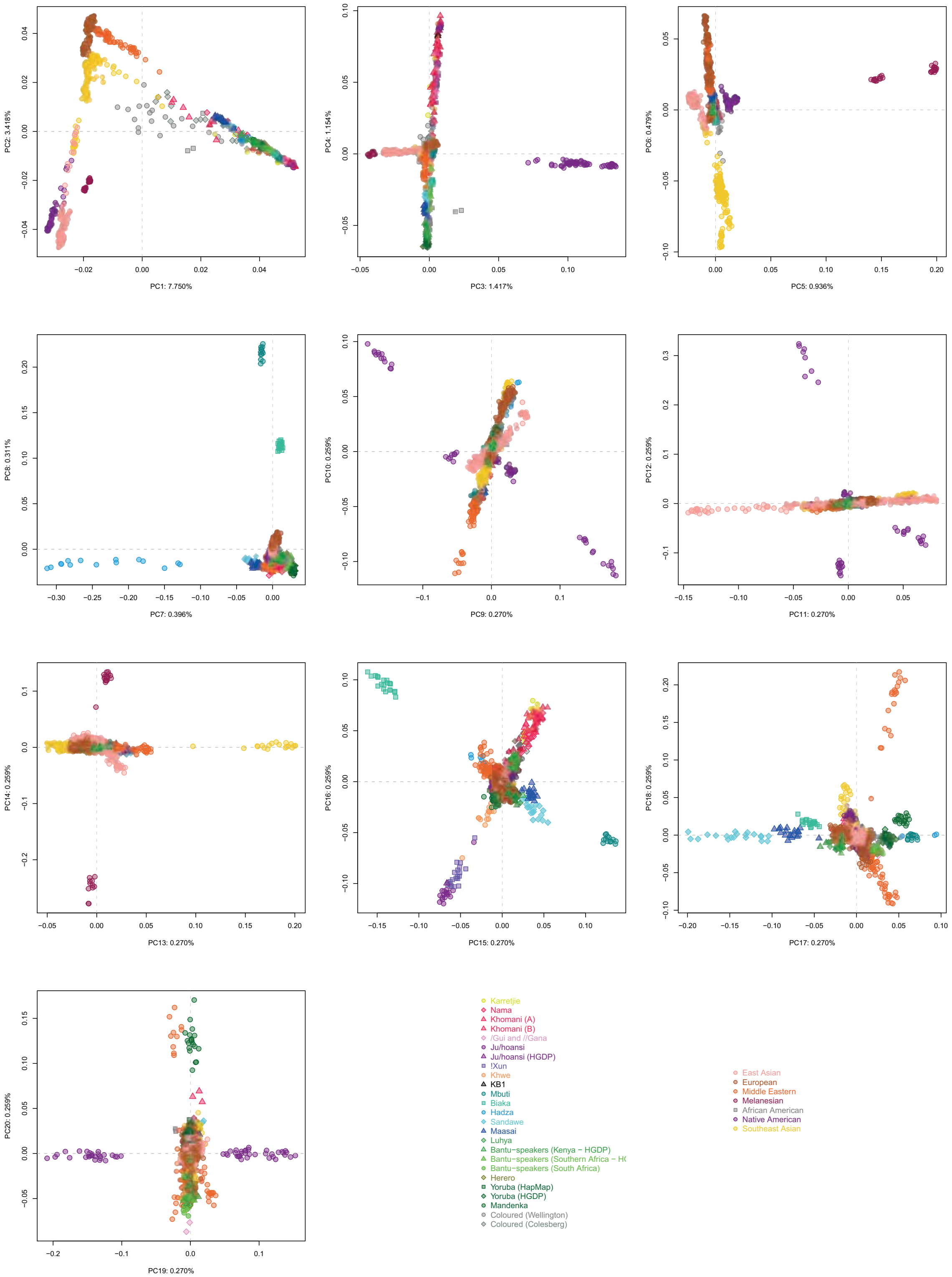

Figure S6: PCA for a world-wide set of individuals based on ∼ 270k SNPs.

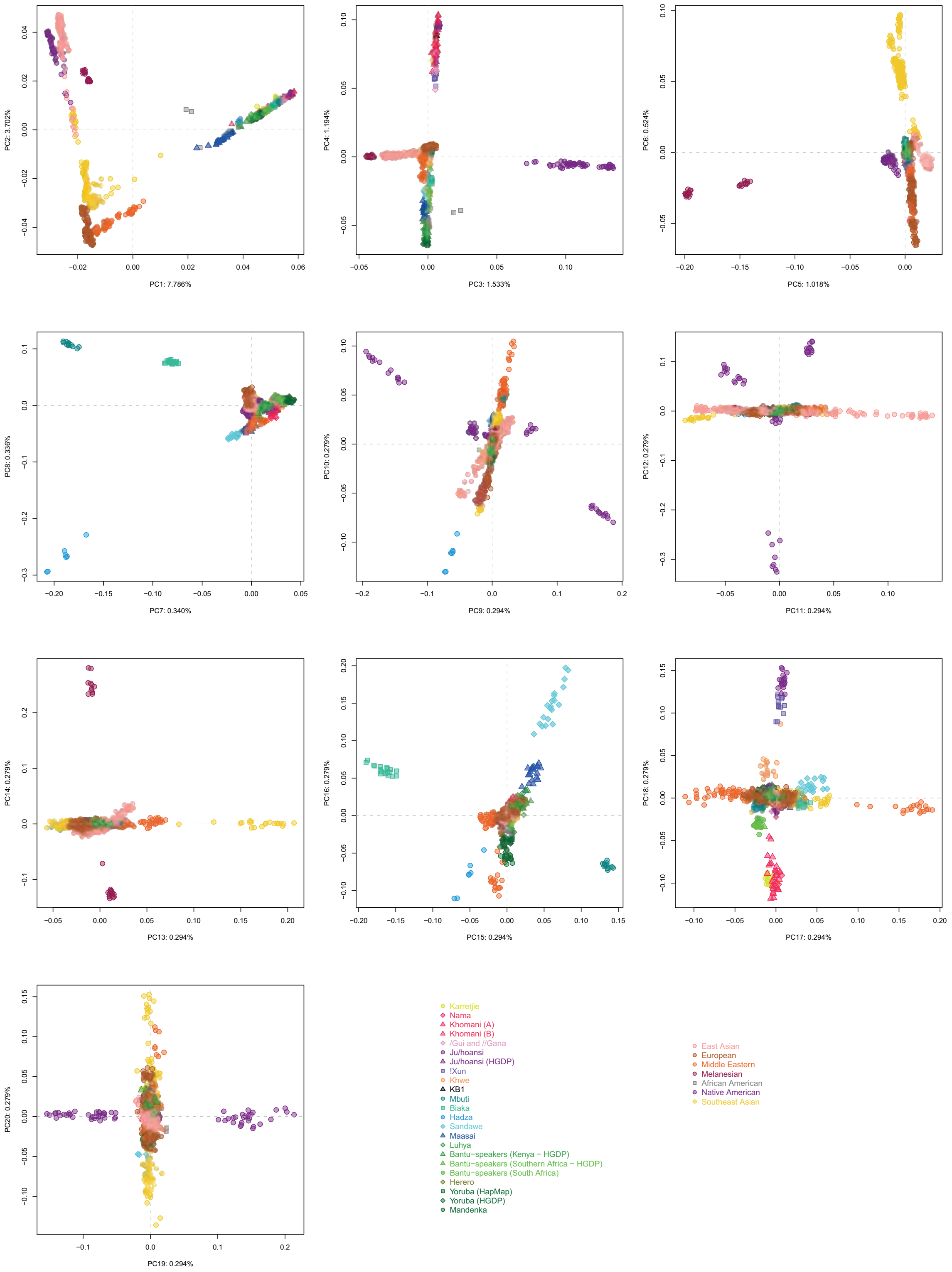

Figure S7: PCA for a world-wide set of individuals (admixed individuals removed) based on ∼ 270k SNPs.

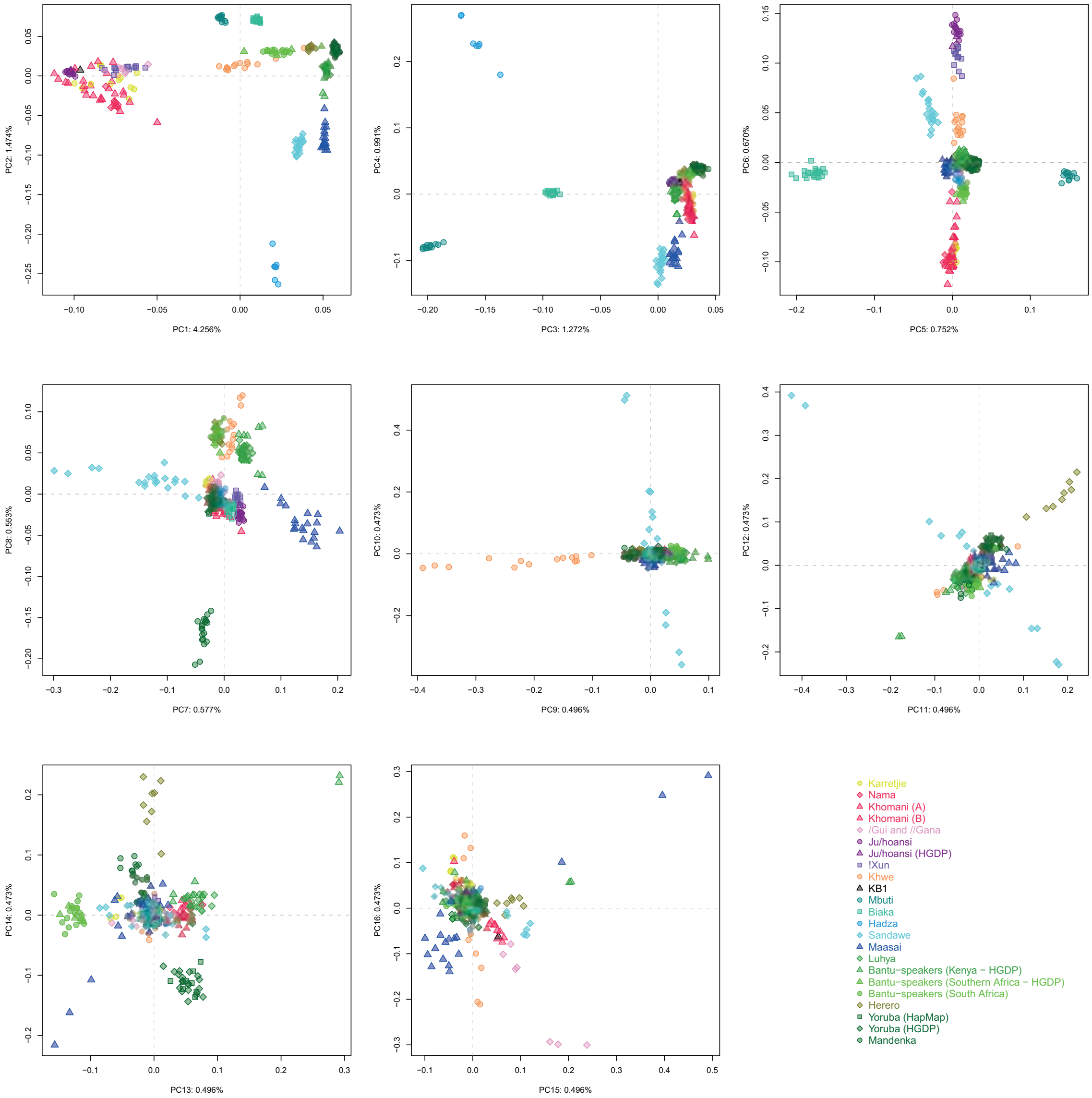

Figure S8: PCA for sub-Saharan individuals (admixed individuals removed) based on ∼ 270k SNPs.

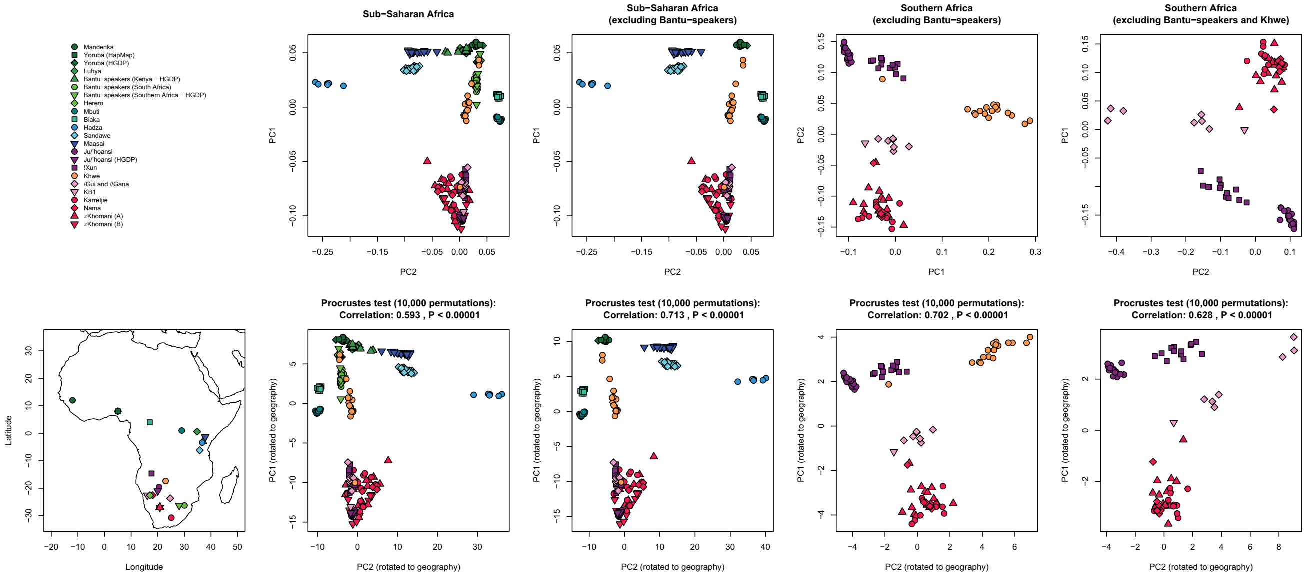

Figure S9: PCA transformed to geography using Procrustes analysis.

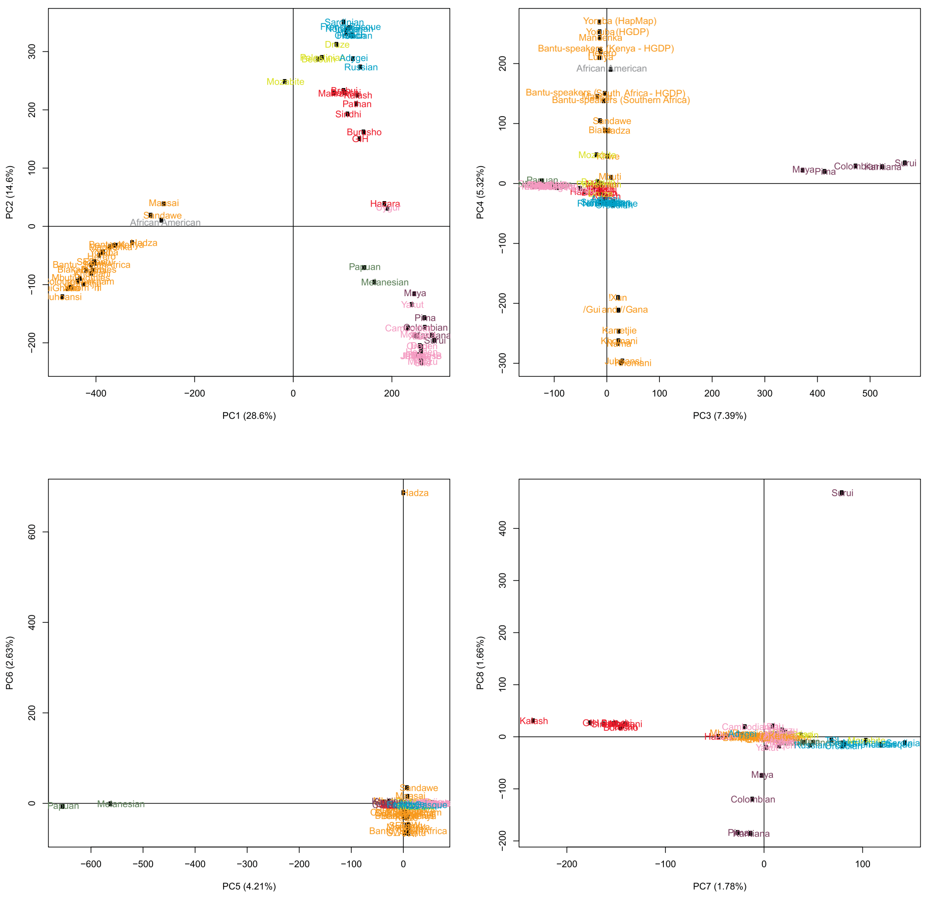

Figure S10: Population PCA for the world-wide set of populations, using ∼ 270k SNPs.

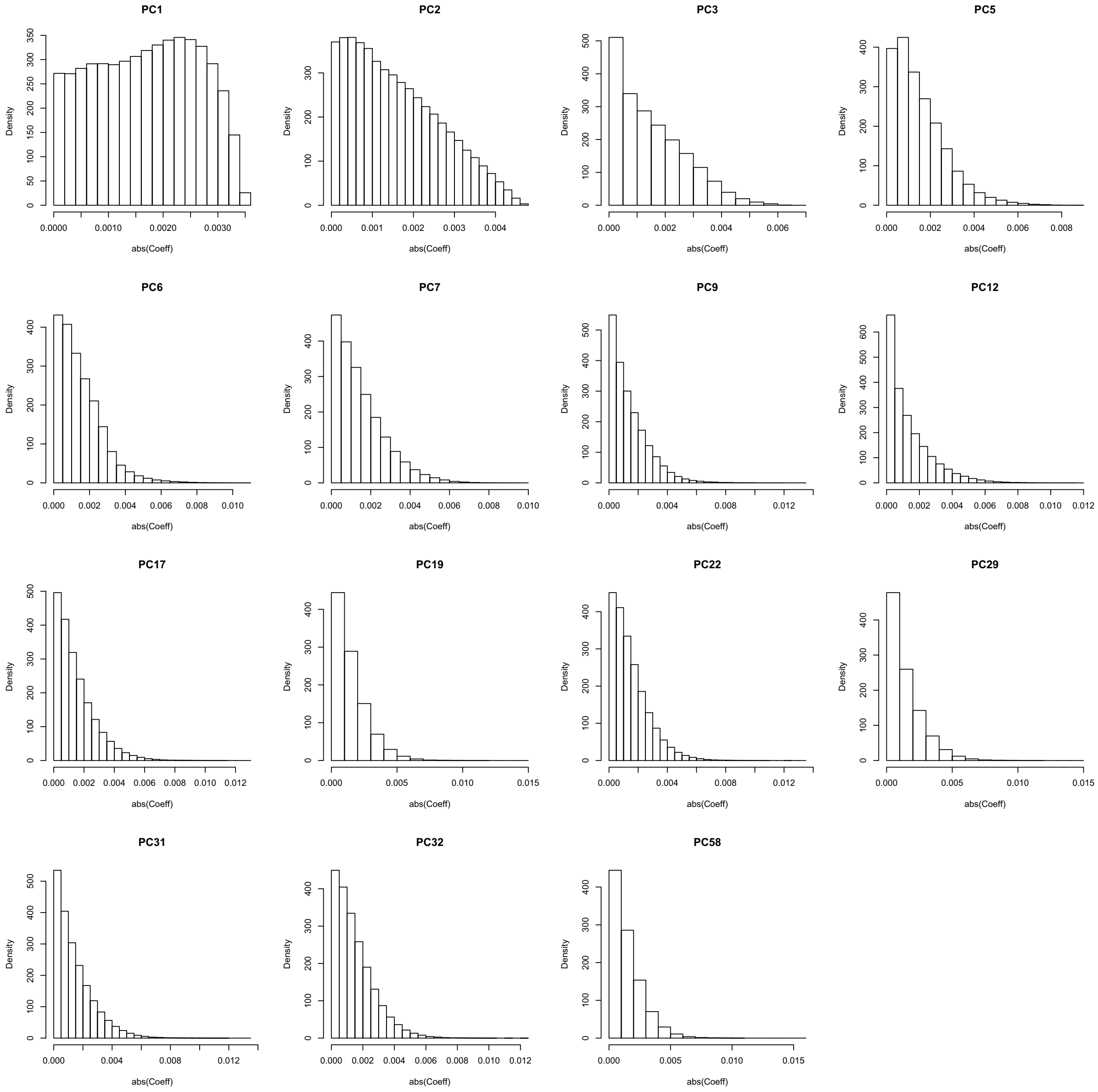

Figure S11: Histograms of the absolute value of the principal components coordinates.

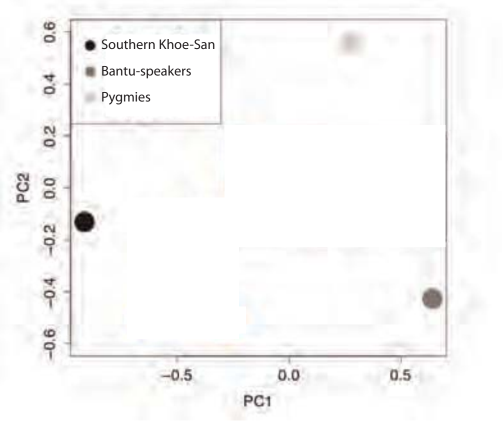

Figure S12: Population PCA for three groups: Southern Khoe-San, Pygmies, and Bantu-speakers.

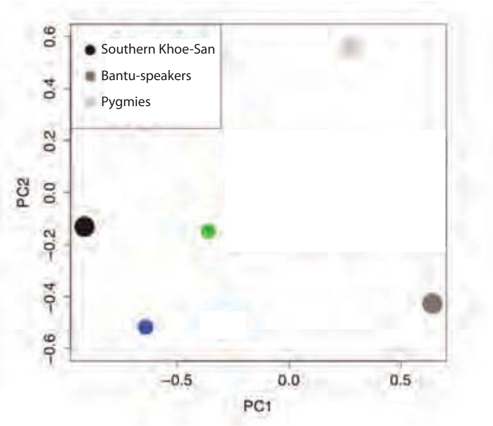

Figure S13: Projecting Khwe (green), and /Gui and //Gana (blue) onto two axes defined by Southern Khoe-San, Pygmies, and Bantu-speakers.

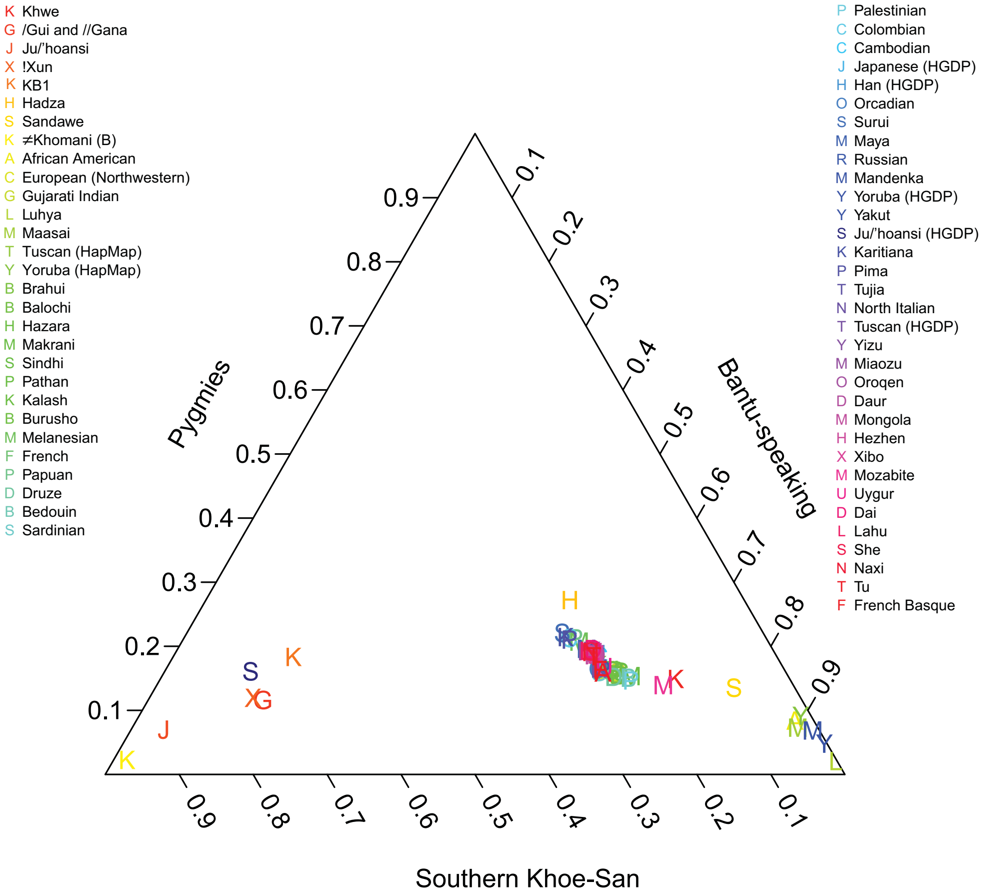

Figure S14: Projecting the global set of populations onto two axes defined by Southern Khoe-San, Pygmies, and Bantu-speakers.

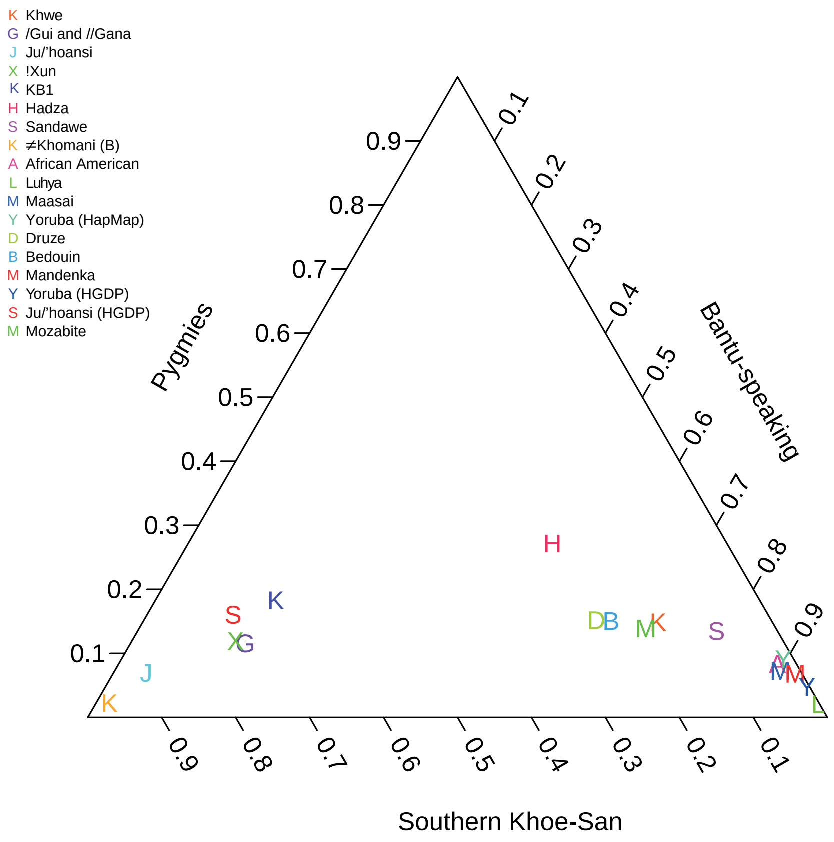

Figure S15: Projecting African and Middle Eastern populations onto two axes defined by Southern Khoe-San, Pygmies, and Bantu-speakers.

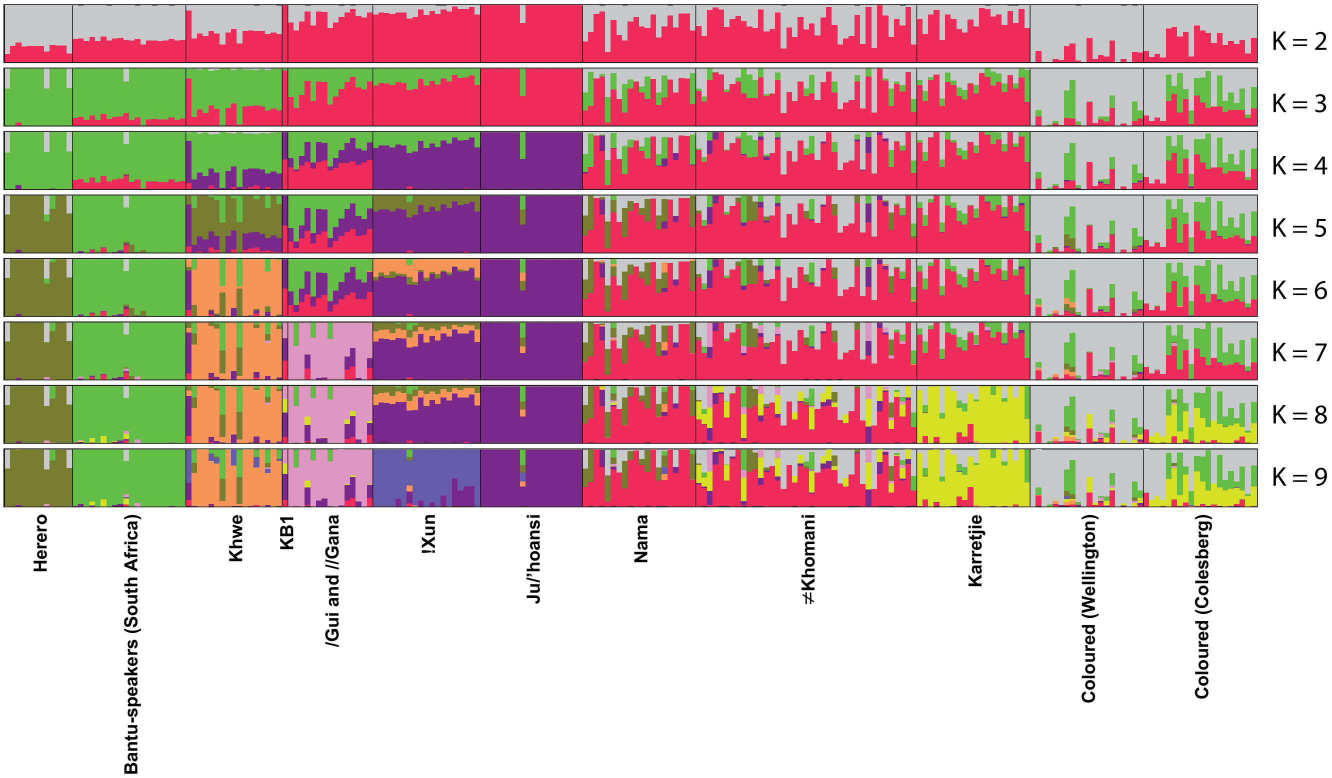

Figure S16: Population structure for Southern African individuals based on $\sim 2.3M$ SNPs and assuming 2 to 9 clusters.

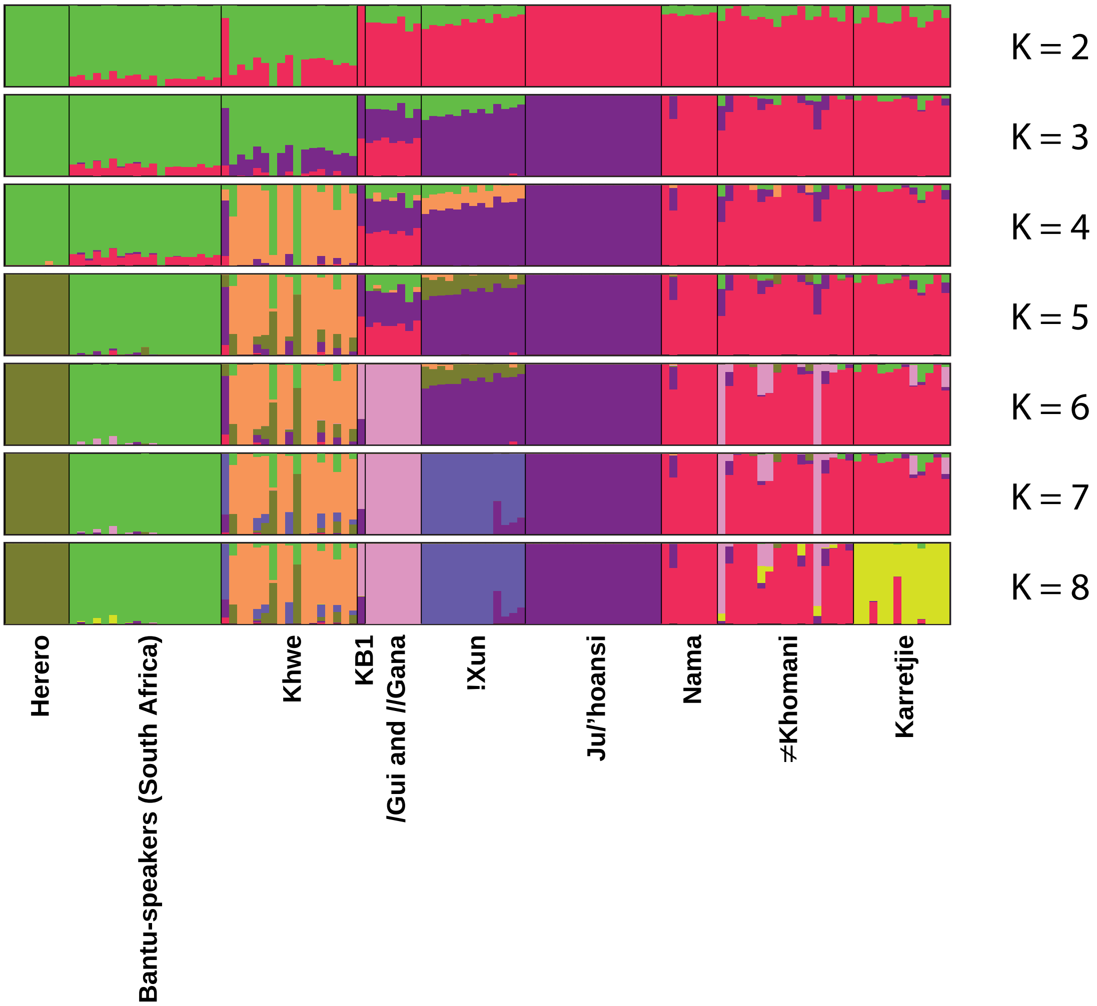

Figure S17: Population structure for Southern African individuals (admixed individuals removed) based on $\sim 2.3M$ SNPs and assuming 2 to 8 clusters.

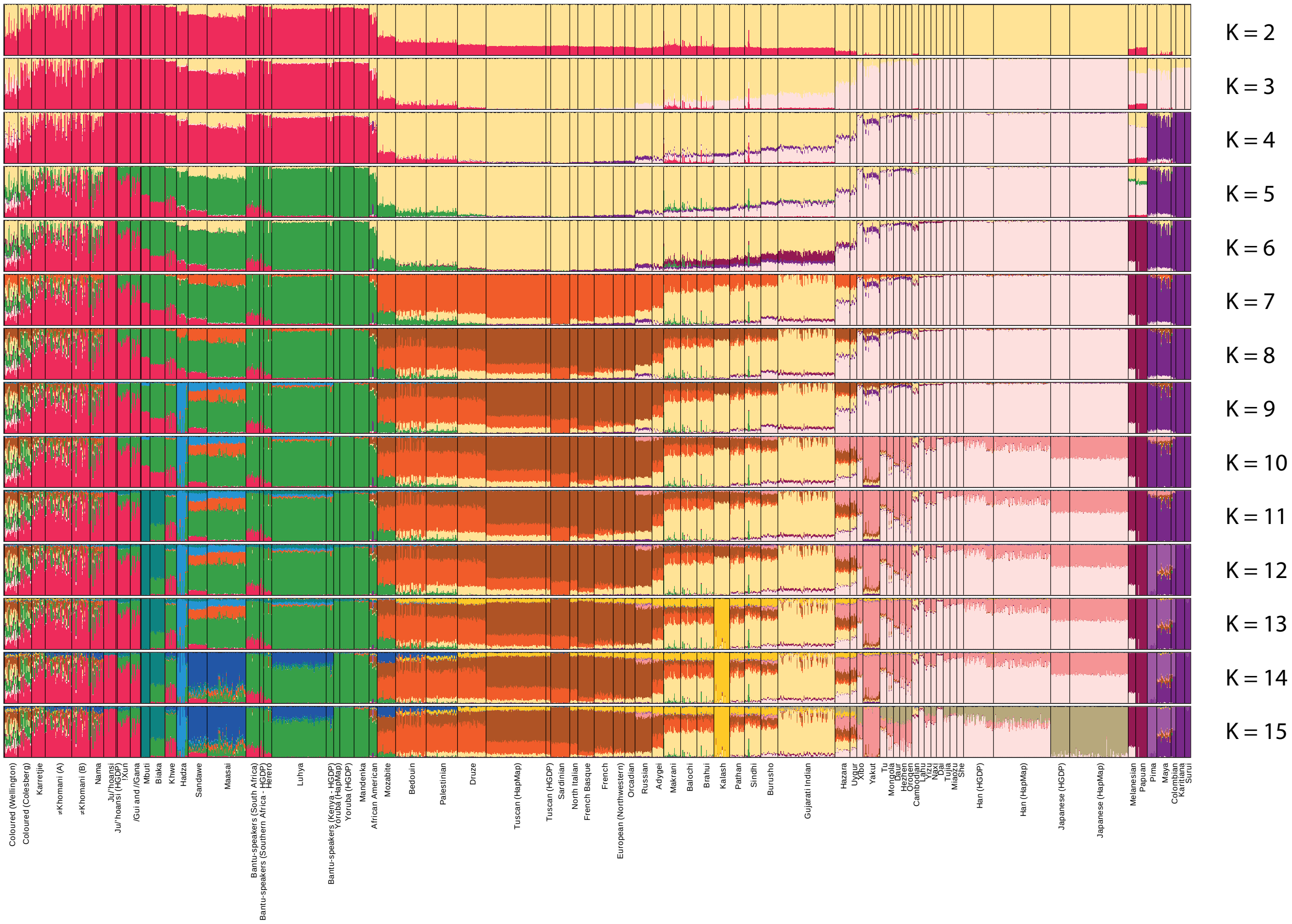

Figure S18: Population structure for a worldwide set of individuals based on ~ 270k SNPs assuming 2 to 15 clusters.

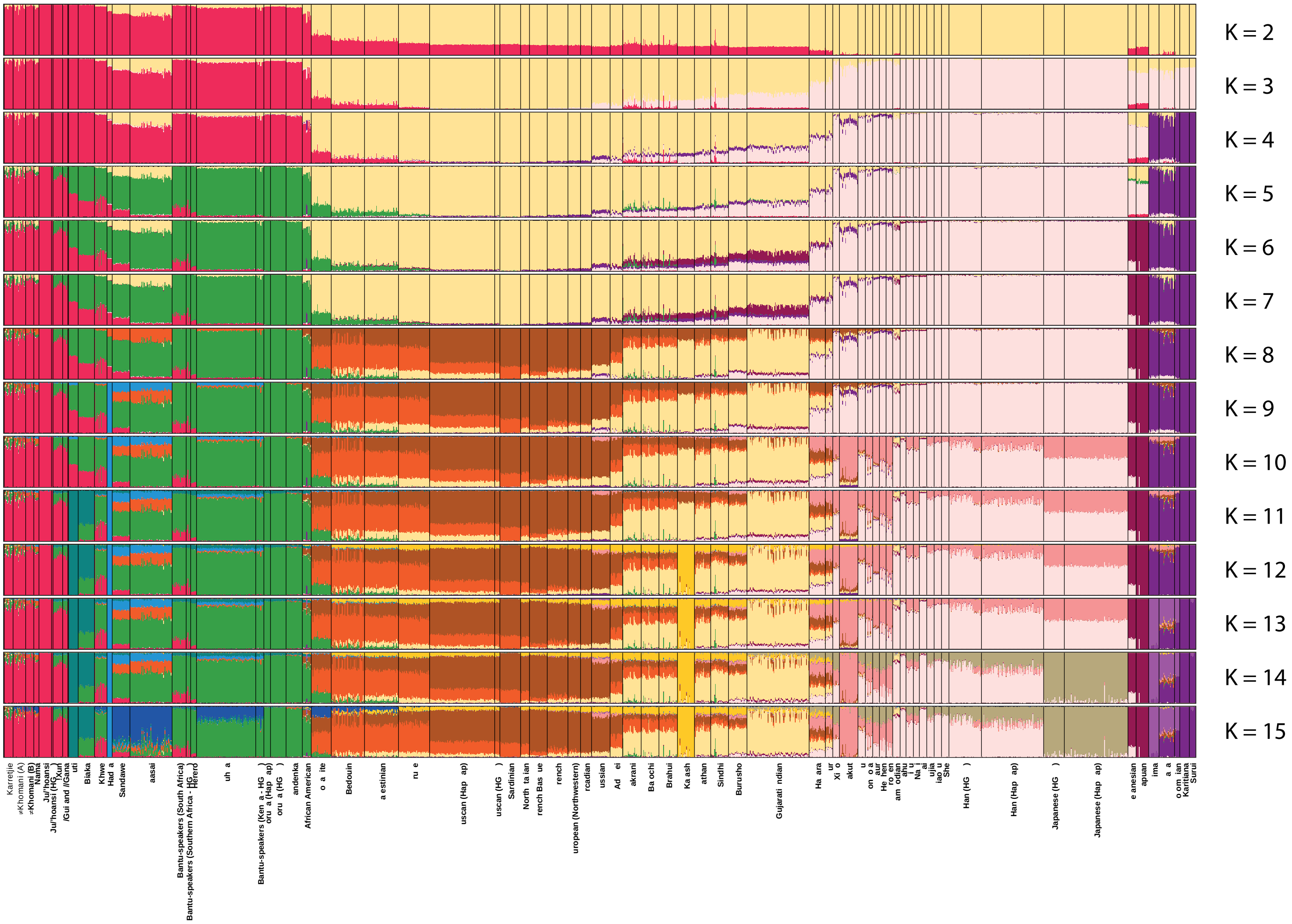

Figure S19: Population structure for a worldwide set of individuals (admixed individuals removed) based on ~ 270k SNPs assuming 2 to 15 clusters.

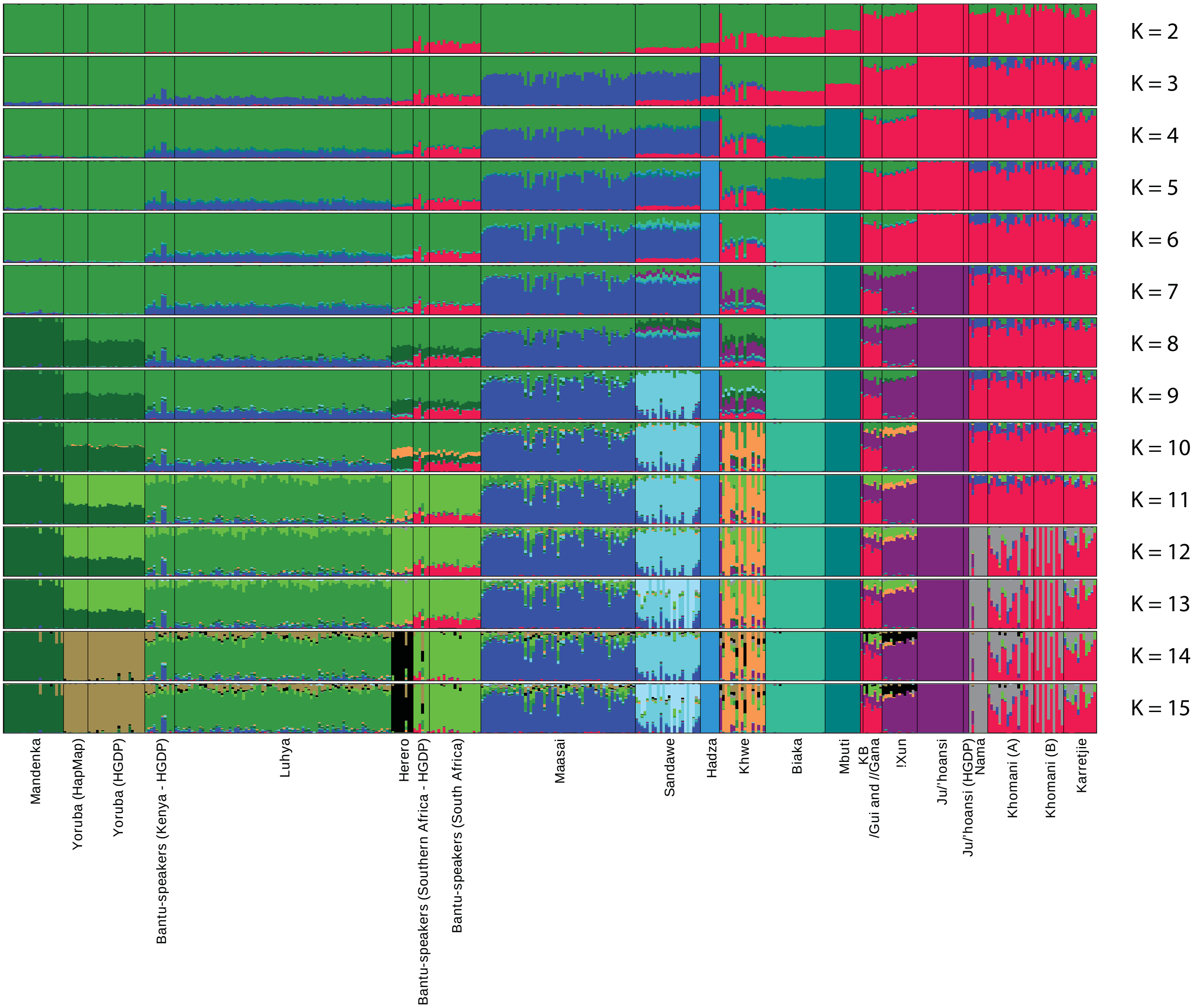

Figure S20: Population structure for sub-Saharan individuals (admixed individuals removed) based on ∼ 270k SNPs assuming 2 to 15 clusters.

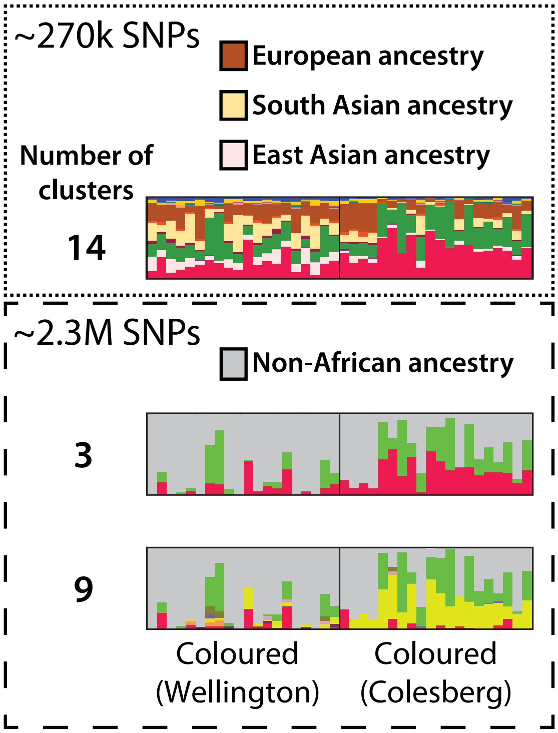

Figure S21: Zoom-in of clustering of admixed "Coloured" individuals. Top pane shows result assuming 14 clusters and including a global set of populations (Figure S18), the bottom panes show the results assuming 3 and 9 clusters. The red, green and yellow clusters correspond to clusters in Figure S17.

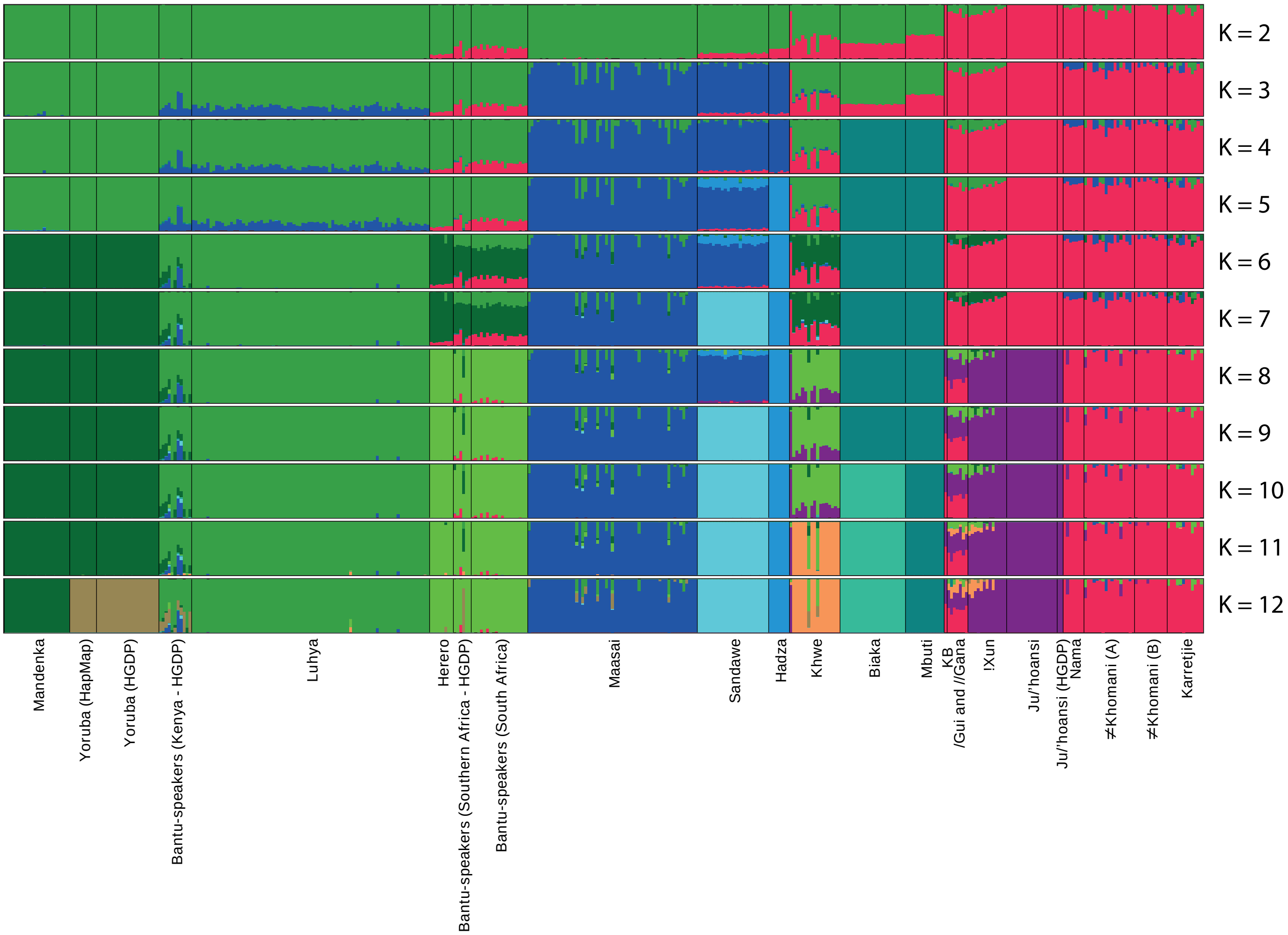

Figure S22: Population structure for sub-Saharan individuals (admixed individuals removed) assuming 2 to 12 clusters. The population structure inference was based on haplotypes, or combinations of SNP variants, instead of SNPs.

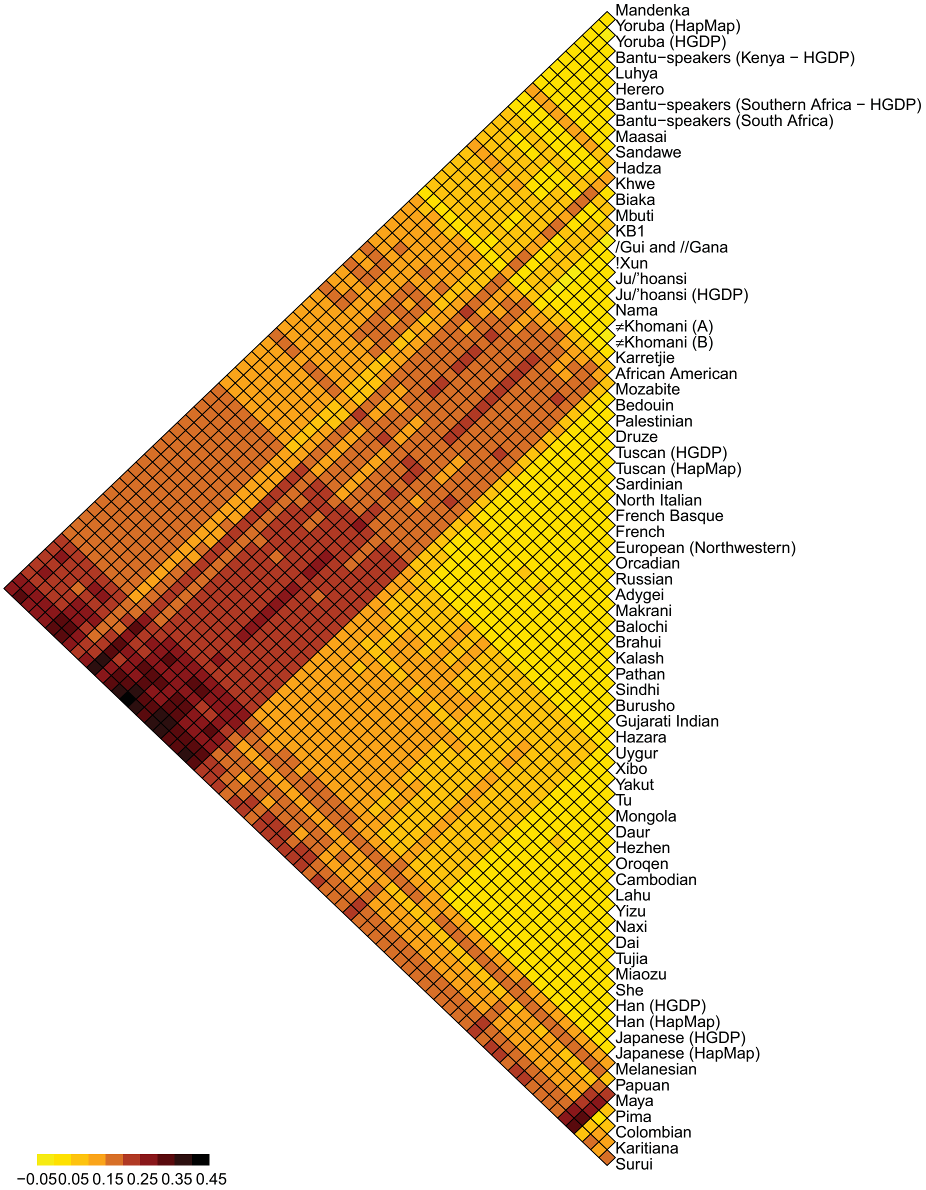

Figure S23: Distograms of the pairwise F_{ST} values computed from $\sim$ 270k SNPs for the global dataset.

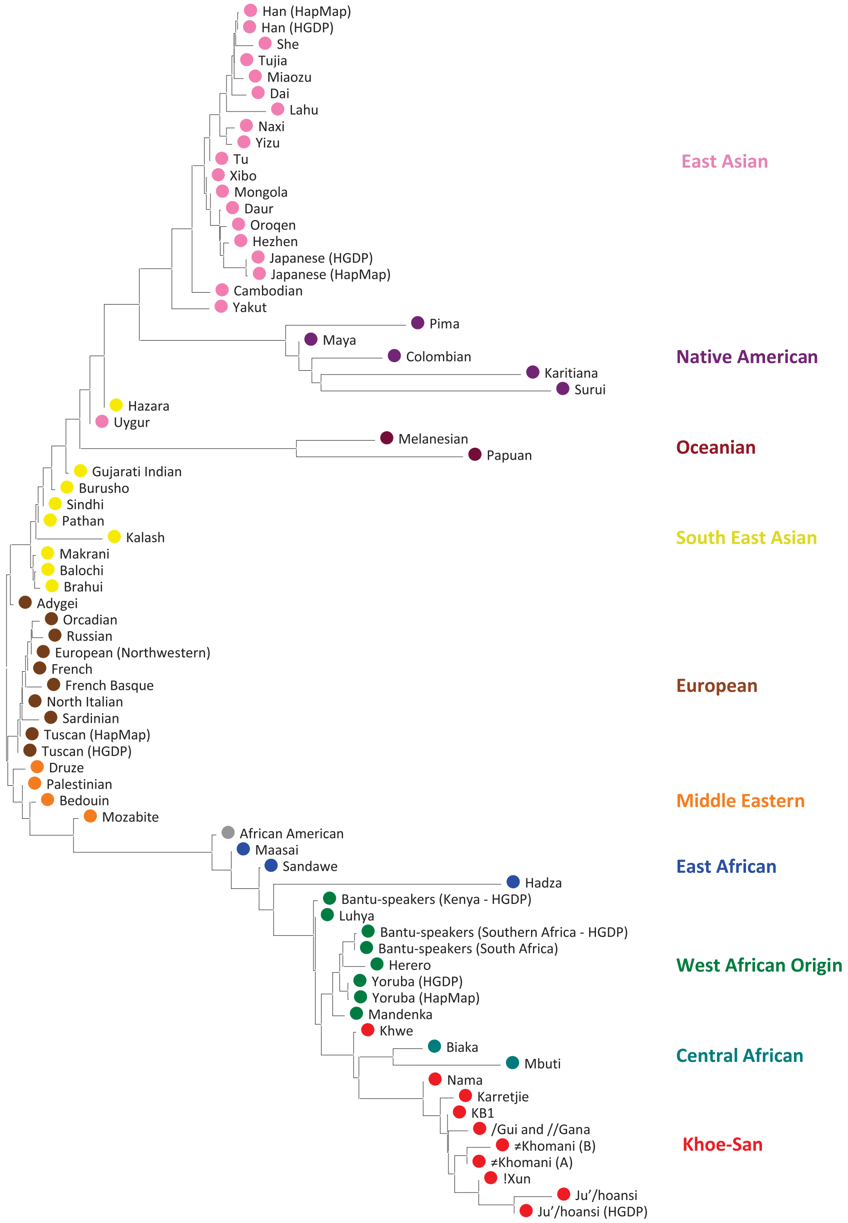

Figure S24: Unrooted neighbor-joining population tree based on pairwise (log transformed: $-\log(1-F_{ST})$) F_{ST} values computed from $\sim$ 270k SNPs. Geographic regions of the populations are indicated in the tree.

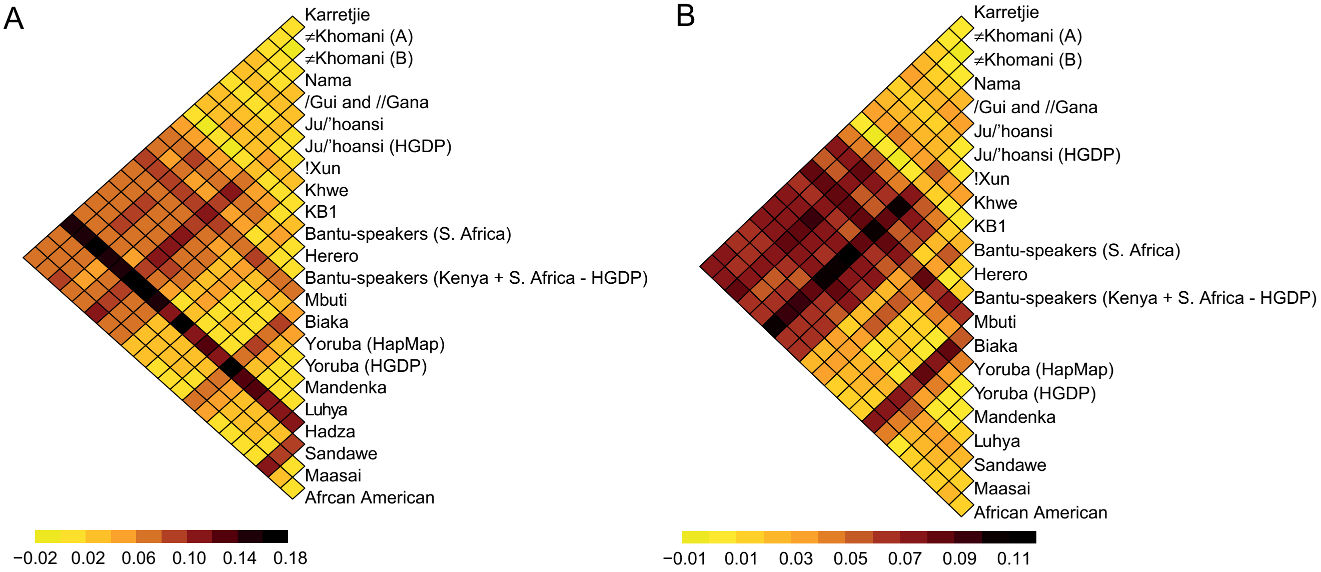

Figure S25: Distograms of the pairwise F_{ST} values for the (A) sub-Saharan dataset and (B) sub-Saharan dataset excluding Hadza.

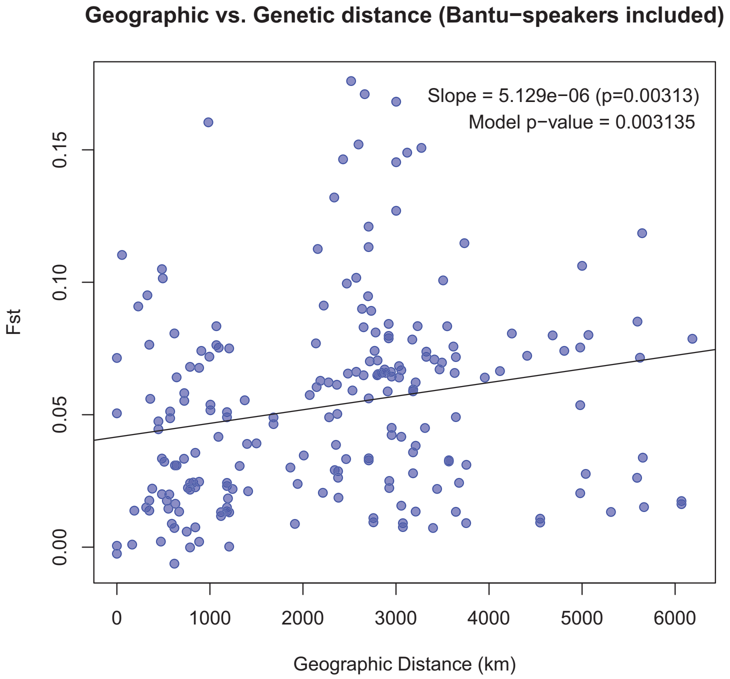

Figure S26: Genetic distance (F_{ST}) as a function of geographic distance (great circle distance) for all sub-Saharan African populations.

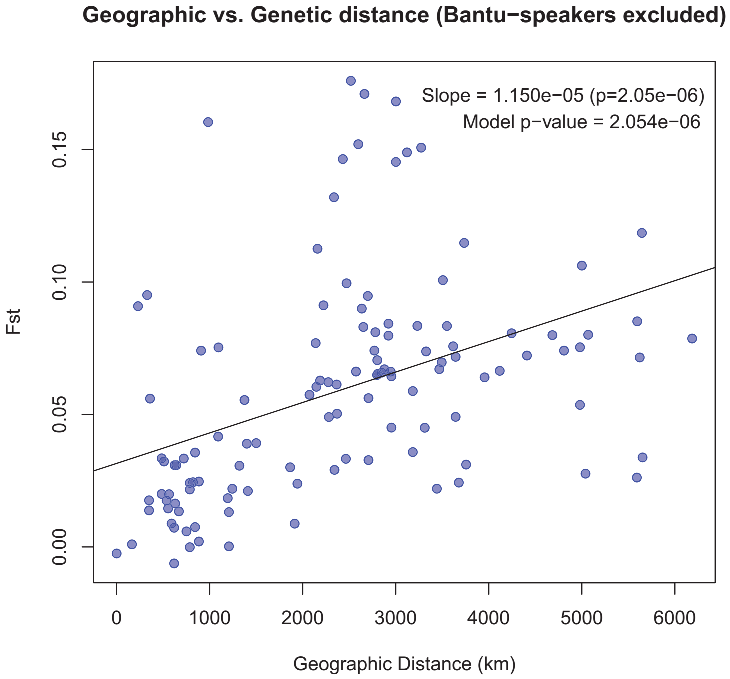

Figure S27: Genetic distance (F_{ST}) as a function of geographic distance (great circle distance) for all sub-Saharan African populations excluding Bantu-speaking populations.

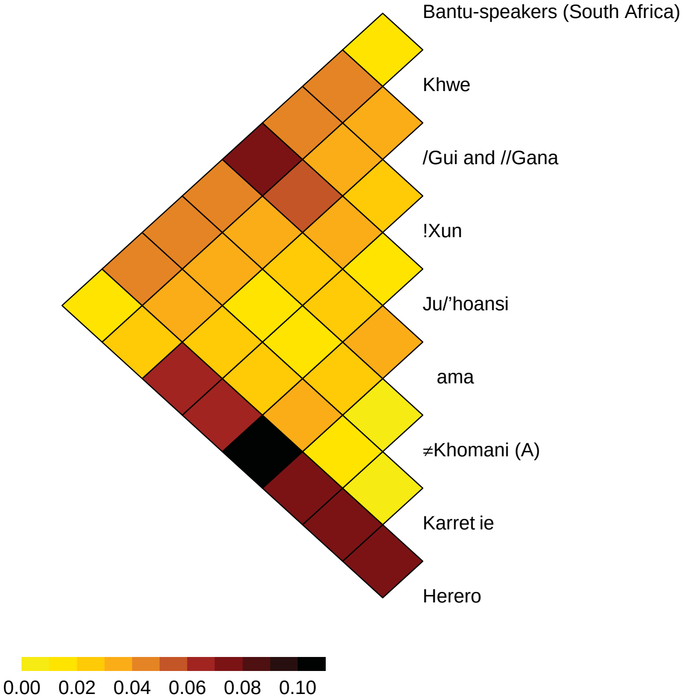

Figure S28: Distograms of the pairwise F_{ST} values for Southern African populations based on ~2.3M SNPs.

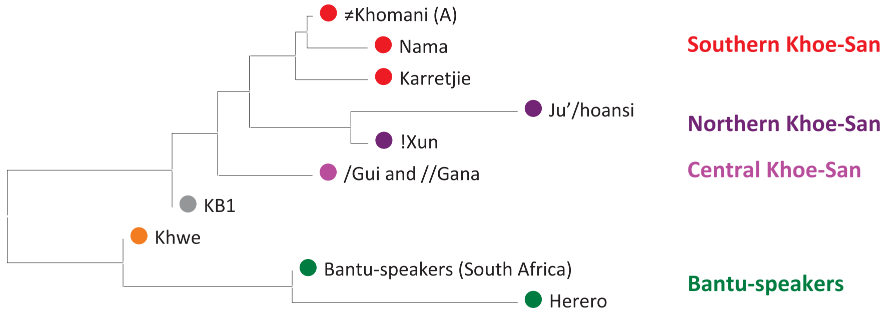

Figure S29: Un-rooted neighbor-joining population tree of Southern African populations based on pairwise (log transformed: $-\log(1 - F_{ST})$) F_{ST} values computed from ~ 2.3M SNPs.

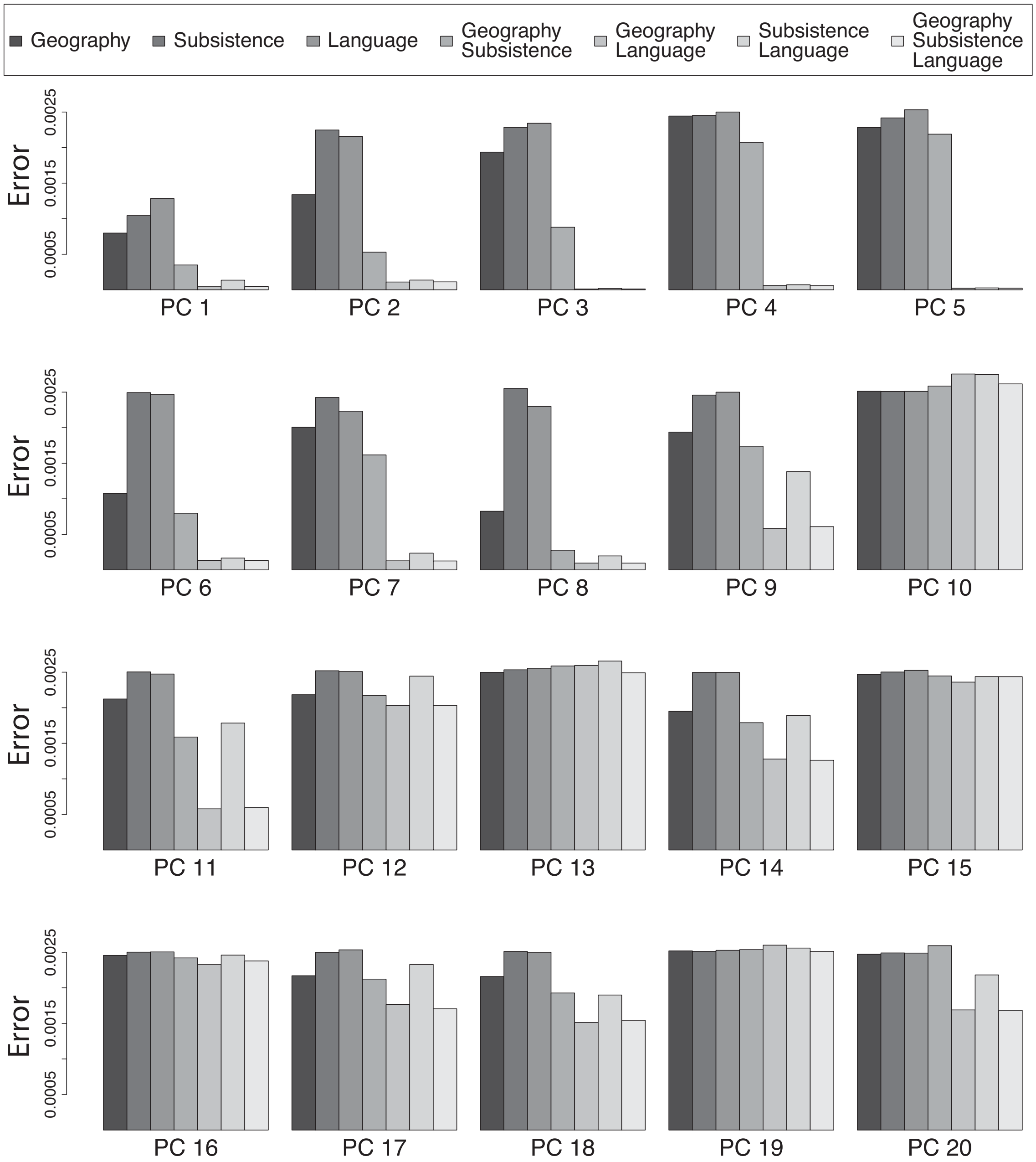

Figure S30: Predictive error for each of the 20 first principal components and each of the 7 models. The error was calculated using a 5-fold cross-validation procedure.

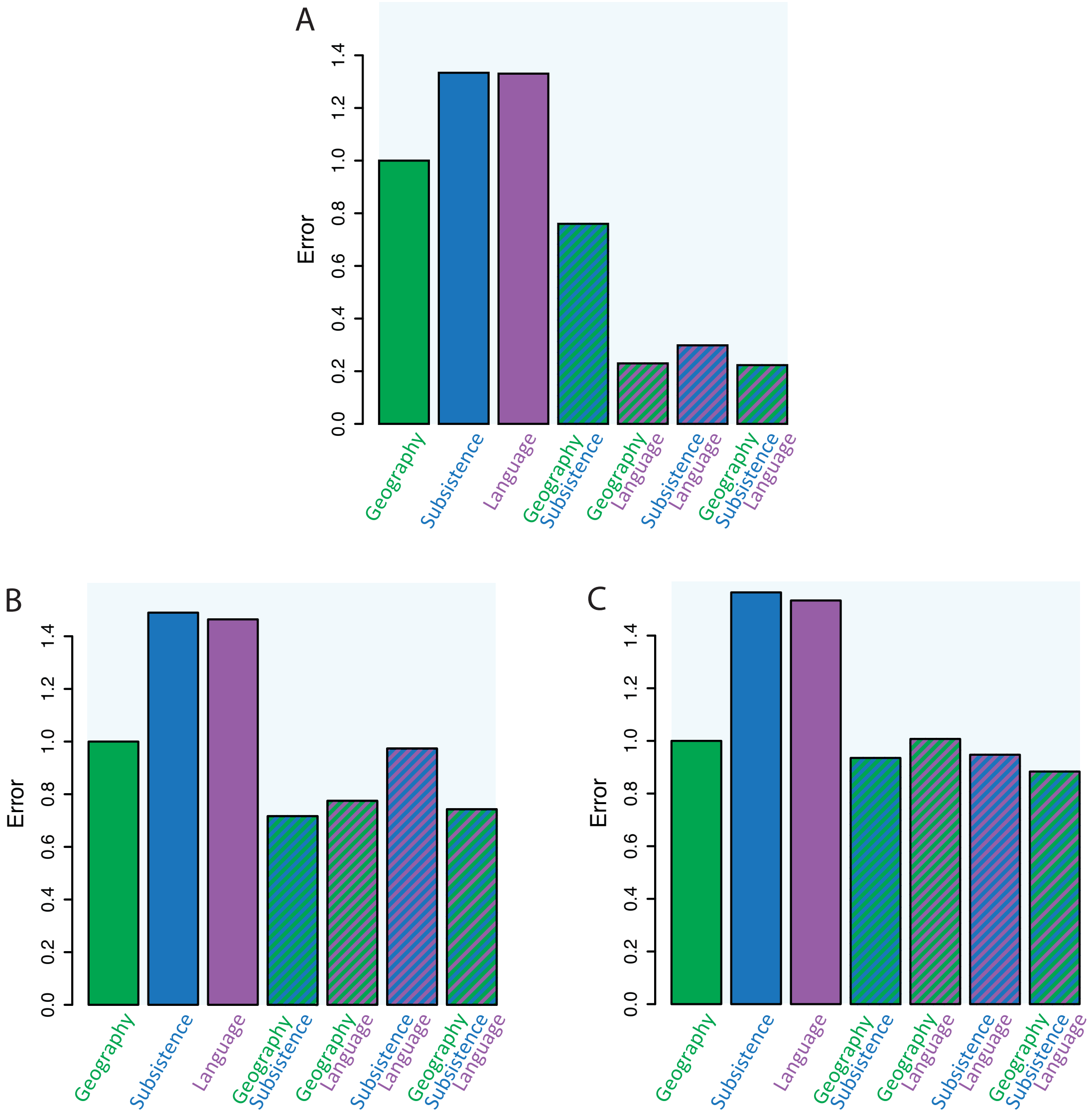

Figure S31: Predictive error averaged over the 10 first principal components for each of the 7 models. A: Sub-Saharan dataset, B: Southern Africa dataset, C: Southern Africa dataset without Bantu-speakers

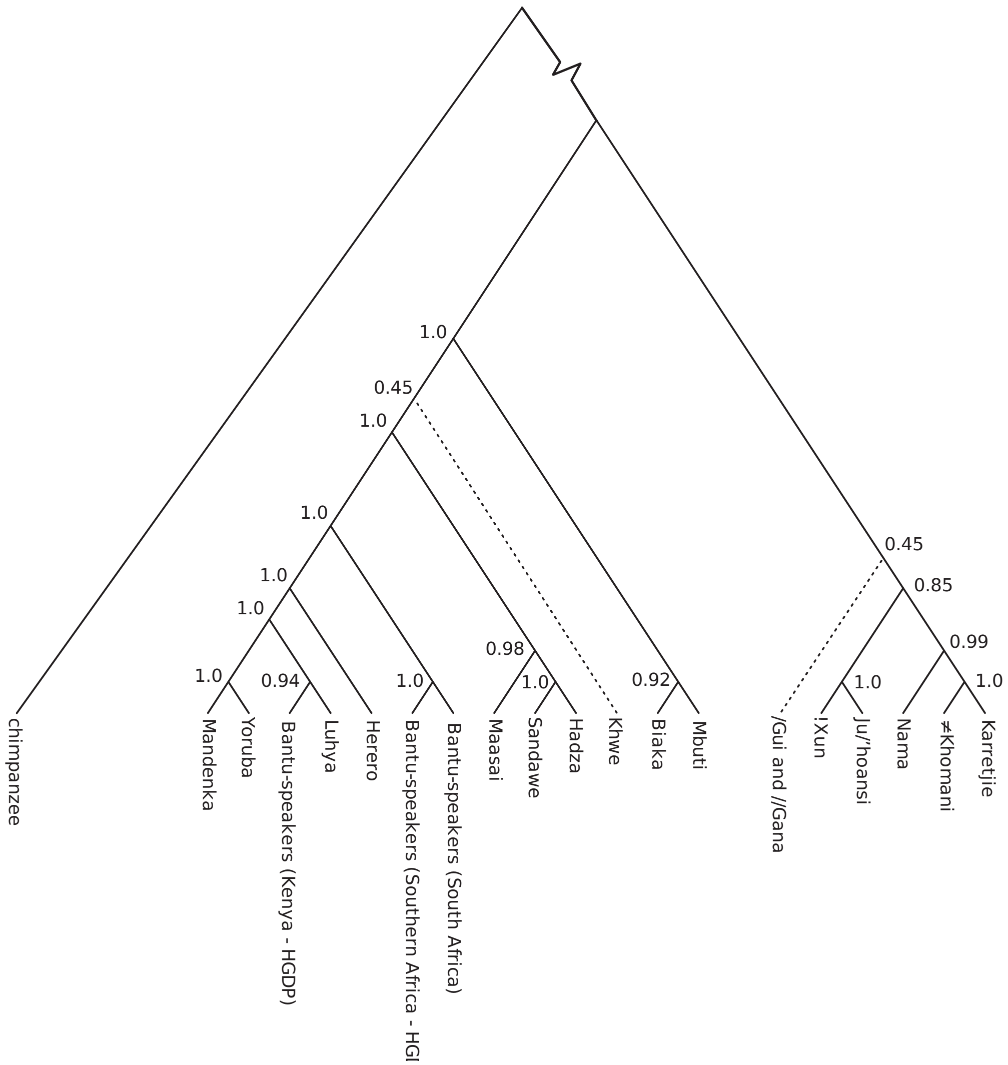

Figure S32: Inferred topology using all sub-Saharan African populations and ∼ 270,000 SNPs. Branch lengths are not related to divergence time or genetic distance.

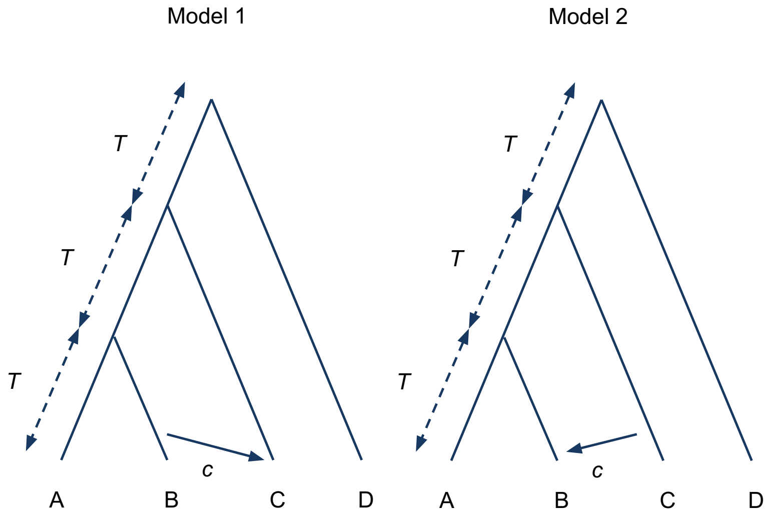

Figure S33: Illustration of models used to investigate the effect of recent admixture on the C-statistic.

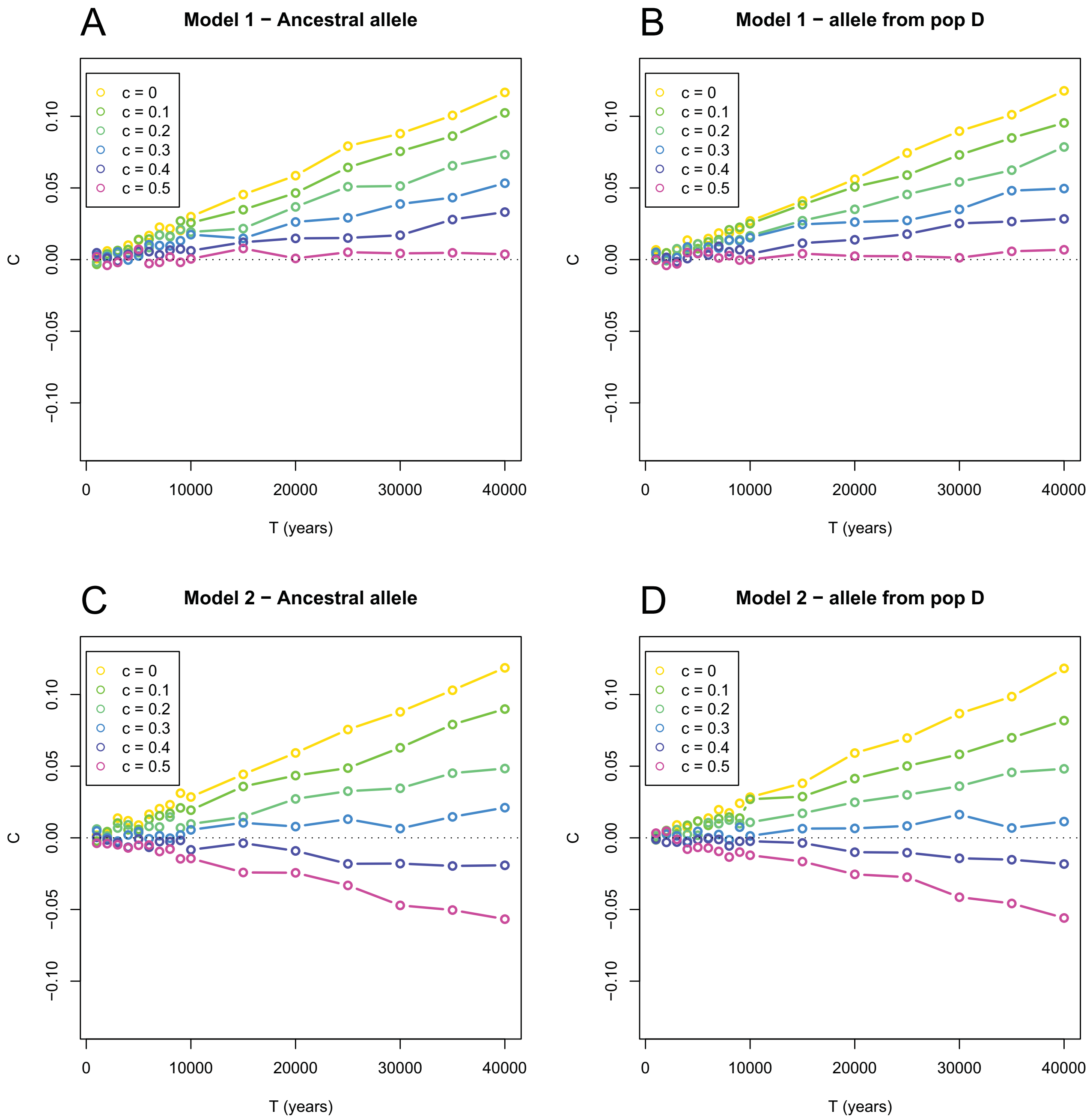

Figure S34: Results of simulations investigating the effect of admixture on the C-statistic. While $C > 0$, the true topology is correctly inferred. For illustration of the models see Figure S33.

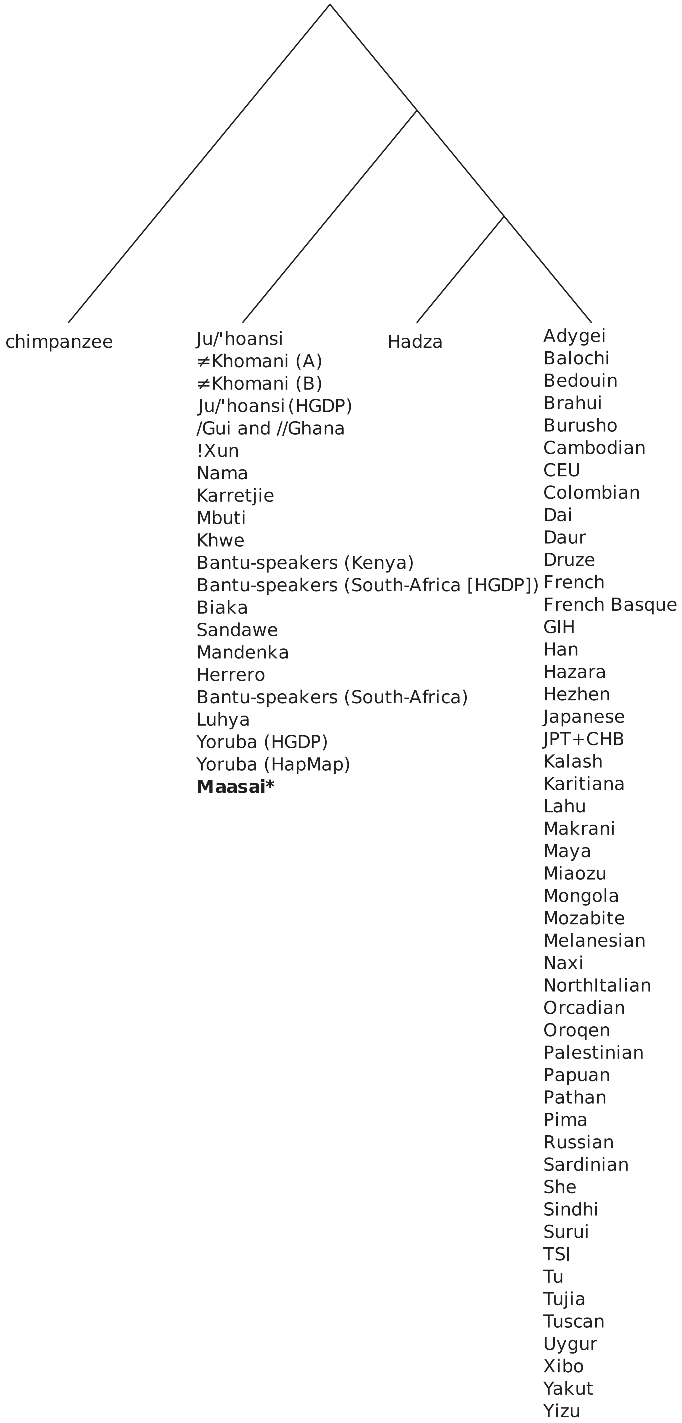

Figure S35: Illustration of the 1,029 four-taxon tests of the hypothesis that East African Hadza are the closest relatives to non-Africans in the data analyzed here. All 980 concordance tests not involving the Maasai as the other African group were significantly supported ($Z > 3$). Note that "Africans" in this Figure are not a monophyletic grouping (see above), but tested individually the grouping *(chimpanzee,(African,(Hadza,non-African))* is robust.

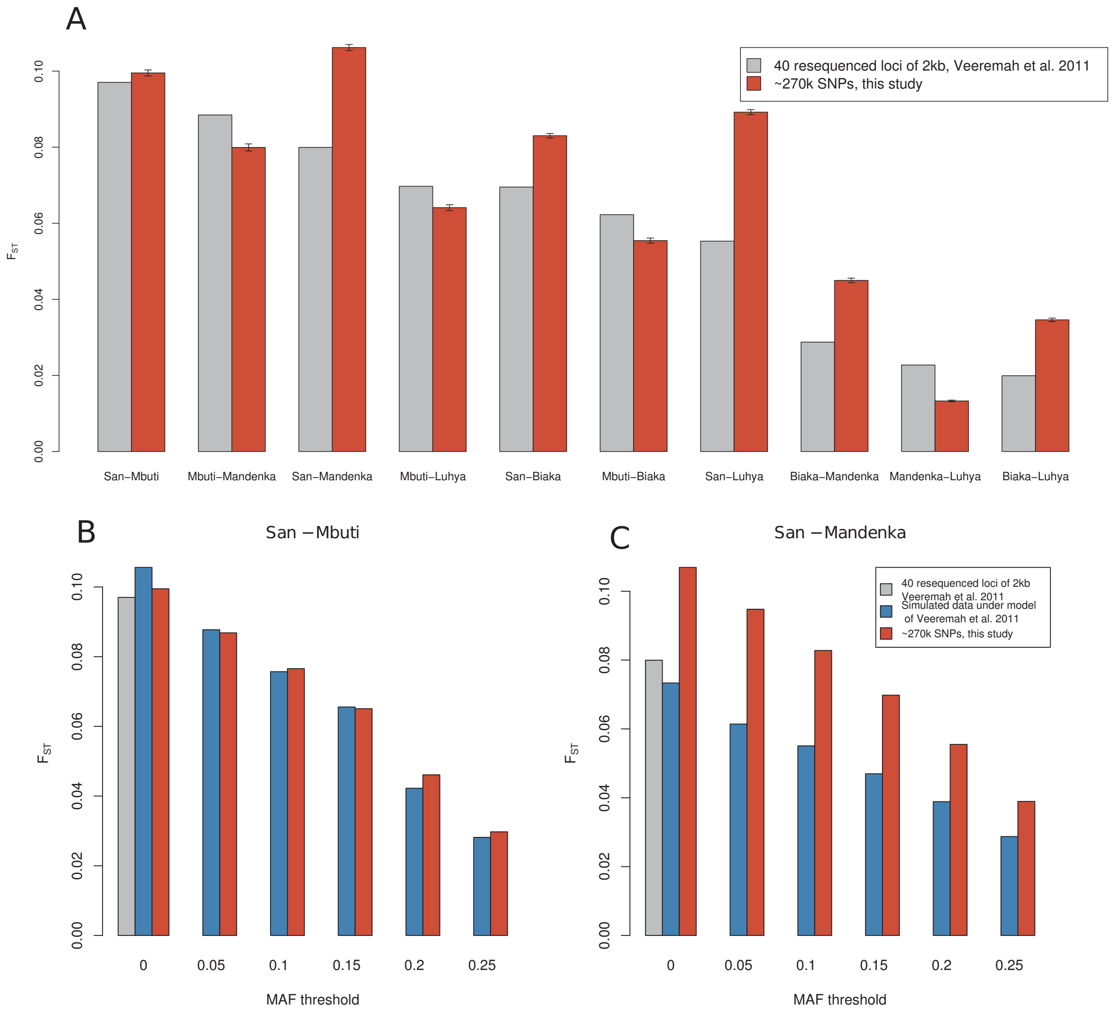

Figure S36: Comparisons between F_{ST} estimated using our SNP data and the results of ref. (*12*). In A), pairwise F_{ST} between all populations present in both this study and ref. (*12*) are presented, with SEs for this study given (estimated using a block jackknife approach). The largest discrepancies are between San and pastoralist populations, but F_{ST}s between San and Mbuti are almost identical. In B), F_{ST} computed based on simulations of a demographic model inferred for San and Mbuti are given stratified by a minor allele frequency threshold in each population (MAF), and compared to our estimates on empirical SNP data using the same threshold. In C), the same comparison as in B) is made between San and Mandenka.

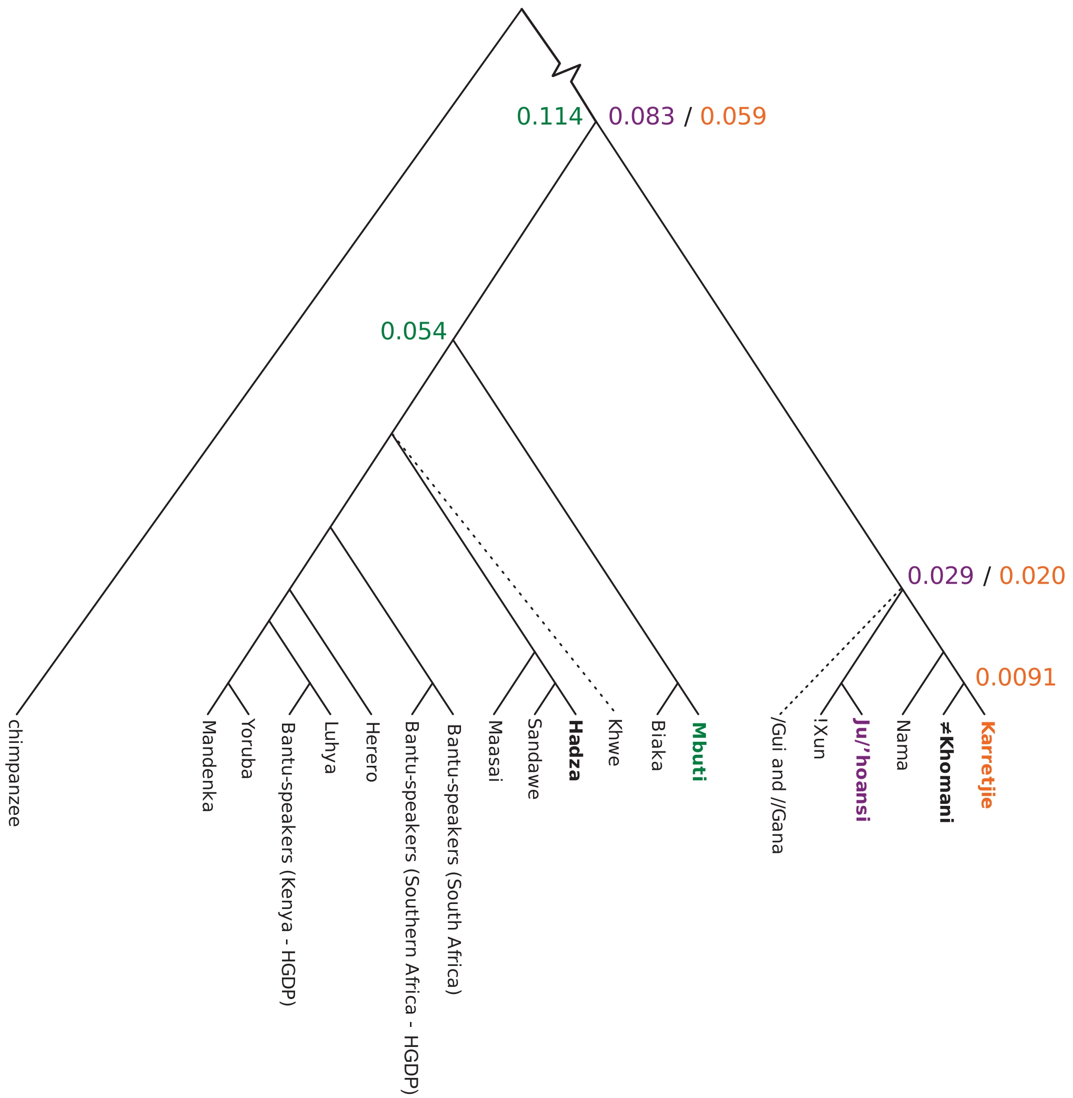

Figure S37: Inferred topology using all sub-Saharan African populations and ∼ 270,000 SNPs. Divergence time estimates for chosen nodes are given. See Table S9 and the text for details. Branch lengths are not related to divergence time or genetic distance.

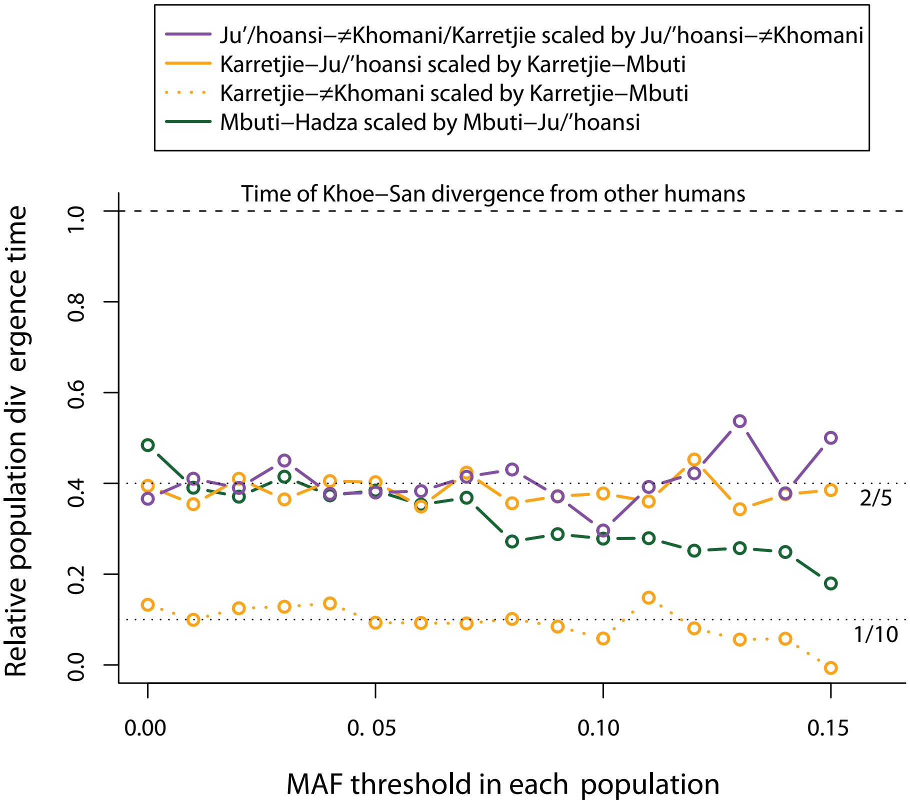

Figure S38: Divergence times between sub-Saharan African hunter-gatherer populations scaled by the time back to the ancestral population of all modern humans.

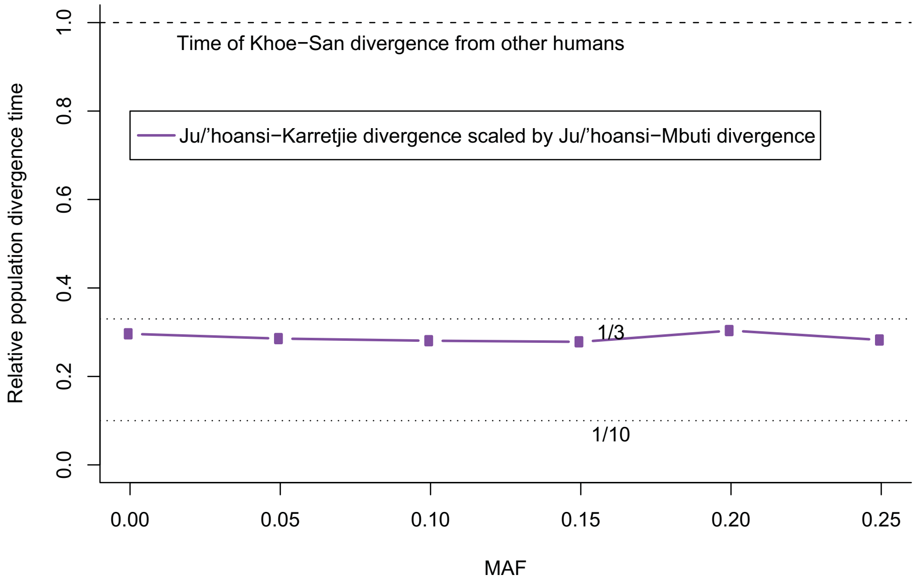

Figure S39: F_{ST}-based divergence time between Ju/'hoansi and Karretjie scaled by the divergence between Ju/'hoansi and Mbuti.

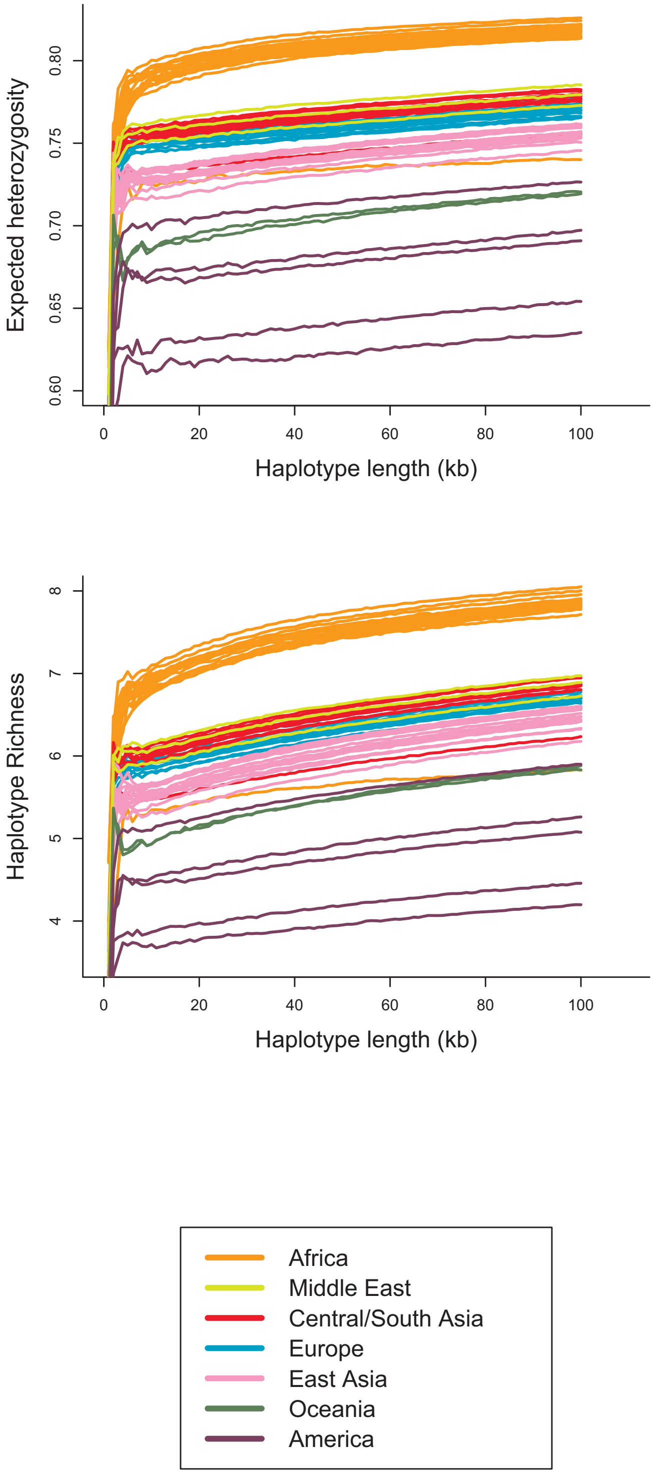

Figure S40: Haplotype heterozygosity (A) and richness (B) as a function of haplotype size in a global set of populations.

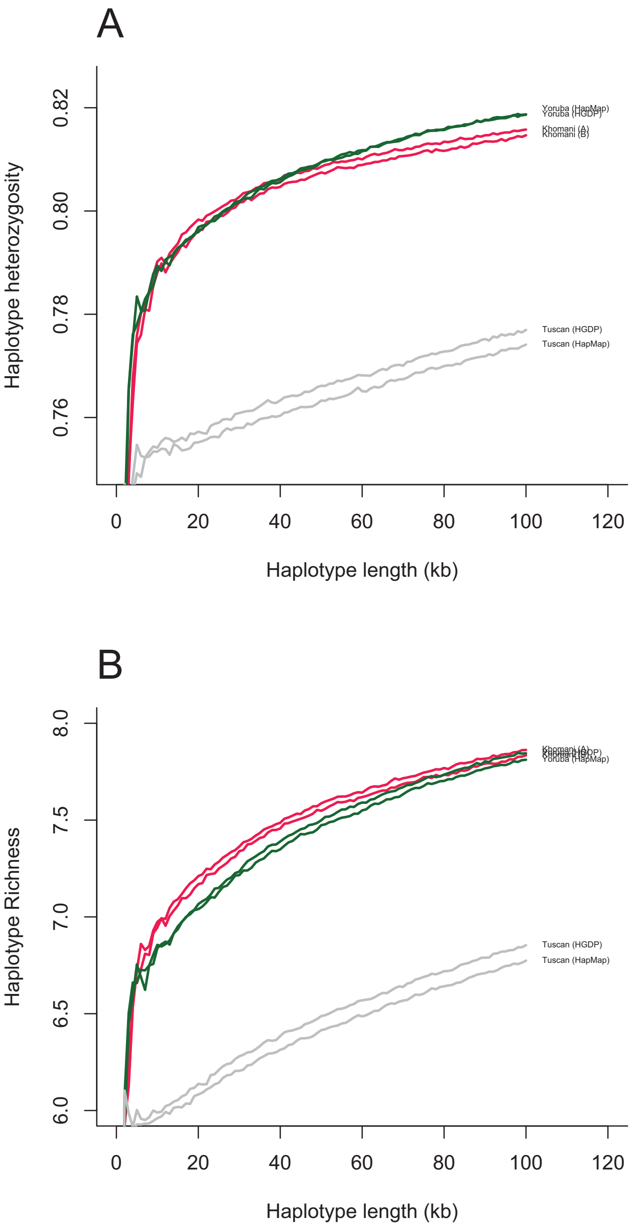

Figure S41: Haplotype heterozygosity (A) and richness (B) for three populations from which we have two independent samples: West African Yoruba, Southern African ≠Khomani, and European Tuscan.

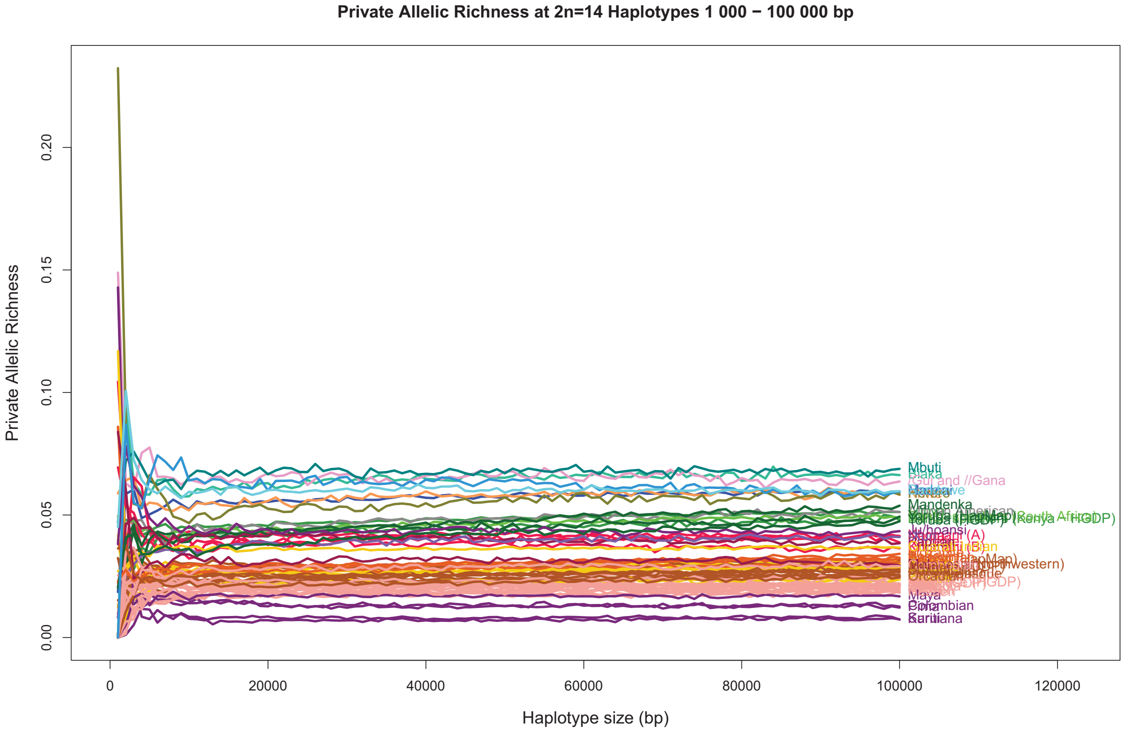

Figure S42: Number of private haplotypes as a function of haplotype size in global set of populations.

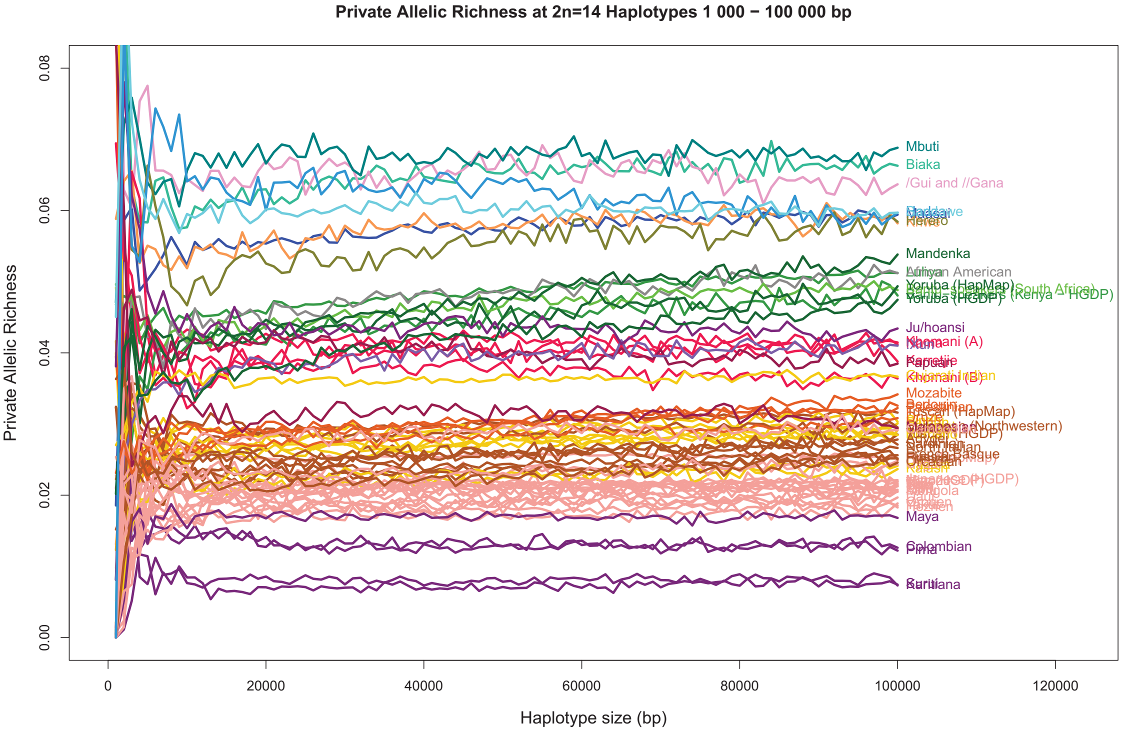

Figure S43: Number of private haplotypes as a function of haplotype size in global set of populations.

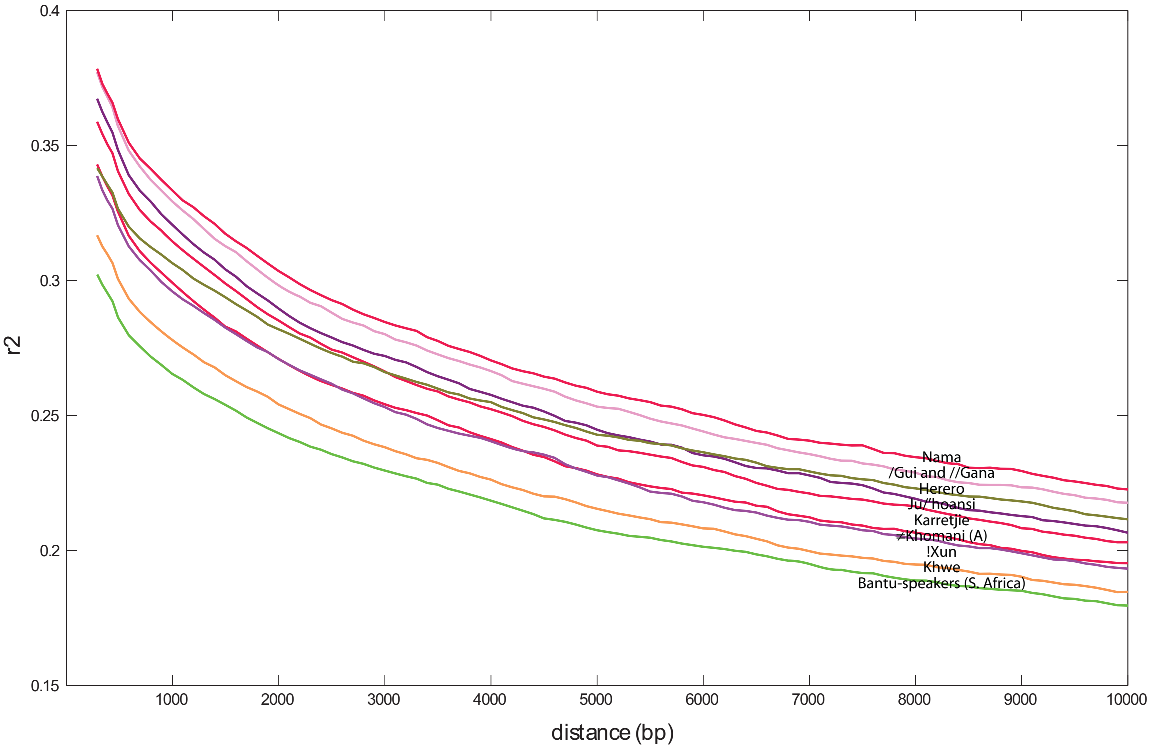

Figure S44: Linkage disequilibrium vs. physical distance (<10kb). r^2 was calculated for each pair of SNPs with minor allele frequency greater than or equal to 0.10. The mean r^2 within a bin is plotted as a function of the mean of the distance between pairs of SNPs within the bin. The bin size was 100bp.

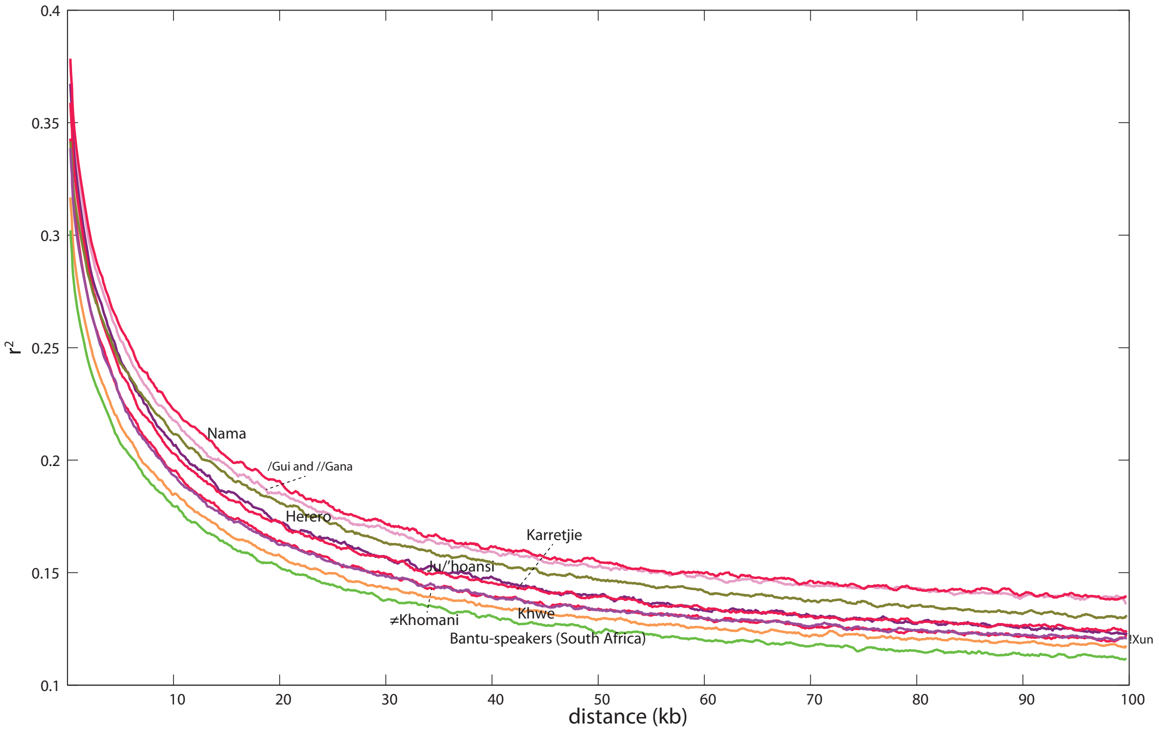

Figure S45: Linkage disequilibrium vs. physical distance (<100kb). r^2 was calculated for each pair of SNPs with minor allele frequency greater than or equal to 0.10. The mean r^2 within a bin is plotted as a function of the mean of the distance between pairs of SNPs within the bin. The bin size was 100bp.

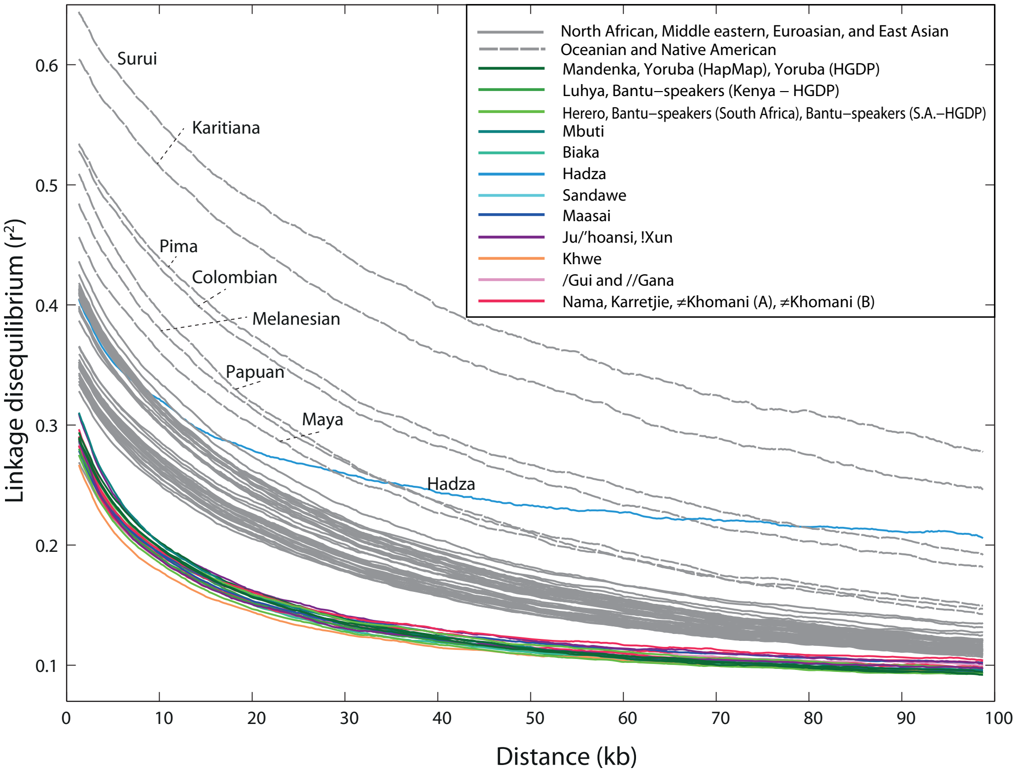

Figure S46: Linkage disequilibrium vs. physical distance (<100kb). r^2 was calculated for each pair of SNPs with minor allele frequency greater than or equal to 0.10. The mean r^2 within a bin is plotted as a function of the mean of the distance between pairs of SNPs within the bin. The bin size was 5kb.

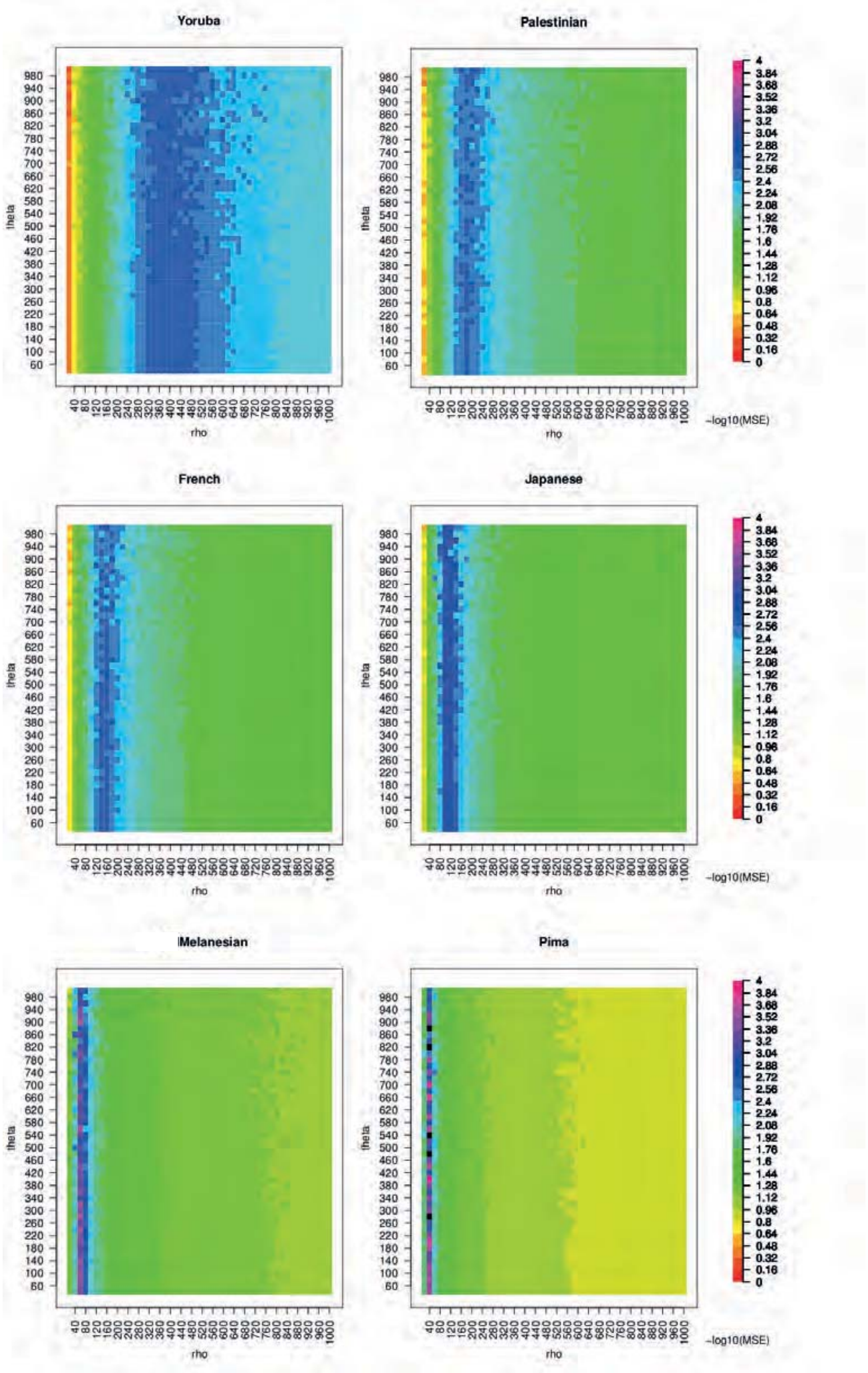

Figure S47: Example of 6 heatmaps representing the $-\log_{10}$ of the Mean Square Error between the observed r^2 curves obtained in section 8.2 and the simulated r^2 curves as described above.

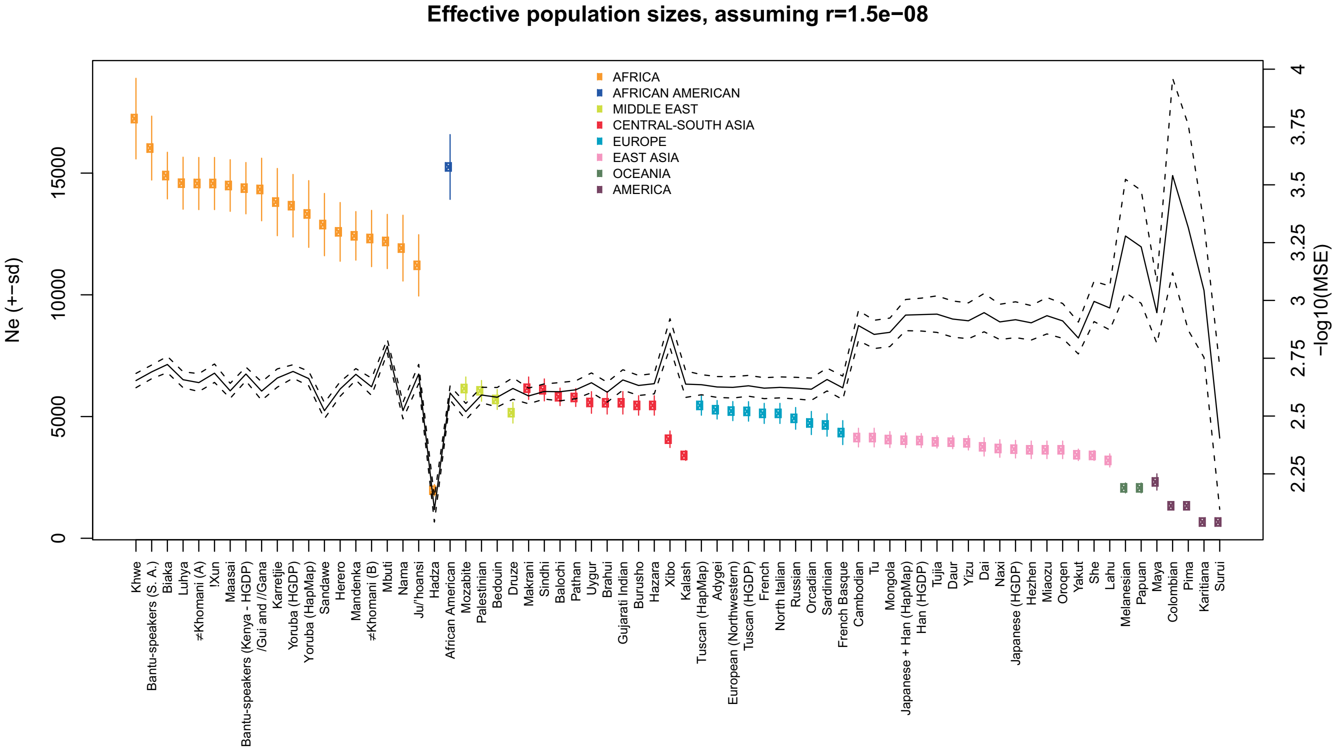

Figure S48: Estimated effective population sizes obtained by fitting simulated data to observed r^2 curves, assuming a recombination rate per site and per generation of 1.5×10^{-8} (filled dots), and $-\log_{10}(\text{MSE})$ of the best fits (black curve). The position of the filled dots represent the average value of the best N_e values obtained for each θ and the bars represent the standard deviation of those values around the mean N_e

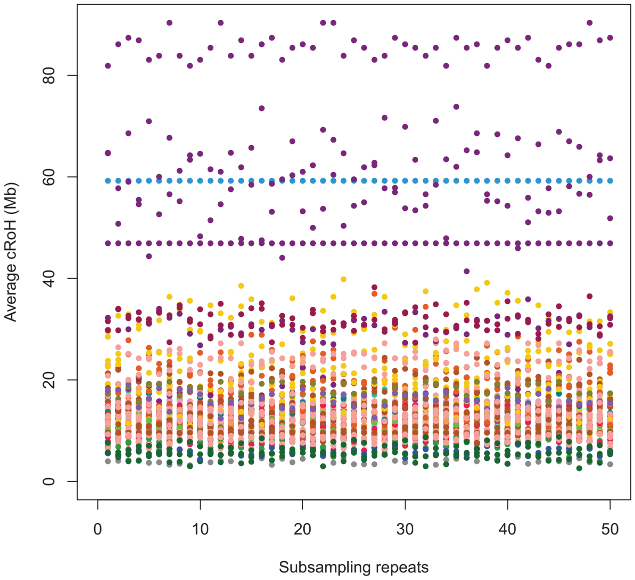

Figure S49: Fifty repeats of *RoH* with random sub-samples (without replacement) of 7 individuals from each population (0.5-1Mb class). Straight lines are due to populations that contain exactly 7 individuals.

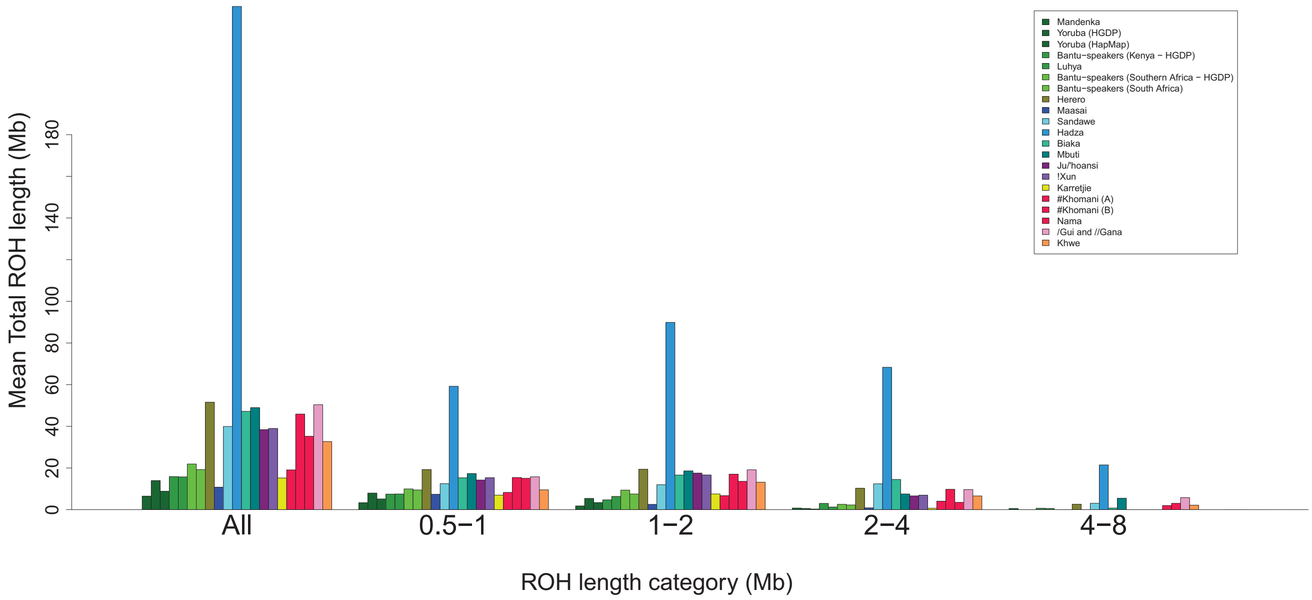

Figure S50: Distribution of long runs of homozygosity for the African dataset across all length classes

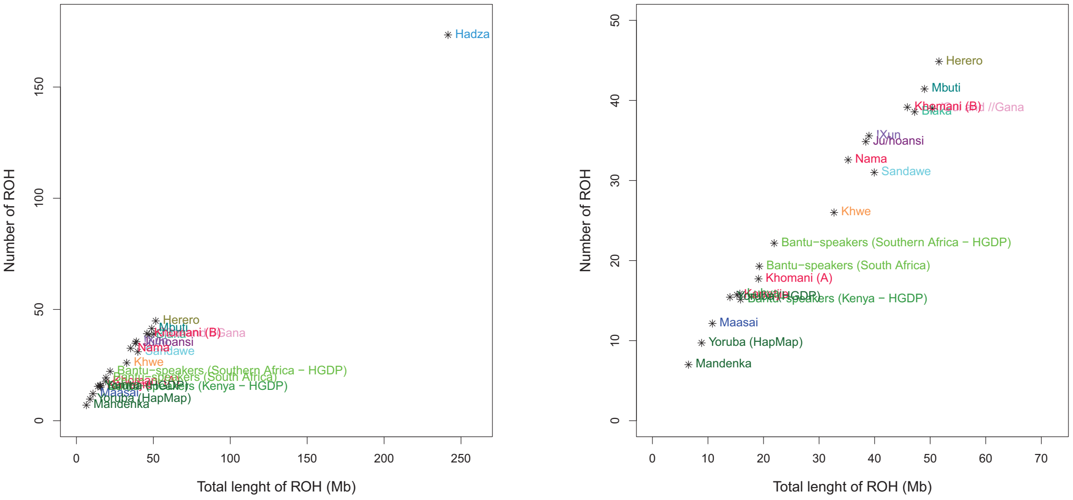

Figure S51: The number of runs of homozygosity compared to the total length in *RoH* for populations

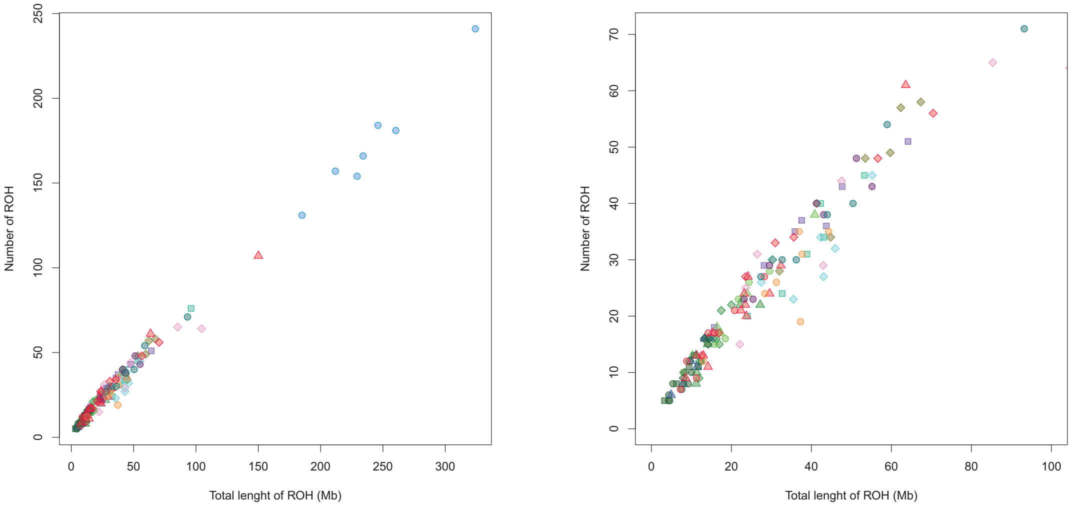

Figure S52: The number of runs of homozygosity compared to the total length in *RoH* for individuals

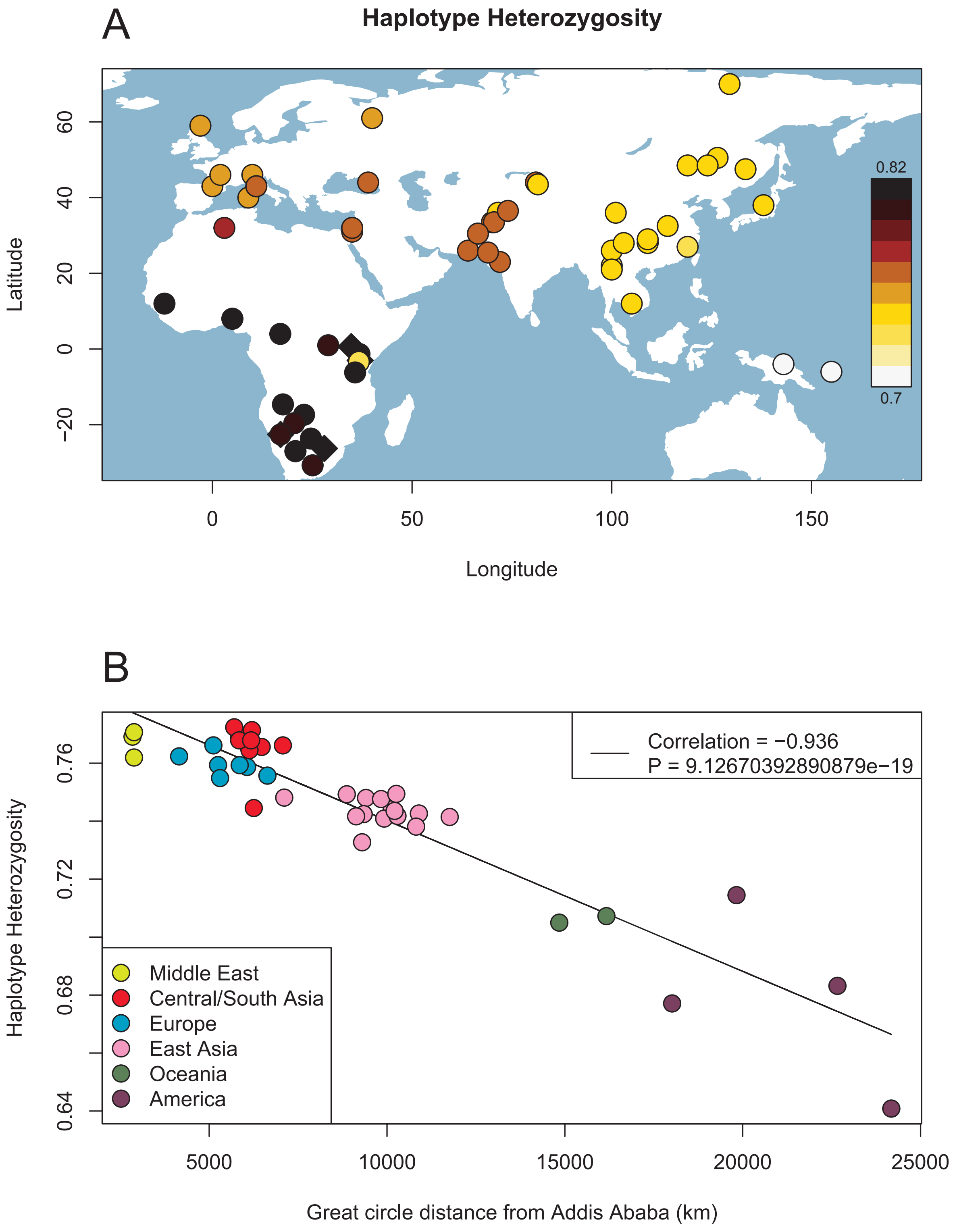

Figure S53: Haplotype heterozygosity at 50kb in the Old World. Geographical distribution (A) and as a function distance from East Africa (Addis Ababa) in non-Africans (B). Populations associated with the Bantu-expansion are shown as diamonds in (A).

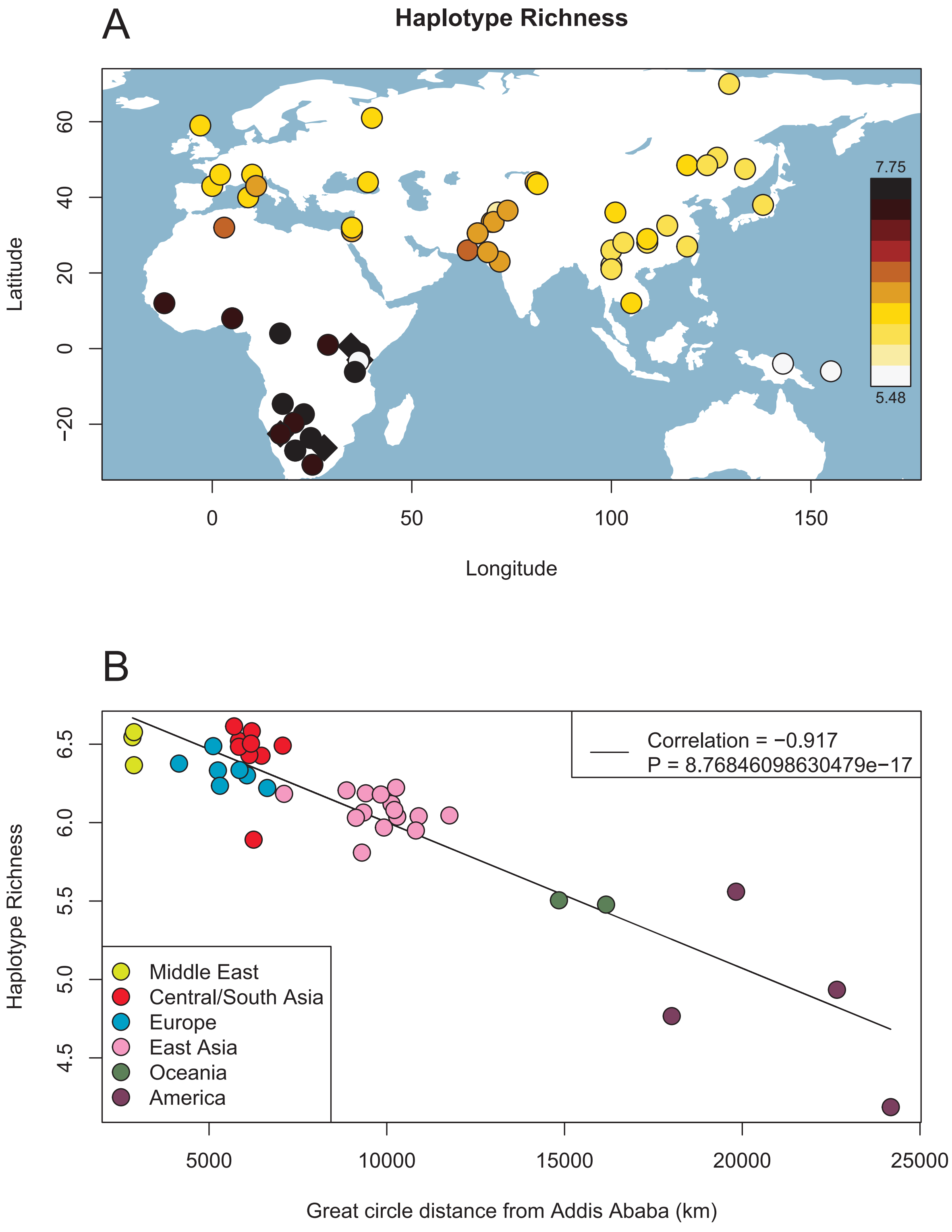

Figure S54: Haplotype richness at 50kb in the Old World. Geographical distribution (A) and as a function distance from East Africa (Addis Ababa) in non-Africans (B). Populations associated with the Bantu-expansion are shown as diamonds in (A).

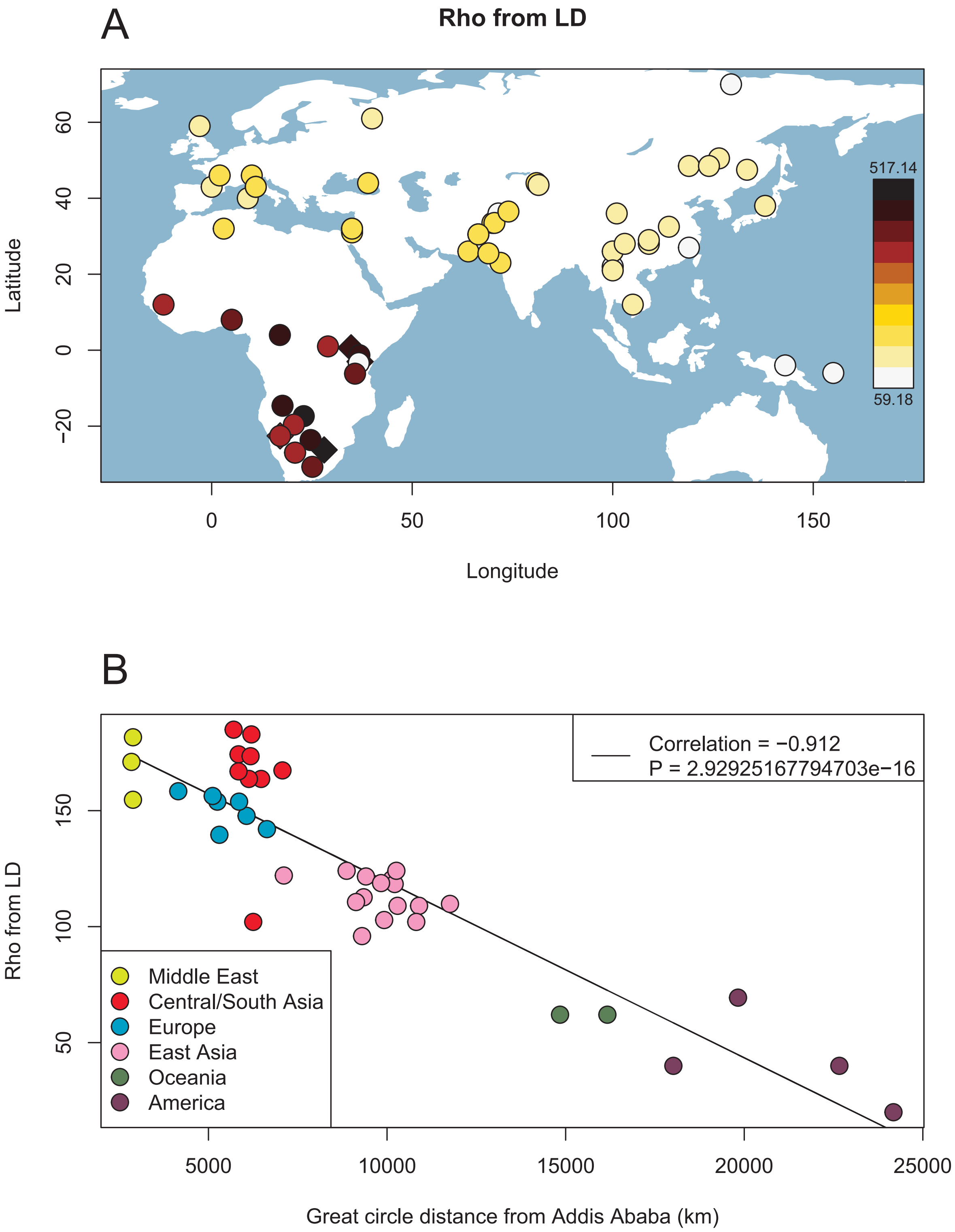

Figure S55: $\rho = 4N_e r$ estimated from LD in the Old World. Geographical distribution (A) and as a function distance from East Africa (Addis Ababa) in non-Africans (B). Populations associated with the Bantu-expansion are shown as diamonds in (A).

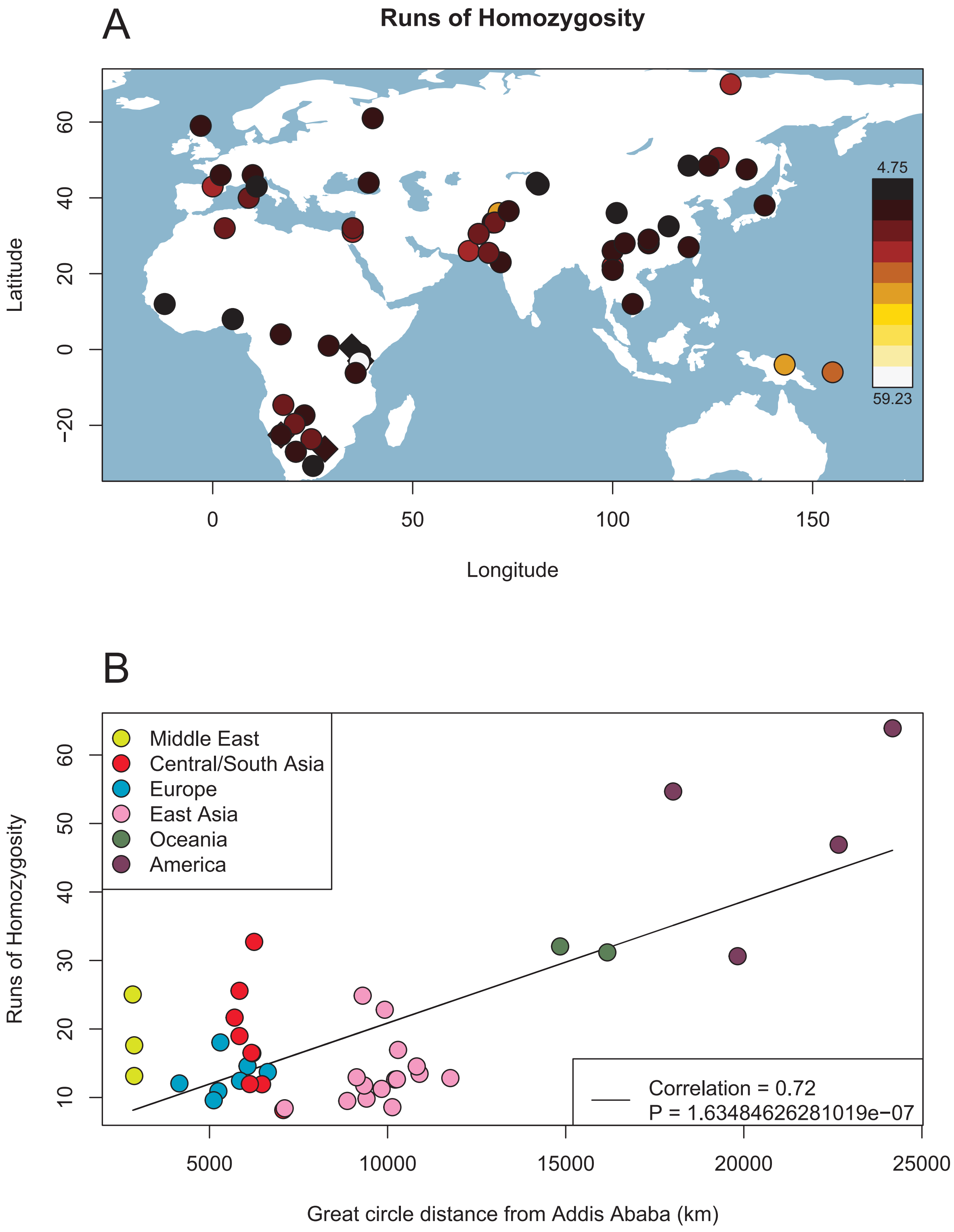

Figure S56: Runs of homozygosity in the Old World. Geographical distribution (A) and as a function distance from East Africa (Addis Ababa) in non-Africans (B). Populations associated with the Bantu-expansion are shown as diamonds in (A).

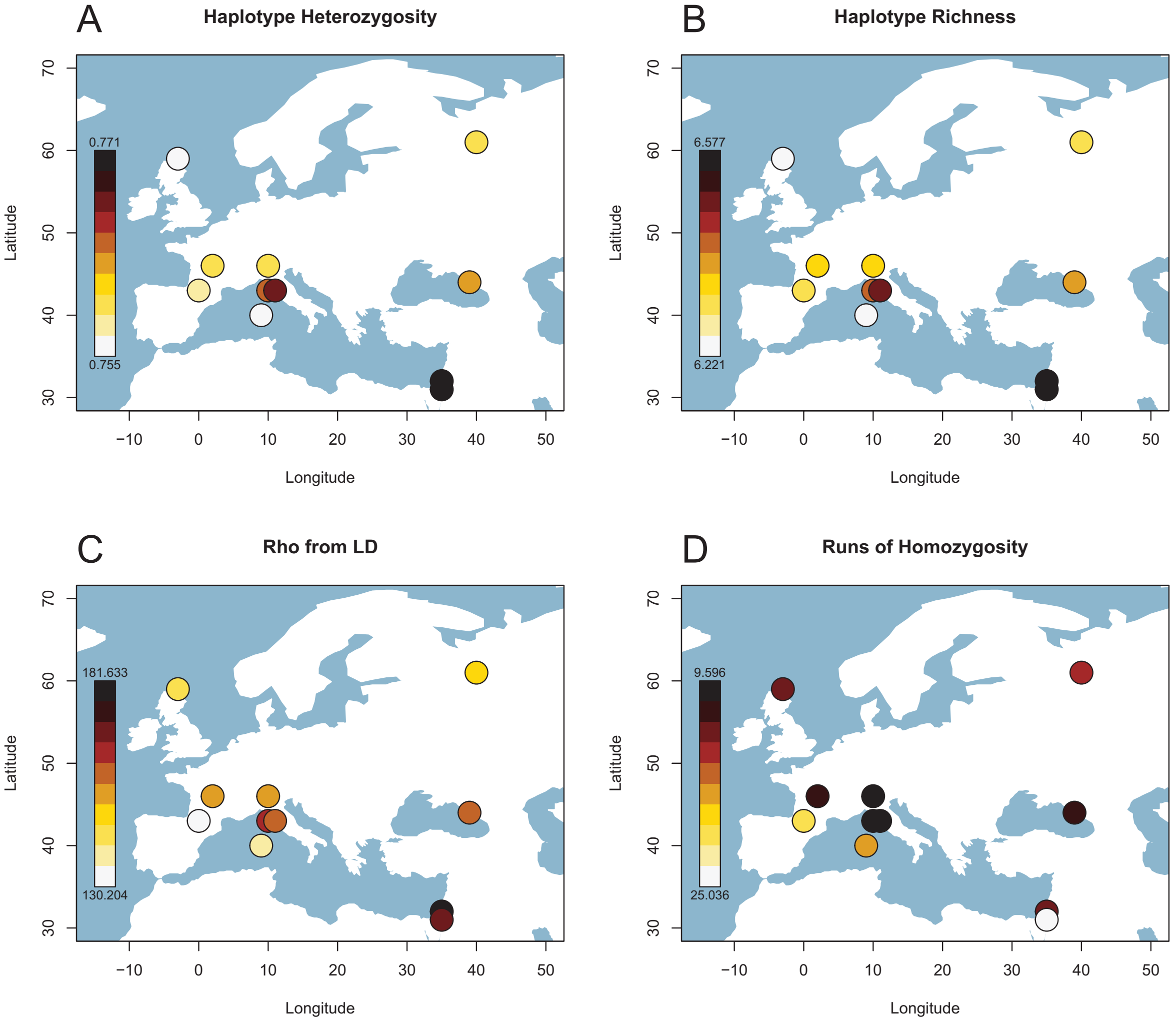

Figure S57: Haplotype (50kb) heterozygosity (A), haplotype richness (50kb) (B), Rho from LD (C), and Runs of Homozygosity averaged over 50 replicates (D) recapitulate previously observed patterns in West Eurasian populations (*126, 129*).

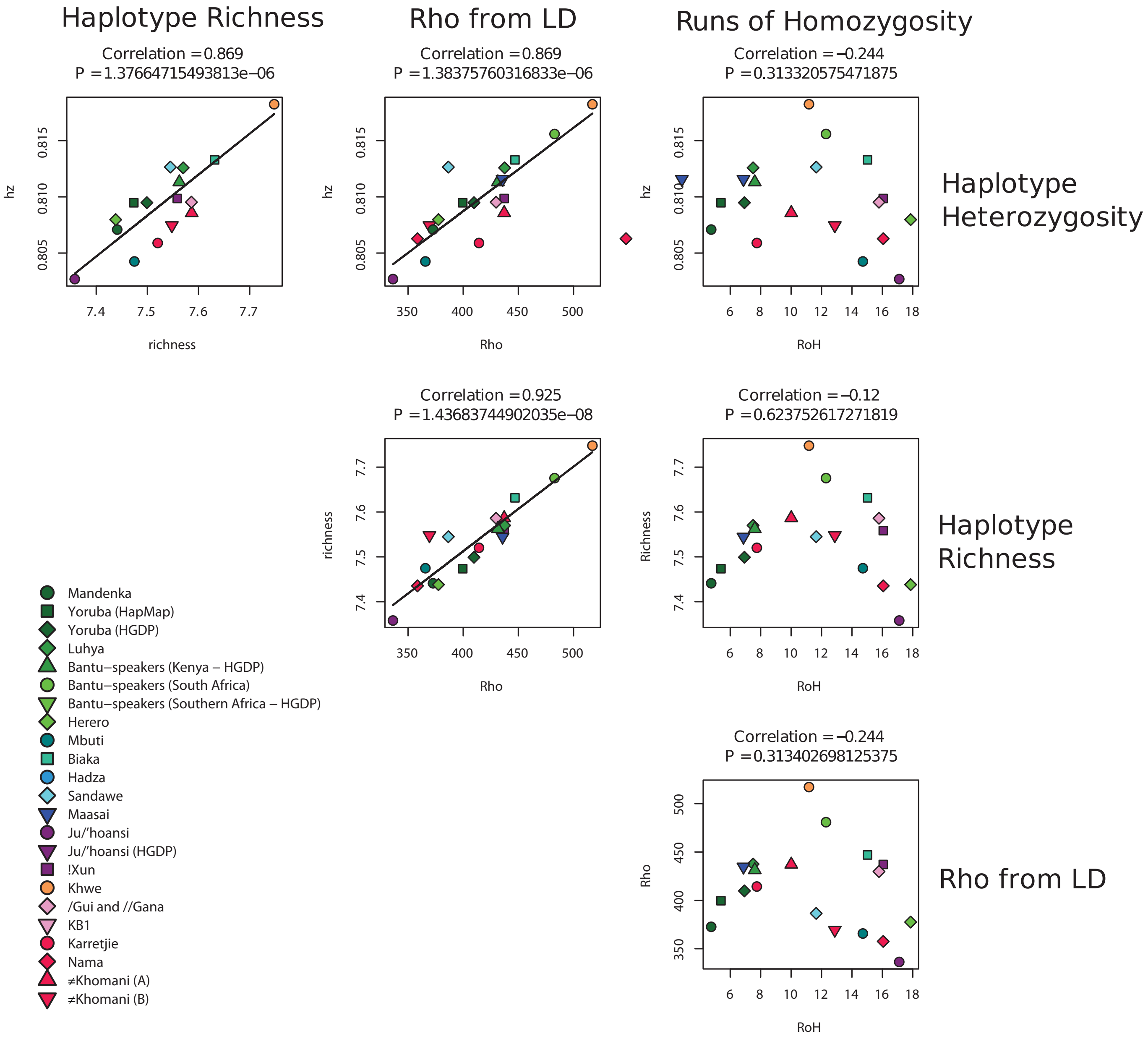

Figure S58: Pairwise correlations between summary statistics in sub-Saharan African populations.

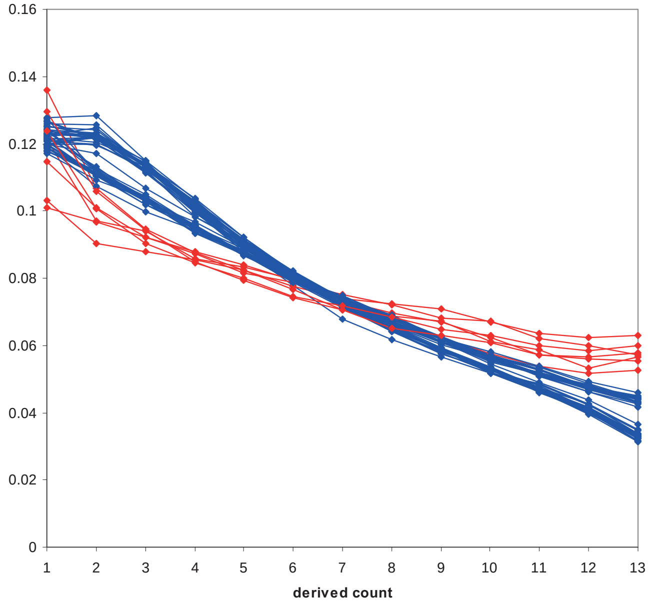

Figure S59: Frequency spectra for non-Africans. Populations (and related populations) used in ascertainment panels (European and Asian) are shown in blue. Populations that are geographically and/or genetically distinct from the ascertained populations (red – Colombian, Karitiana, Pima, Surui, Melanesian, and Papuan) display frequency spectra that are less flat, presumably due to the weaker effect of ascertainment bias in these populations.

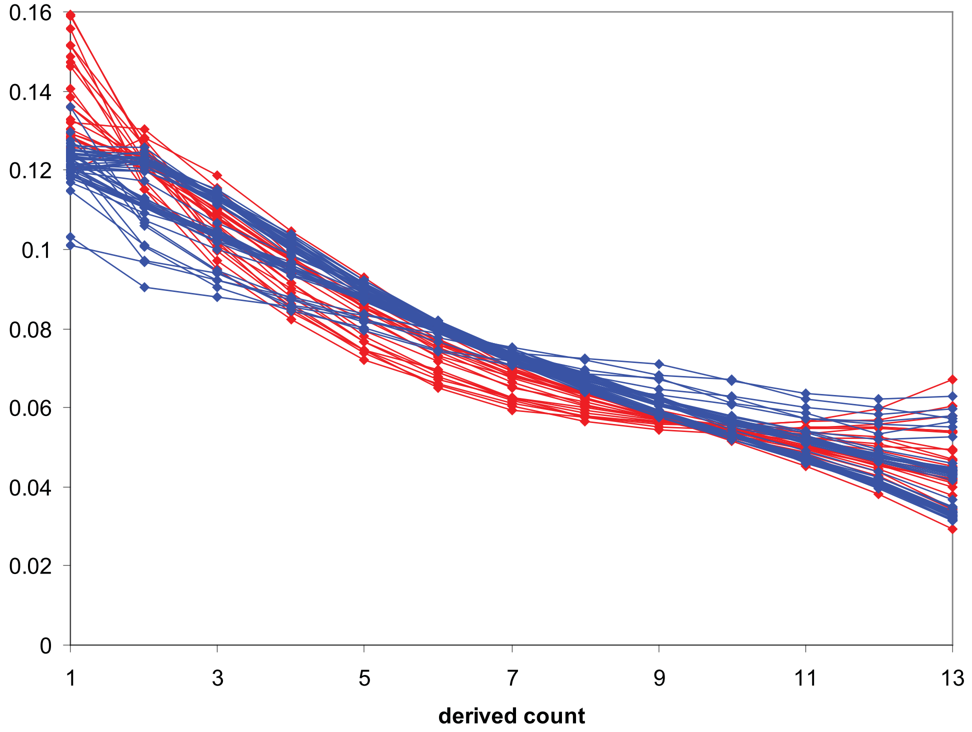

Figure S60: Frequency spectra for sub-Saharan African populations (red) and non-sub-saharan African populations (blue) based on ∼270k SNPs.

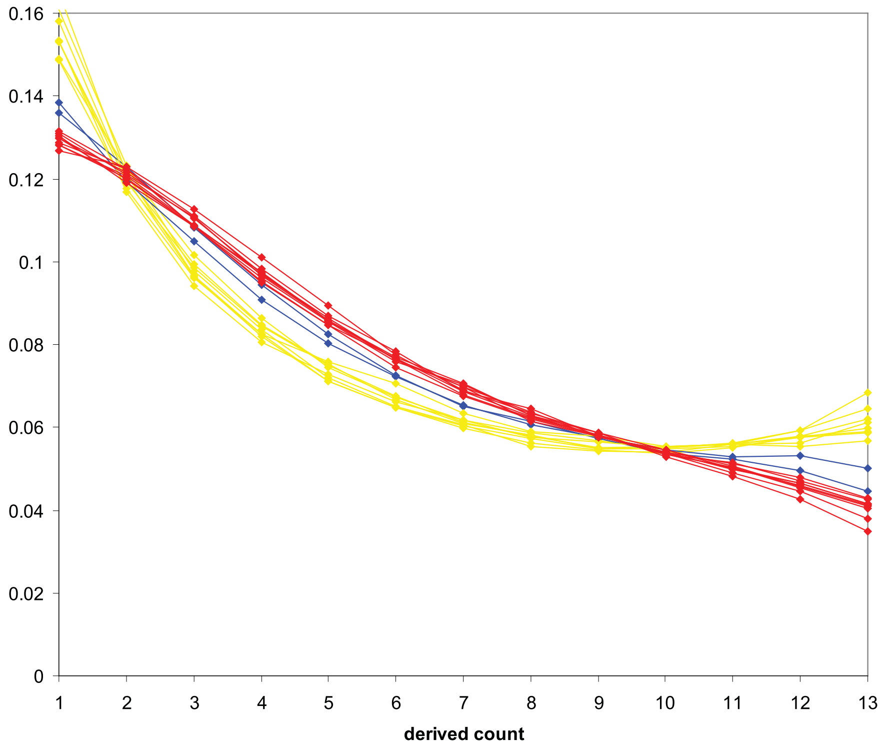

Figure S61: Frequency spectra for African populations based on ∼270k SNPs. The populations (red) displaying the most "flat" spectra are West and East African populations (or Bantu-speaking populations from Southern African of West African origin): Sandawe, Luhya, Bantu-speakers (Kenya – HGDP), Bantu-speakers (Southern Africa), Herero, Yoruba (HapMap), Mandenka, Yoruba (HGDP), and Massai. Intermediate spectra are found for the Biaka and Khwe populations (blue), which are hypothesized to have substantial (non-recent) admixture with Bantu-speaking groups. The indigenous Southern African Khoe-San populations (≠Khomani (A), ≠Khomani (B), /Gui and //Gana, Ju/'hoansi, Karretjie, Nama, !Xun) and the central African Mbuti display the least flat frequency spectra (yellow).

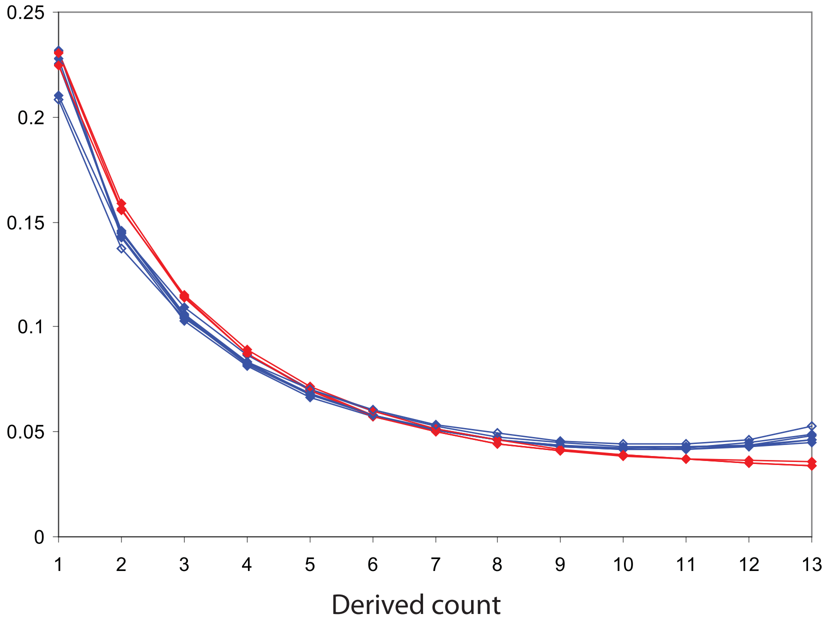

Figure S62: Frequency spectra for Southern African populations and based on ∼2.3M SNPs. Ju/'hoansi, Nama, ≠Khomani (A), Karretjie, !Xun, /Gui and //Gana (blue), Bantu-speakers (Southern Africa), Herero, and Khwe. Note that the Bantu-speaking populations and the Khoe-San populations show much more similar frequency spectra for this 10 times larger set of SNPs, compare *e.g.* with Figure S61, possibly an indication of that for the larger set of SNPs, the effect on the Bantu-speaking populations is less pronounced.

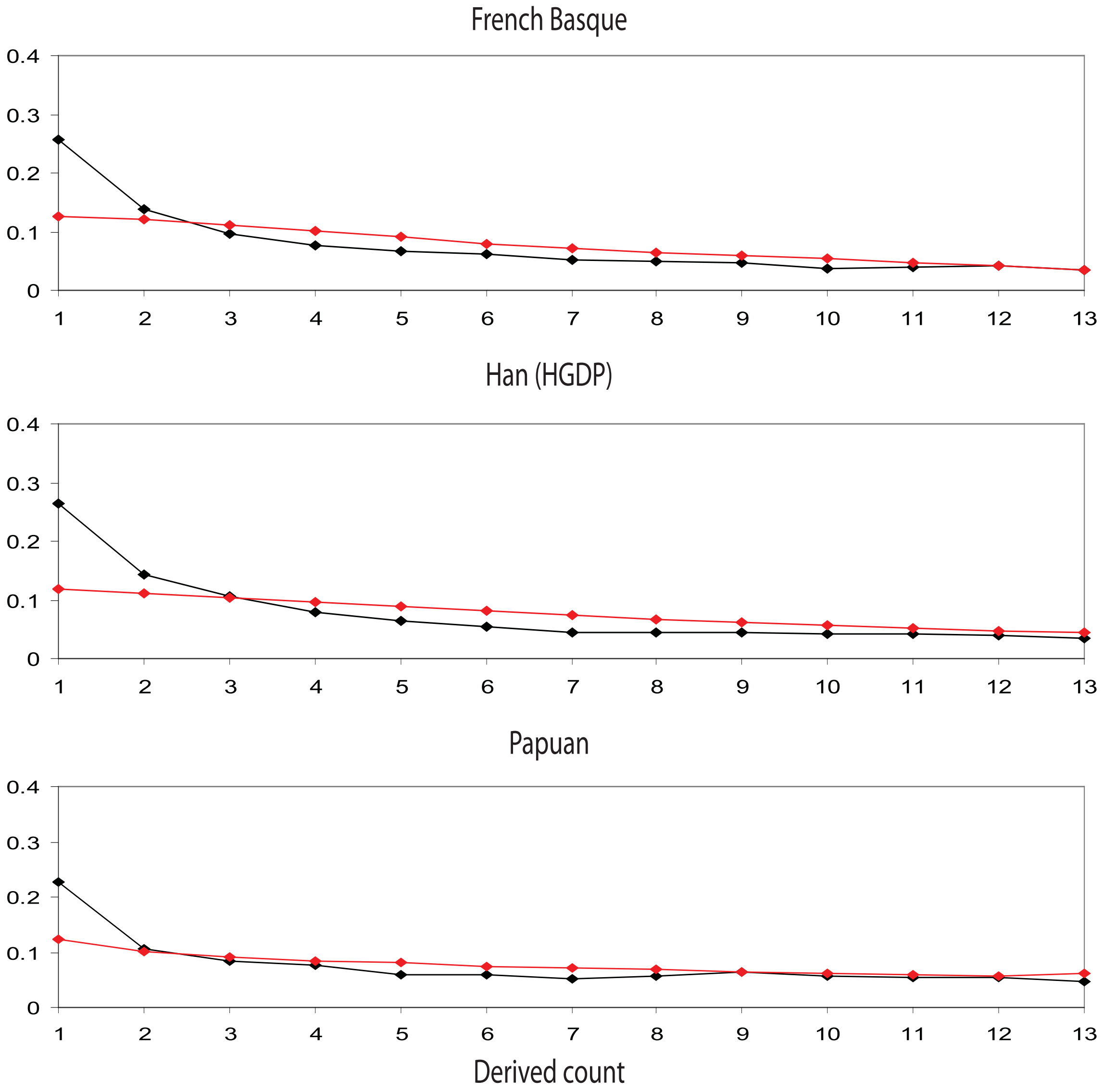

Figure S63: Comparison of frequency spectra for non-African populations. Spectra from sequence data is shown in black, and spectra from ∼270k SNP data is shown in red.

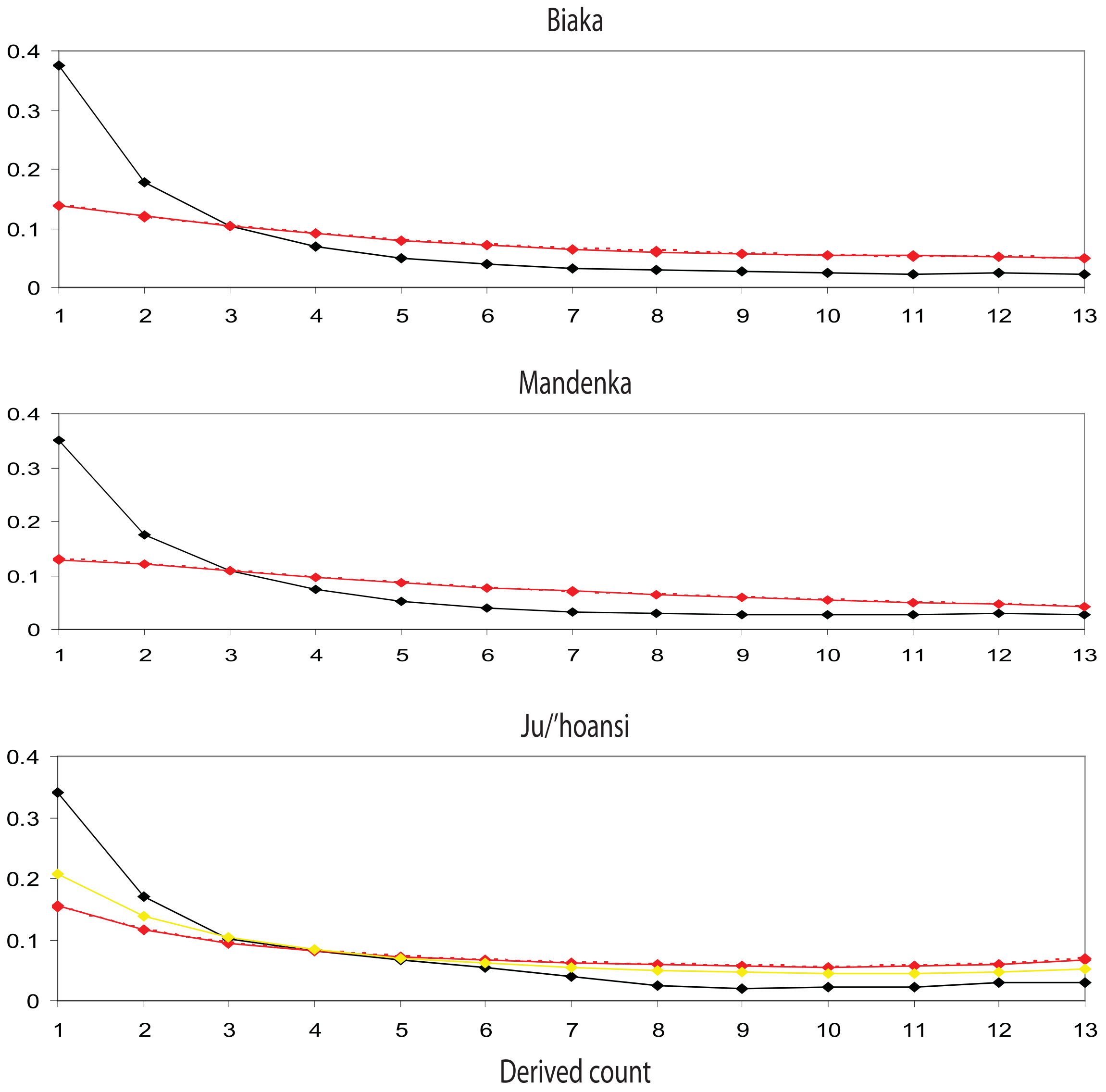

Figure S64: Comparison of frequency spectra for African populations. Spectra from sequence data is shown in black, spectra from ∼270k SNP data is shown in red, and spectra from ∼2.3M SNP data is shown in yellow. Including admixed individuals has basically no effect on the spectra (dashed vs solid line).

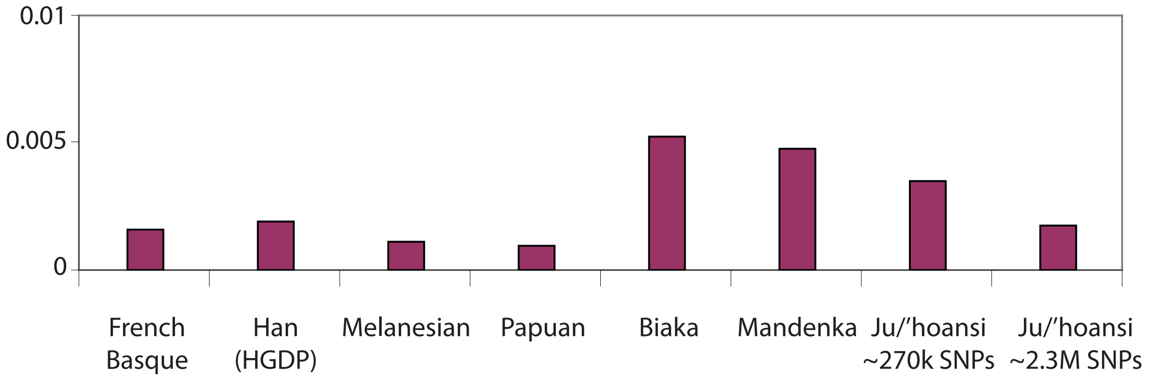

Figure S65: The Sum of Squared Differences (SSD) between the HOMINID sequence data and SNP-chip data.

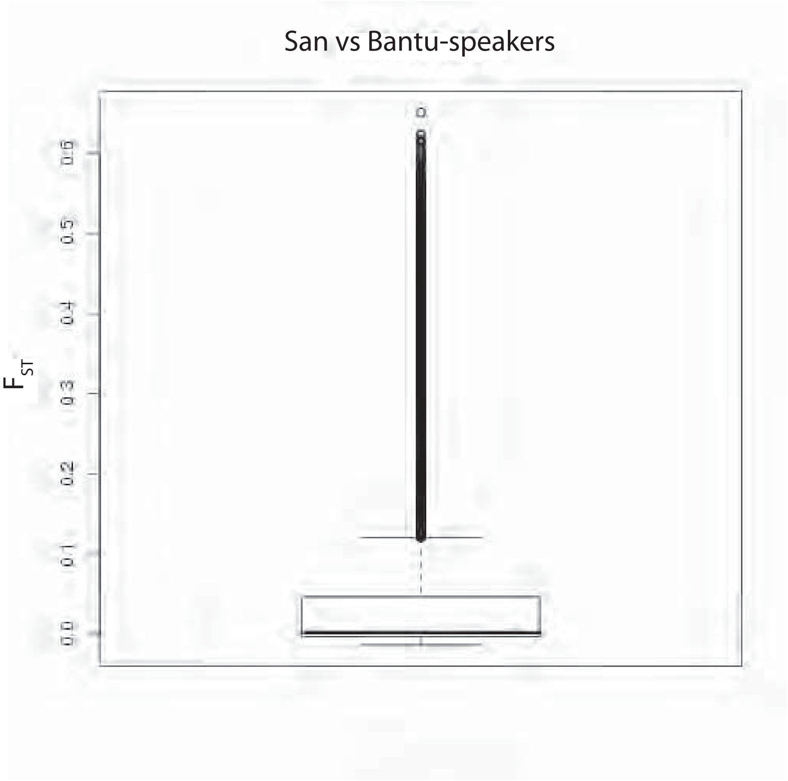

Figure S66: Box plot with the distribution of the F_{ST} values for Khoe-San vs Bantu-speakers

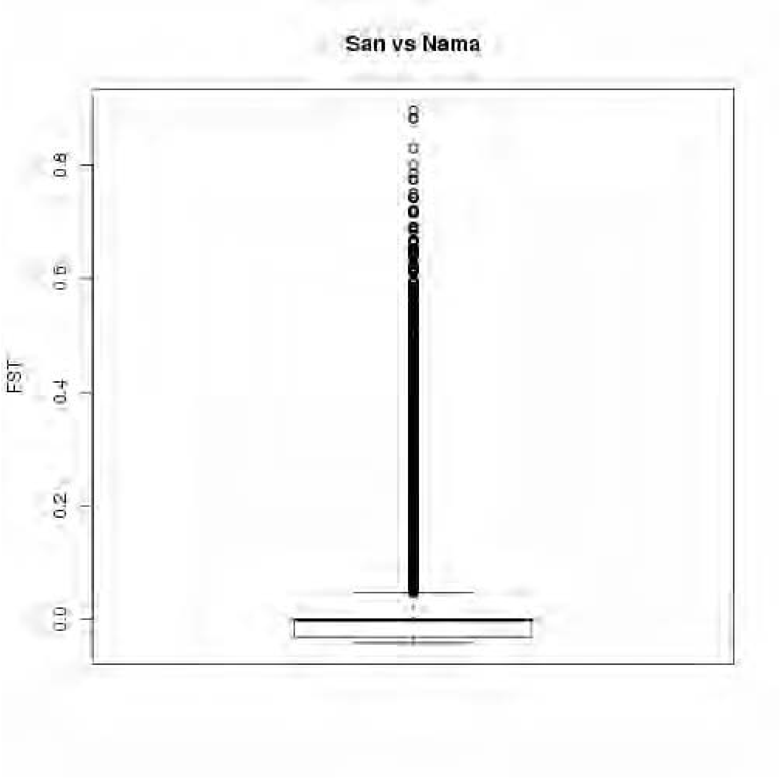

Figure S67: Box plot with the distribution of the F_{ST} values for Nama vs San.

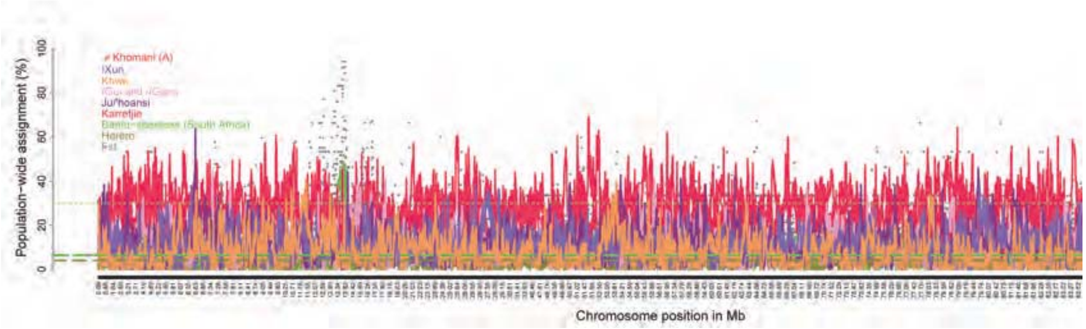

Figure S68: The chromosome-wide assignment of Nama to each of the eight donor populations. F_{ST} is represented as grey dots, population assignment over chromosome regions is shown with donor population colors according to the legend. Thick broken lines are chromosome-wide averages. Thin broken green line is upper 99th percentile line for assignments to the Bantu-speakers (South Africa) group

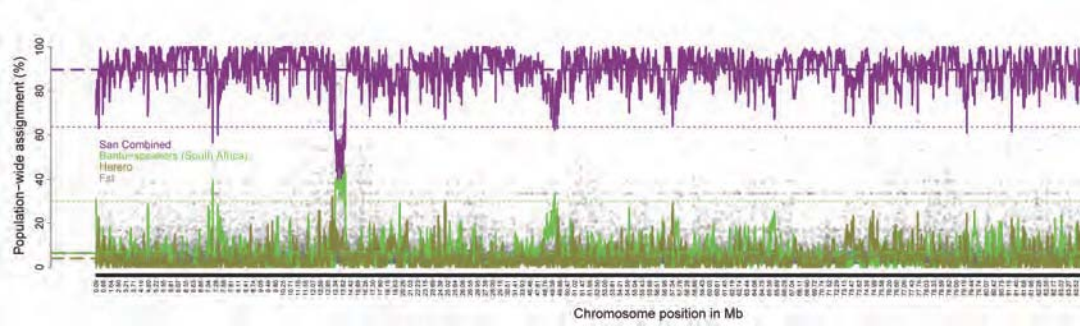

Figure S69: The chromosome-wide assignment of Nama to the 6 San donor population summarized (purple) and the two Bantu-speaking populations (green). F_{ST} is represented as grey dots. Thick broken lines are chromosome-wide averages. Thin broken green line is 99th percentile line for assignments to Bantu-speakers (South Africa) group. Thin broken purple line is lower 99th percentile line for assignments to San groups.

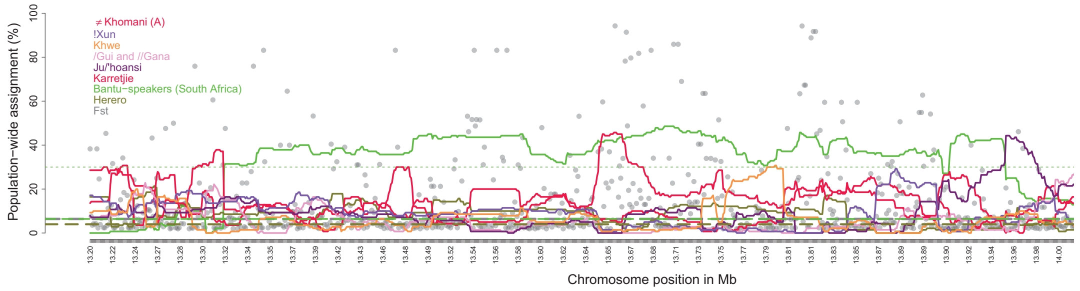

Figure S70: The enlarged view of the region of interest showing the assignment of Nama to each of the eight donor populations separately. F_{ST} is represented as grey dots, population assignment over chromosome regions is shown with donor population colors according to the legend. Thick broken lines are chromosome-wide averages. Thin broken green line is upper 99th percentile line for assignments to Bantu-speakers (South Africa) group.

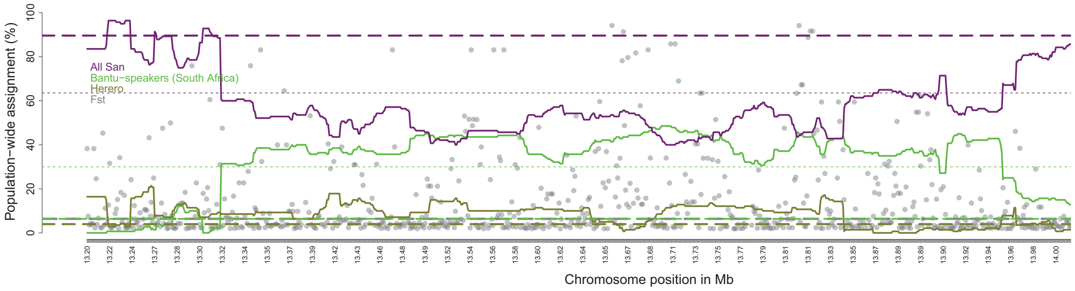

Figure S71: The enlarged view of the region of interest showing the assignment of Nama to each of the donor populations with San donor populations summarized into one value. F_{ST} is represented as grey dots, population assignment over chromosome regions is shown with donor population colors according to the legend. Thick broken lines are chromosome-wide averages. Thin broken green line is 99th percentile line for assignments to Bantu-speakers (South Africa) group. Thin broken purple line is lower 99th percentile line for assignments to San groups.

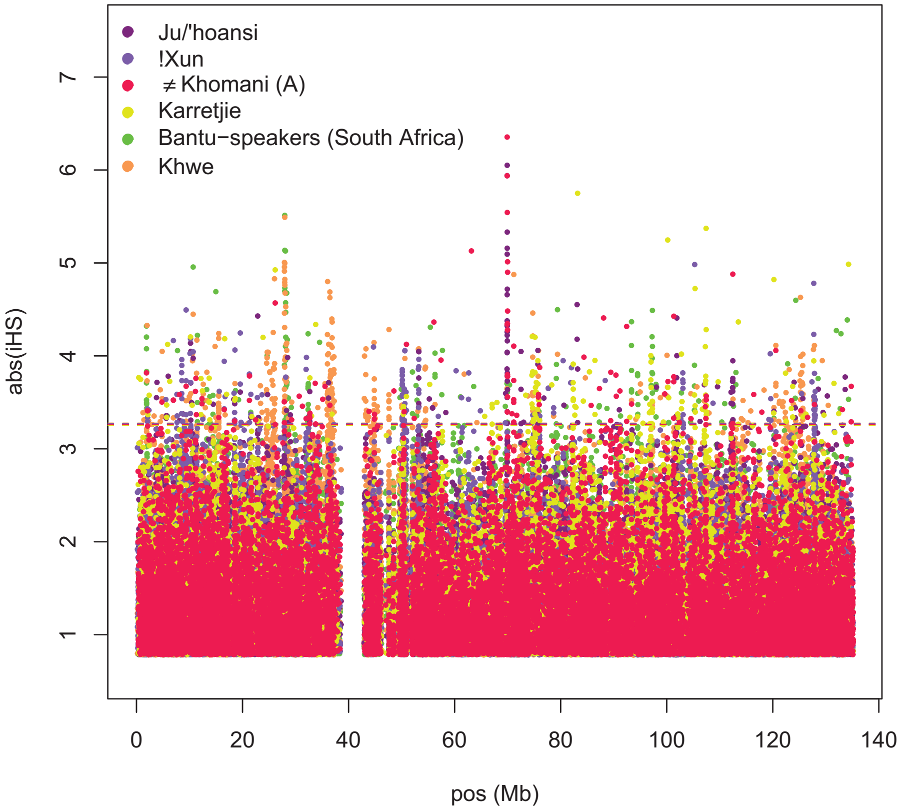

Figure S72: $|iHS|$ values for Khoe-San and Bantu-speaking populations for chromosome 10

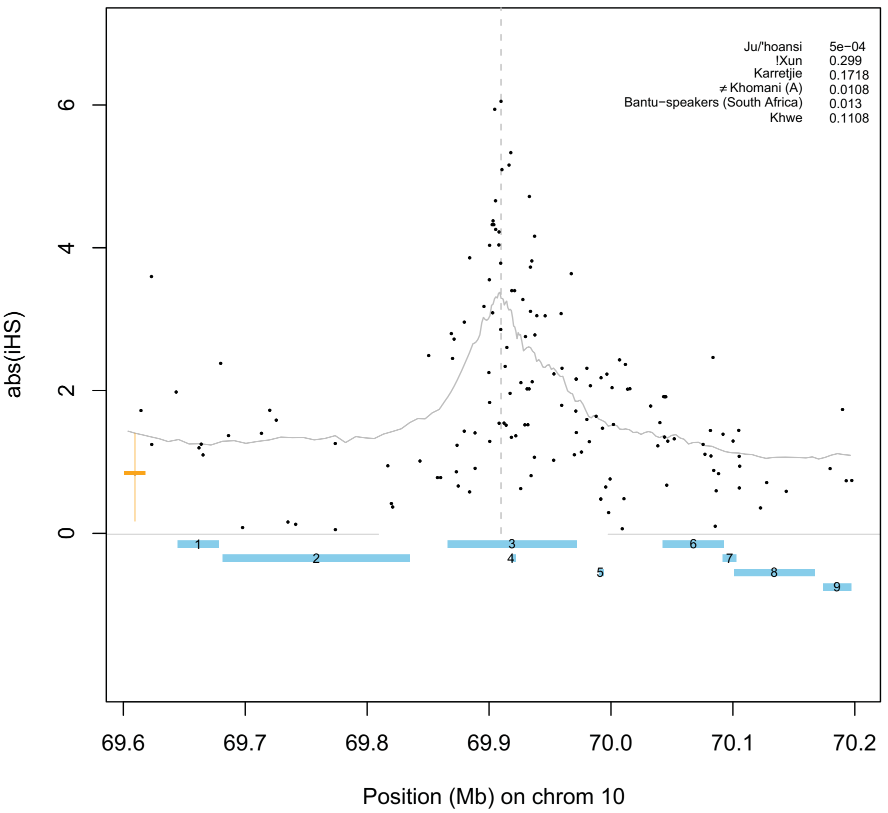

Figure S73: $|iHS|$ for each SNP (black dots) from position 69.6Mb to 70.2Mb on chromosome 10 in Ju/'hoansi surrounding the muscle gene *MYPN*. The vertical dashed line shows the position of the maximum $|iHS|$ value. The orange horizontal bars give the genome wide $|iHS|$ average for the particular population and the vertical lines indicates the standard deviation. The empirical p-values for 200kb regions centered on the peak are also given for each population. The locations of genes in the region are shown by blue rectangles and the gene names are given in Table S17.

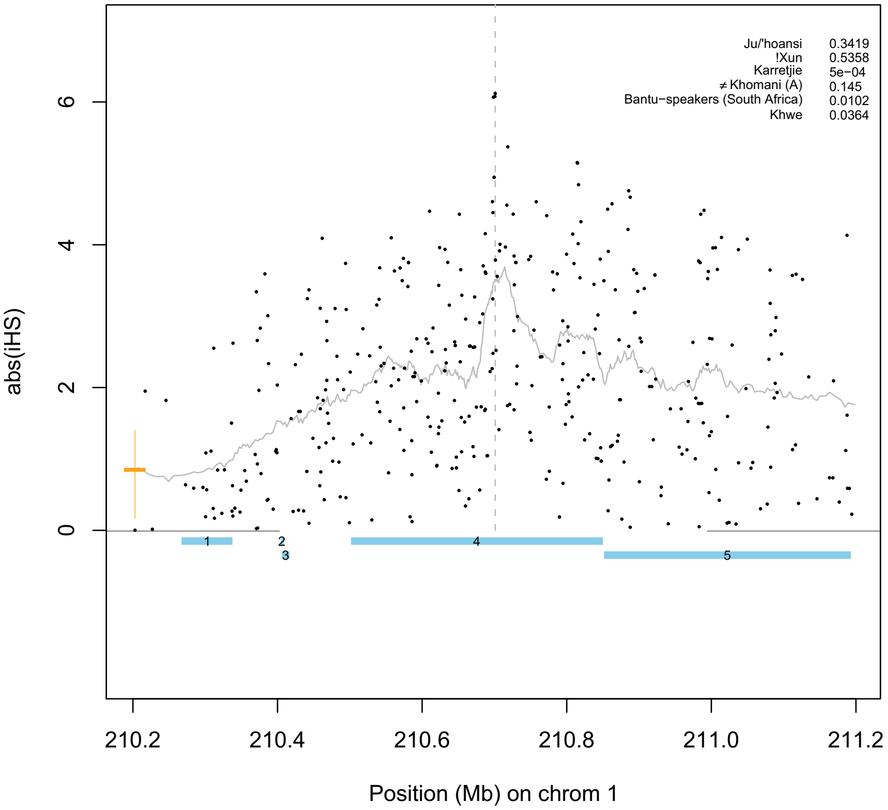

Figure S74: The genes on chromosome 1 from position 210.4-211 Mb corresponding to a candidate windows in Karretjie. The vertical dashed line shows the position of the maximum $|iHS|$ value. The orange horizontal bars give the genome wide $|iHS|$ average for the particular population and the vertical lines indicates the standard deviation. The empirical p-values for 200kb regions centered on the peak are also given for each population. The locations of genes in the region are shown by blue rectangles and the gene names are given in Table S18.

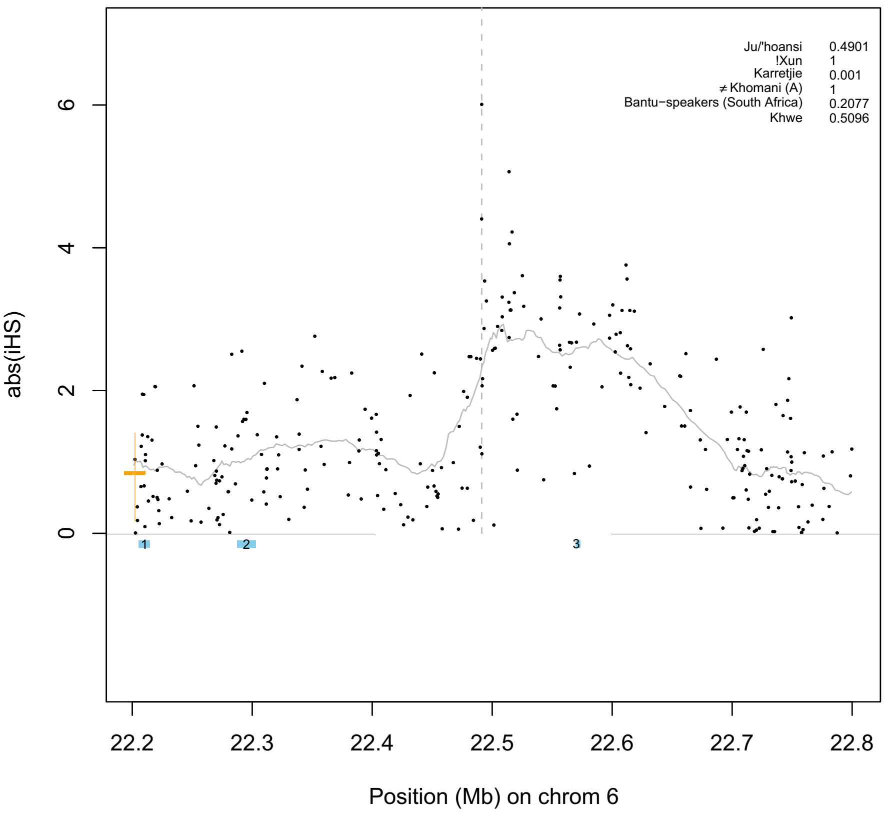

Figure S75: The genes on chromosome 6 from position 22.4-22.6Mb corresponding to a candidate window in Karretjie. The vertical dashed line shows the position of the maximum $|iHS|$ value. The orange horizontal bars give the genome wide $|iHS|$ average for the particular population and the vertical lines indicates the standard deviation. The empirical p-values for 200kb regions centered on the peak are also given for each population. The locations of genes in the region are shown by blue rectangles and the gene names are given in Table S19.

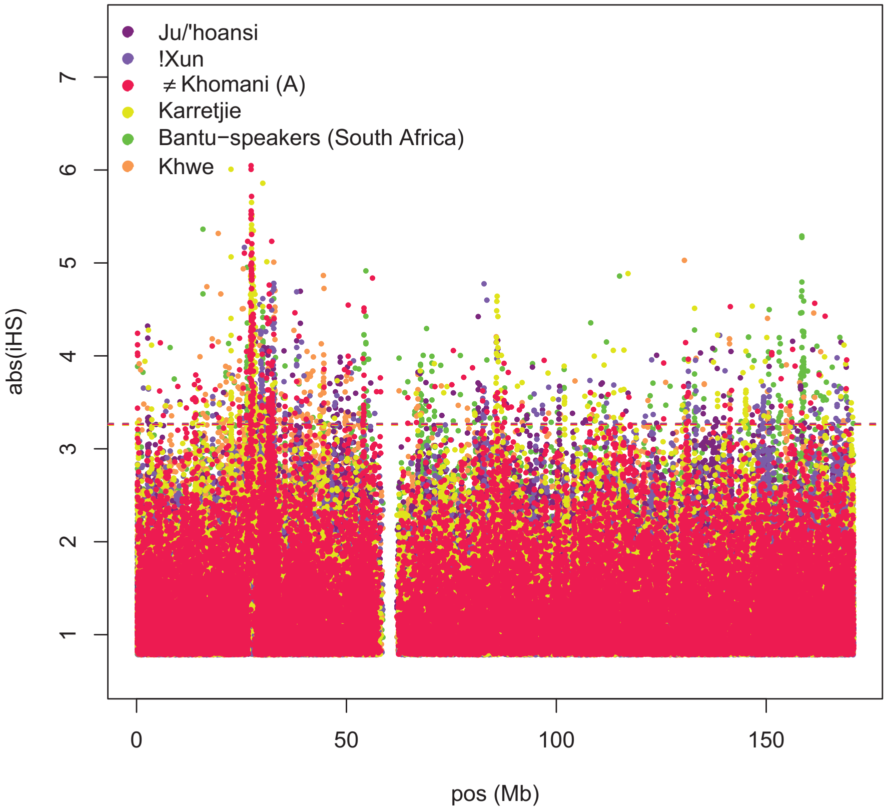

Figure S76: $|iHS|$ values for Khoe-San and Bantu-speaking populations for chromosome 6

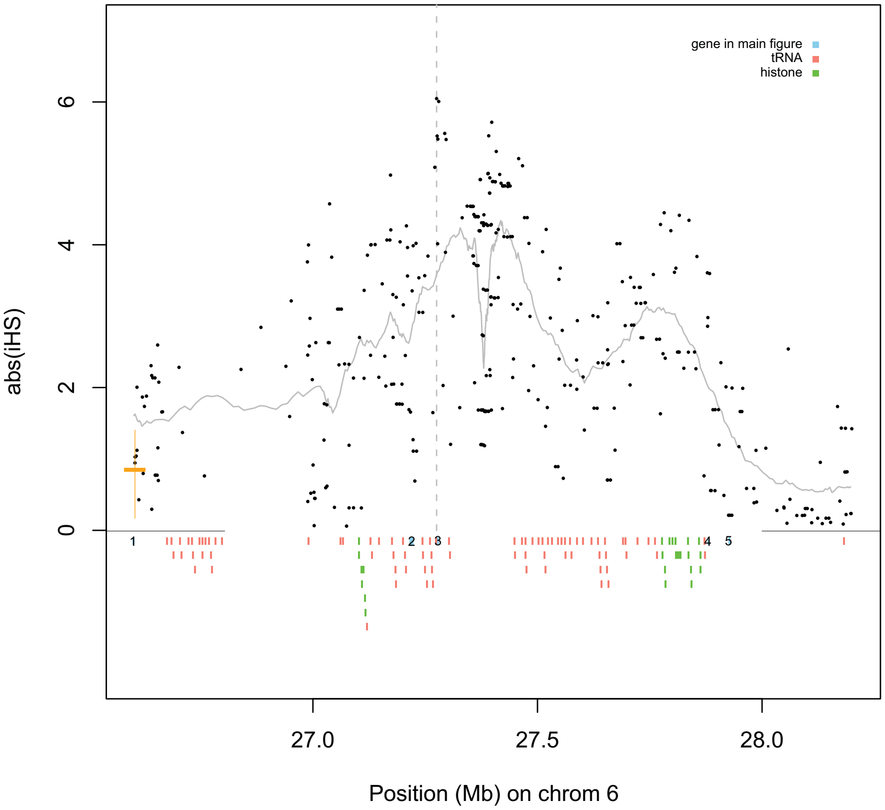

Figure S77: The genes on chromosome 6 from position 26.8-28.0Mb corresponding to a candidate window in ≠Khomani (A) surrounding the immune system genes PRSS16 and POM121L2, including the numerous tRNA (n=83) and histone (n=21) genes in this region. The vertical dashed line shows the position of the maximum $|iHS|$ value. The orange horizontal bars give the genome wide $|iHS|$ average for the particular population and the vertical lines indicates the standard deviation. The empirical p-values for 200kb regions centered on the peak are also given for each population. The locations of genes in the region are shown by blue rectangles and the gene names are given in Table S20.

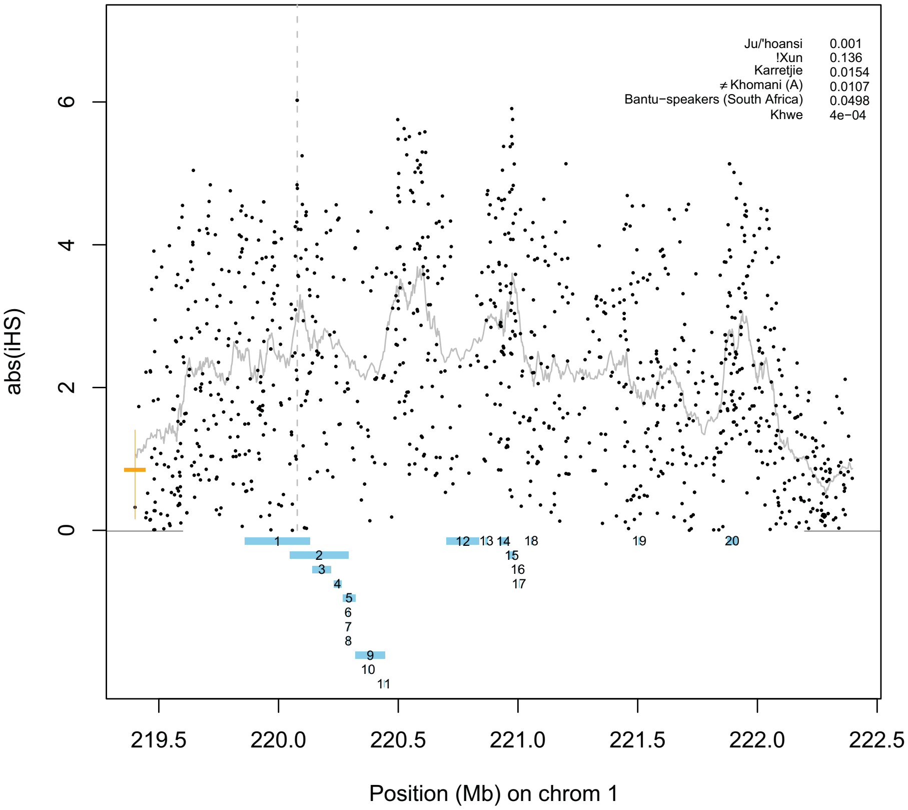

Figure S78: The genes on chromosome 1 from position 219.6-222.2Mb corresponding to a candidate windows in Khwe. The vertical dashed line shows the position of the maximum $|iHS|$ value. The orange horizontal bars give the genome wide $|iHS|$ average for the particular population and the vertical lines indicates the standard deviation. The empirical p-values for 200kb regions centered on the peak are also given for each population. The locations of genes in the region are shown by blue rectangles and the gene names are given in Table S21.

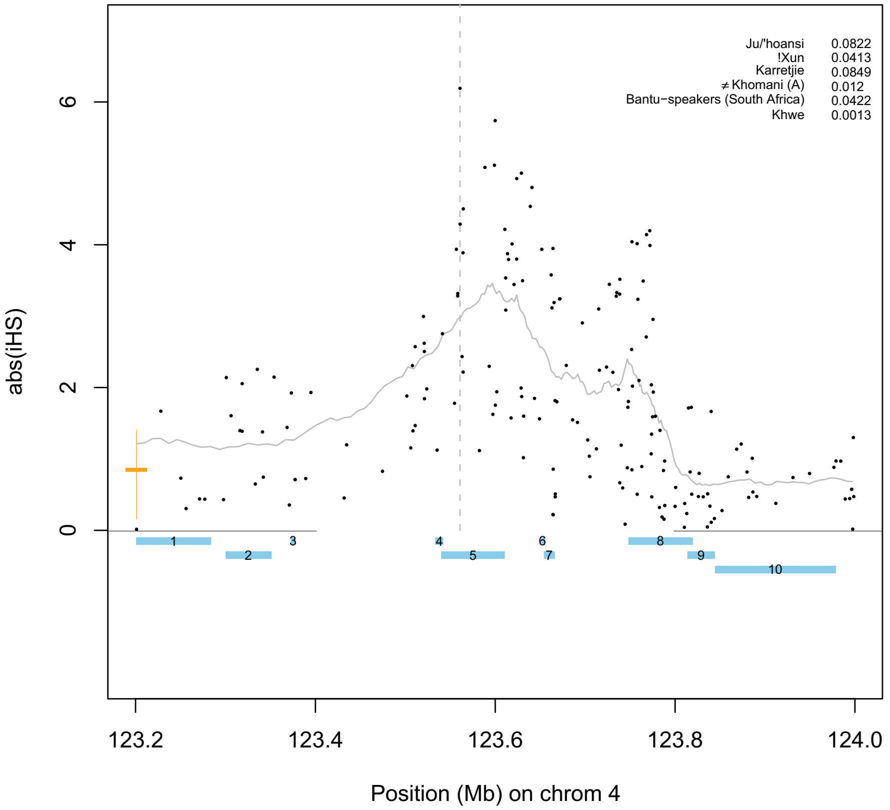

Figure S79: The genes on chromosome 4 from position 123.4-123.8Mb corresponding to a candidate window in Khwe. The vertical dashed line shows the position of the maximum $|iHS|$ value. The orange horizontal bars give the genome wide $|iHS|$ average for the particular population and the vertical lines indicates the standard deviation. The empirical p-values for 200kb regions centered on the peak are also given for each population. The locations of genes in the region are shown by blue rectangles and the gene names are given in Table S22.

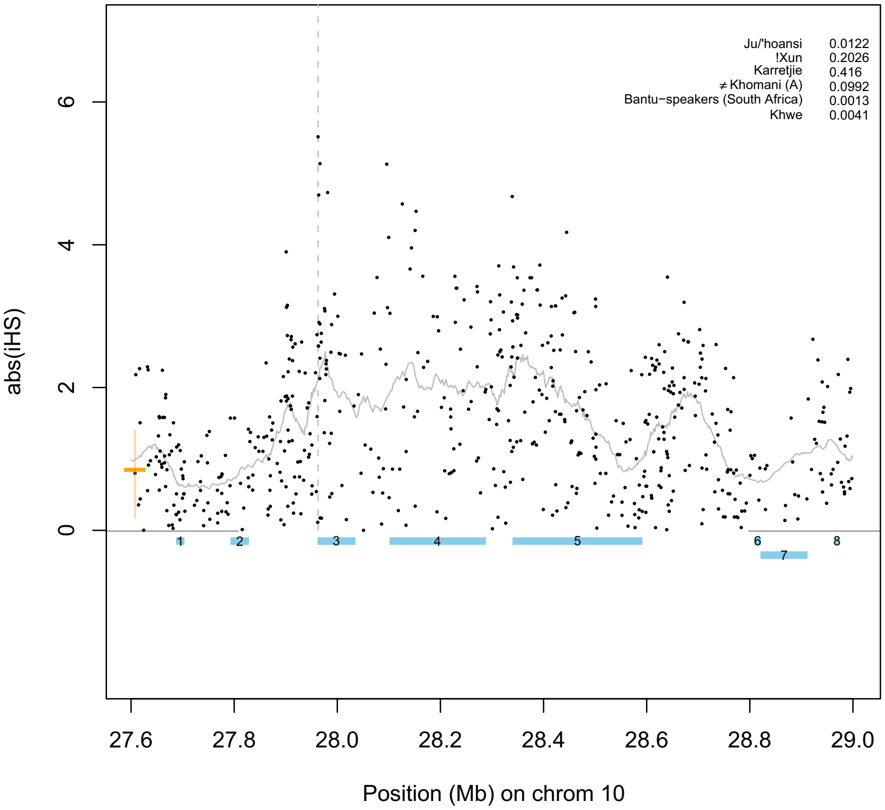

Figure S80: The genes on chromosome 10 from position 27.8-28.8Mb corresponding to a candidate window in Bantu-speakers (South Africa). The vertical dashed line shows the position of the maximum $|iHS|$ value. The orange horizontal bars give the genome wide $|iHS|$ average for the particular population and the vertical lines indicates the standard deviation. The empirical p-values for 200kb regions centered on the peak are also given for each population. The locations of genes in the region are shown by blue rectangles and the gene names are given in Table S23.

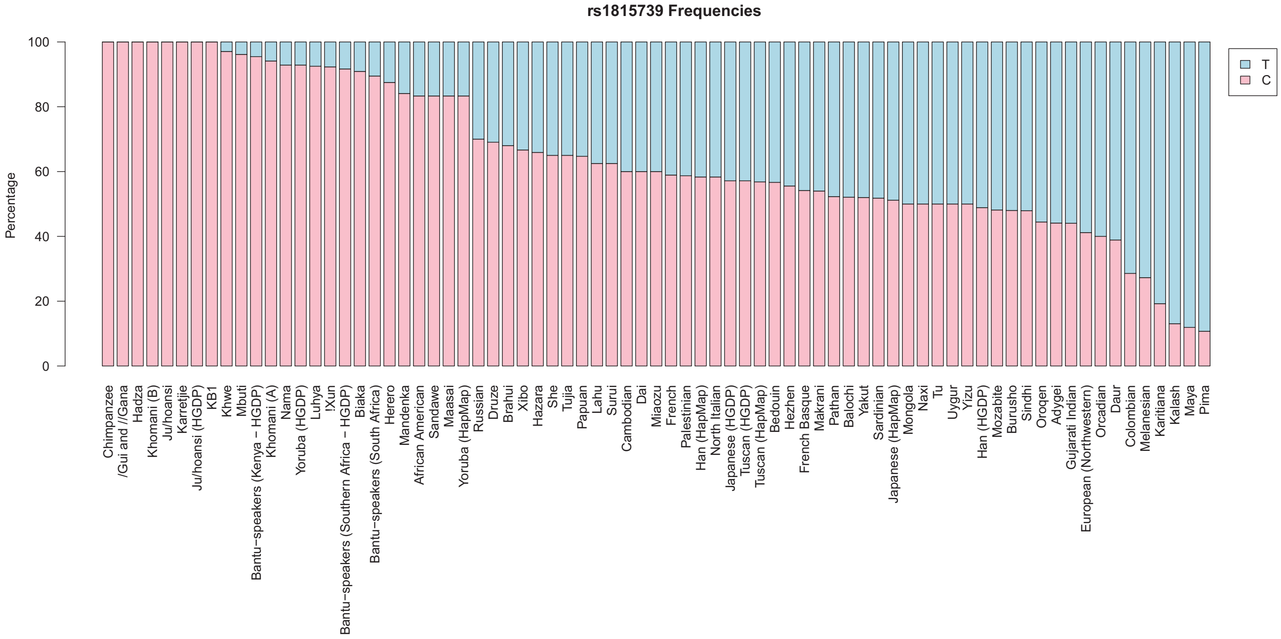

Figure S81: Frequencies of the rs1815739 C variant sorted in decreasing order.

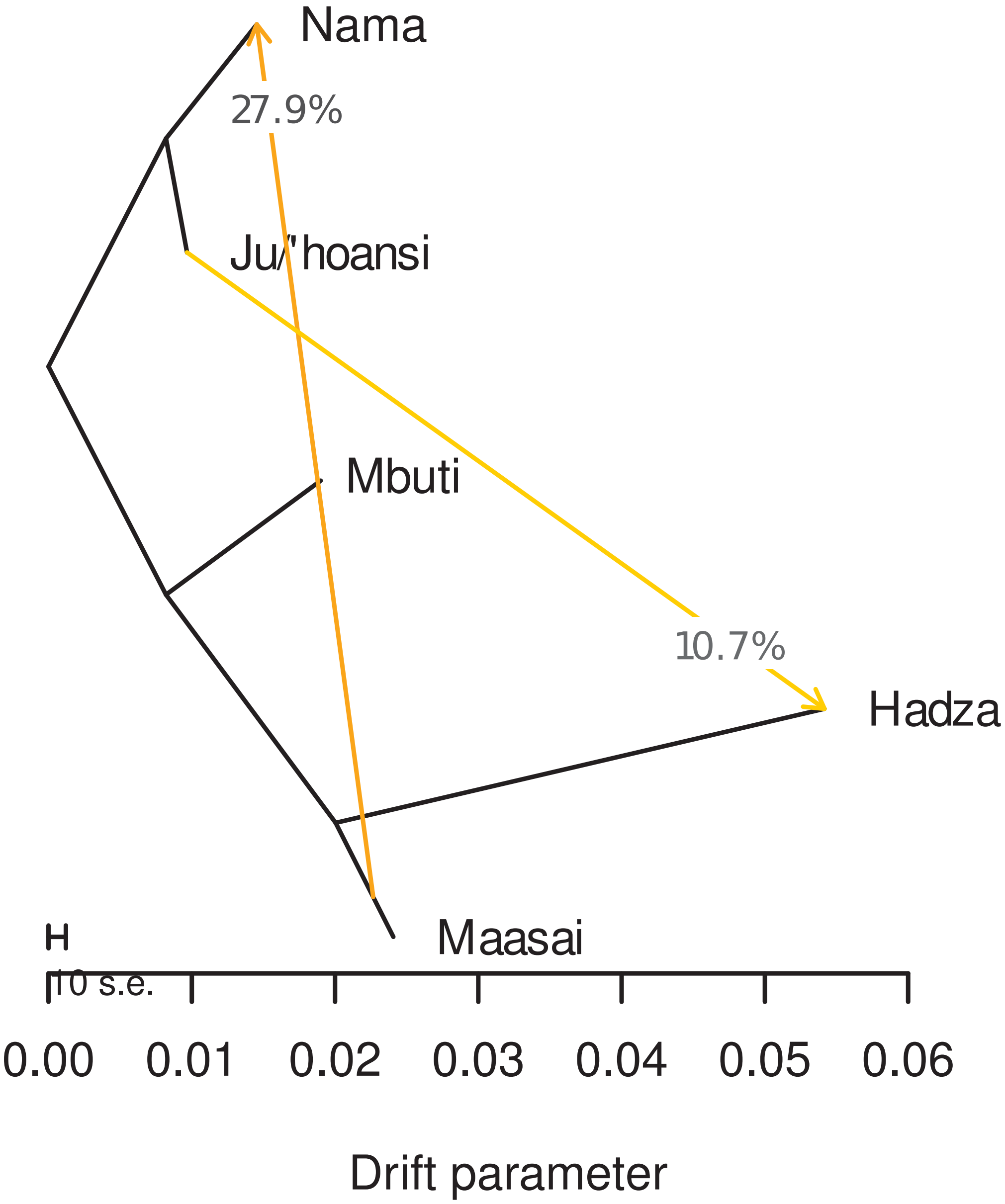

Figure S82: Ancestry graph estimated using `TREEMIX` suggesting gene flow between East African groups and Khoe-San groups

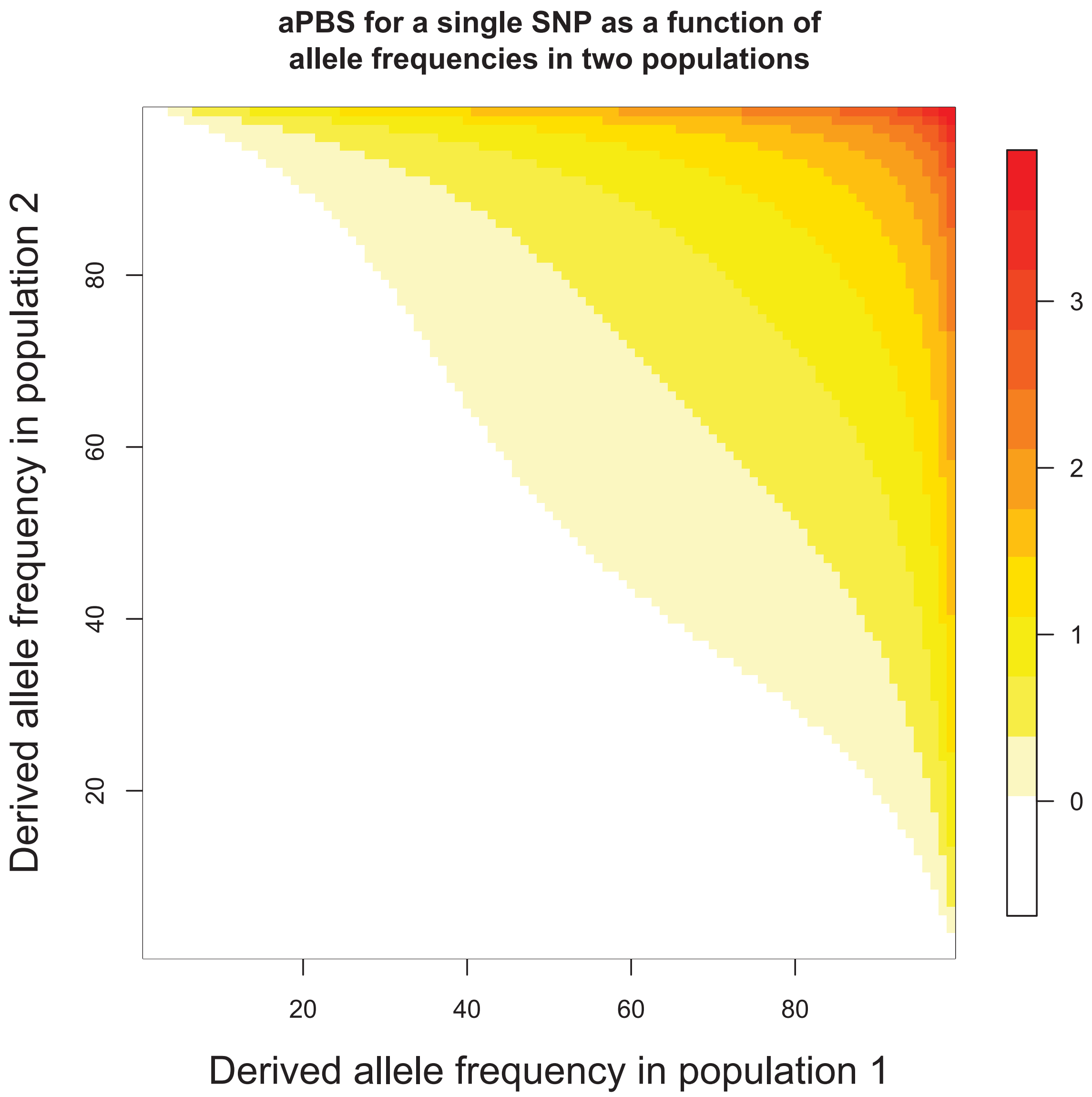

Figure S83: The *aPBS* statistic as a function of allele frequencies in the two descendant populations. Consecutive runs of SNPs with high frequency in two diverged populations are expected to be rare genome-wide, and can be used as a footprint of selective sweeps in the ancestral population. The *aPBS* is designed to capture this type of pattern.

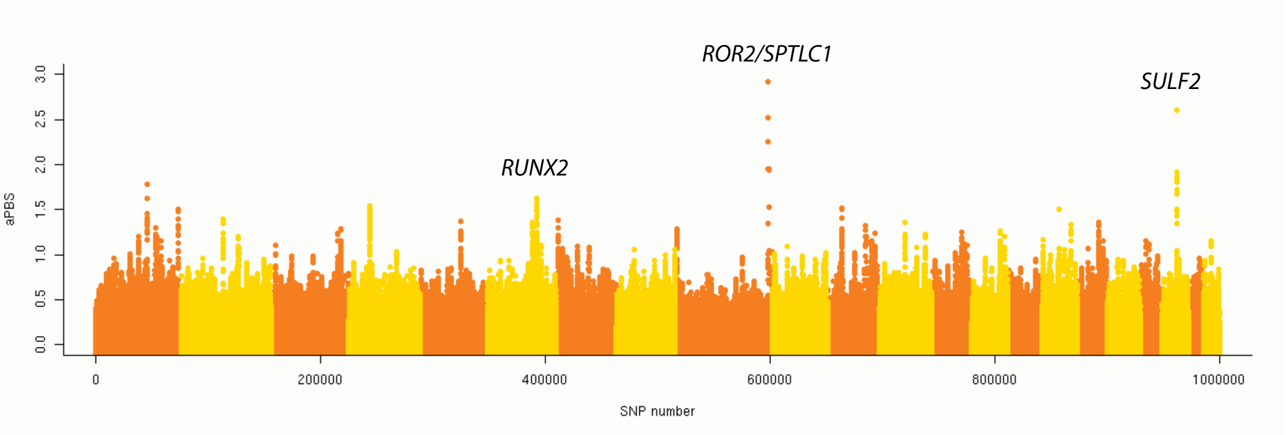

Figure S84: Manhattan plot of $aPBS$ values for each SNP, which were computed in a window extending 10 SNPs to each side (total 21 SNPs in each window).

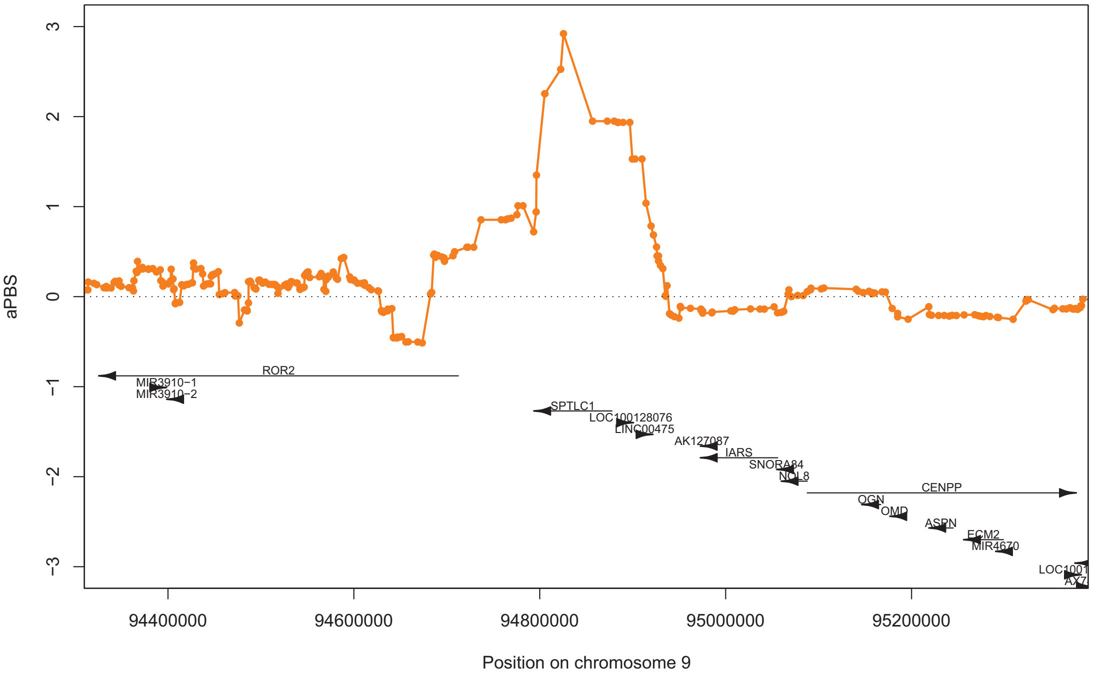

Figure S85: The top candidate region for selective sweeps in the ancestors of modern humans identified by the $aPBS$ statistic.

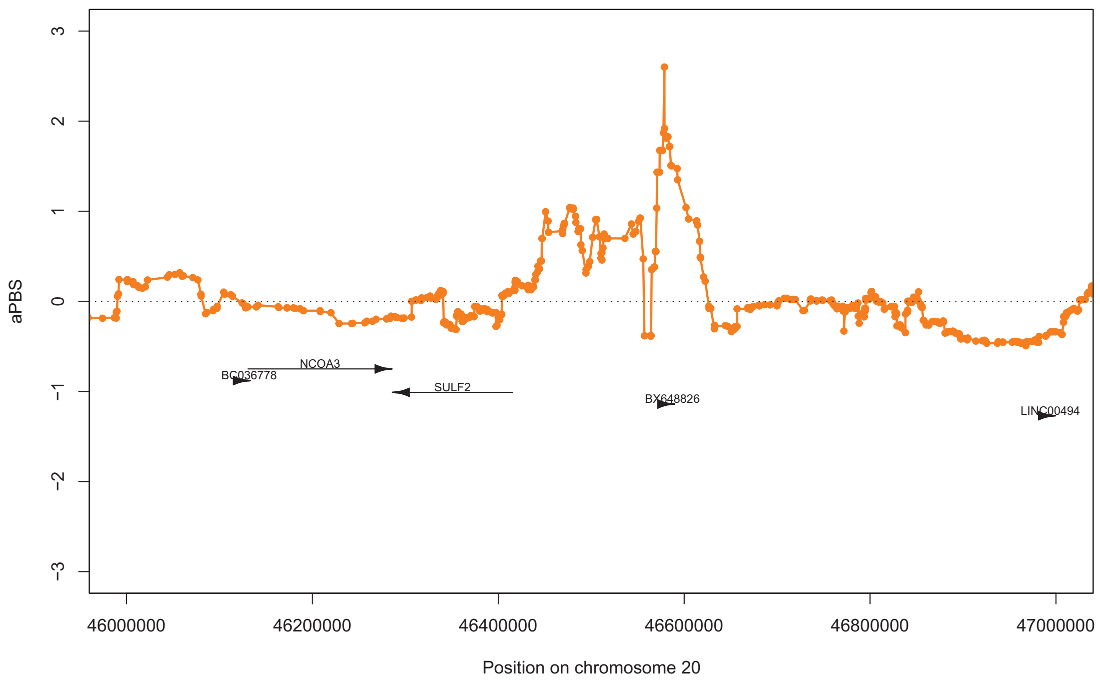

Figure S86: The second top candidate region for selective sweeps in the ancestors of modern humans identified by the $aPBS$ statistic.

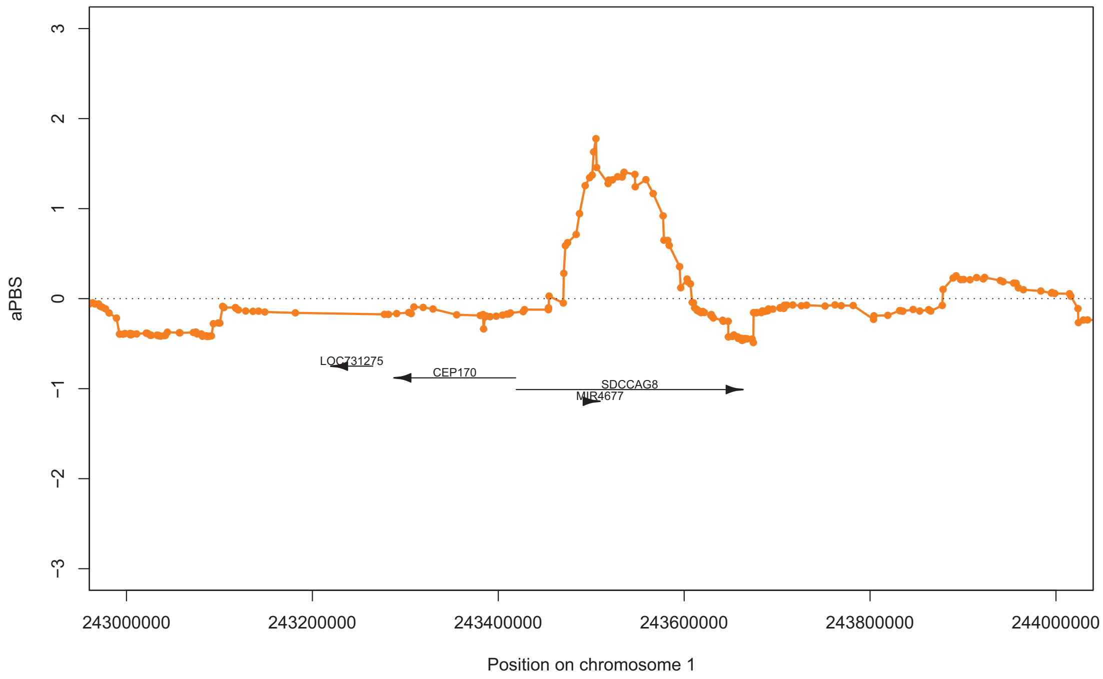

Figure S87: The third top candidate region for selective sweeps in the ancestors of modern humans identified by the $aPBS$ statistic.

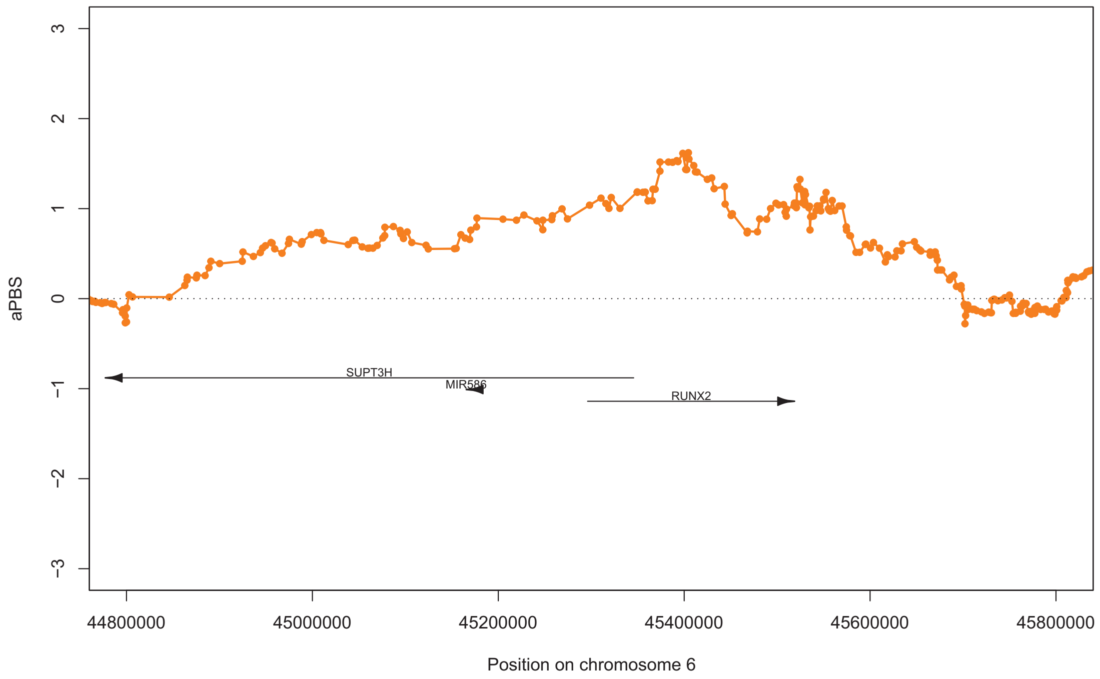

Figure S88: The fourth top candidate region for selective sweeps in the ancestors of modern humans identified by the $aPBS$ statistic.

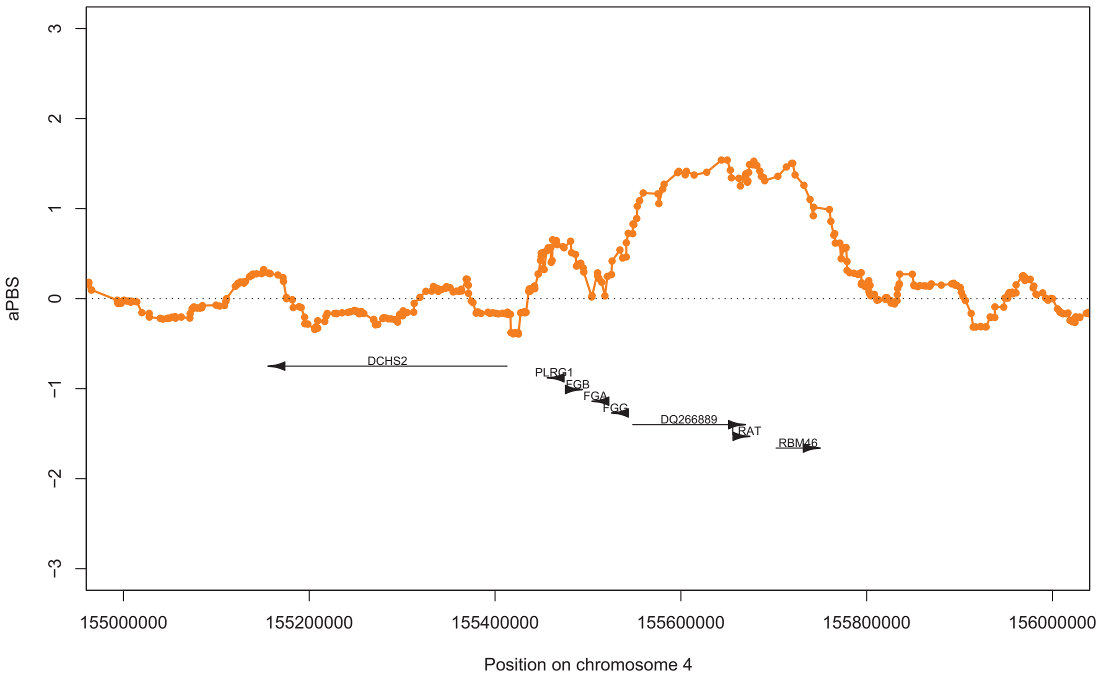

Figure S89: The fifth top candidate region for selective sweeps in the ancestors of modern humans identified by the $aPBS$ statistic.

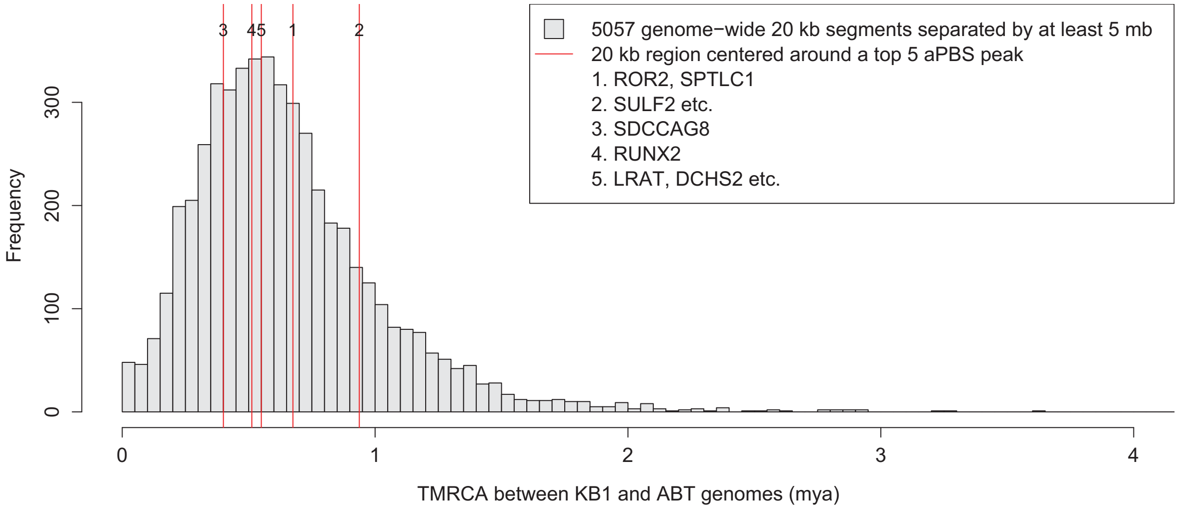

Figure S90: Time to most recent common ancestor (TMRCA) of 20 kb regions in the KB1 and ABT genomes.

Tables

Table S1: A description of sample groups, including group name, group membership, place of sampling and origin, and number of individuals. Country Abbreviations: AN - Angola, BT - Botswana, NM - Namibia, and SA - South Africa. A detailed description of the populations and the sampling process is described in section 2.

Ethnic group name	Main population group	Place of sampling	Place of origin	Geno-	QC	Adm.
		(Country)	(If different from sampling)	typed		rem.
Karretjie People	San	Colesberg (SA)		20	20	12
Colesberg Coloured	Coloured	Colesberg (SA)		20	20	0
Wellington Coloured	Coloured	Wellington (SA)		20	20	0
Nama	Khoe	Windhoek (NM)		20	20	7
/Gui and //Gana	San	Kutse Game reserve (BT)		20	15	7
Ju/'hoansi	San	Tsumkwe (NM)		20	18	17
!Xun	San	Omega camp (NM), Schmidtsdrift (SA)	Around Menongue (AN)	20	19	13
Khwe	San	Omega camp (NM), Schmidtsdrift (SA)	Caprivi strip (NM, AN, BT)	18	17	17
≠Khomani	San	Askham (SA)		40	39	17
Herero	Southwestern Bantu-speakers	Windhoek (NM)		12	12	8
Bantu-speakers (S. Africa)	Southeastern Bantu-speakers (Zulu, Southern Sotho, Tswana)	Various (SA)		20	20	19
Total				230	220	100

Table S2: Number of individuals removed due to admixture.

Group	Panel	Comment	Removed inds.	Remaining inds.
/Gui and //Gana	KSP	Removed due to European and Bantu-speaking admixture	8	7
Ju/'hoansi	KSP	Removed due to European and Bantu-speaking admixture	1	17
Karretjie	KSP	Removed due to European and Bantu-speaking admixture	8	12
≠Khomani (A)	KSP	Removed due to European and Bantu-speaking admixture	22	17
Nama	KSP	Removed due to European and Bantu-speaking admixture	13	7
!Xun	KSP	Removed due to European and Bantu-speaking admixture	6	13
Bantu-speakers (S. Africa)	KSP	Removed due to European and Khoe-San admixture	1	19
Herero	KSP	Removed due to European and Khoe-San admixture	4	8
Coloured (Colesberg)	KSP	Whole group removed	20	0
Coloured (Wellington)	KSP	Whole group removed	20	0
Hadza	HENN	Removed due to European and Bantu-speaking admixture	10	7
≠Khomani (B)	HENN	Removed due to European and Bantu-speaking admixture	16	11
Sandawe	HENN	Removed due to European and Bantu-speaking admixture	4	24
Total			133	142

Table S3: Population information.

Panel	Population	Country	Region	Sample size	Excl. adm.	Lat (N)	Long (E)
Henn	≠Khomani (B)	South Africa	AFRICA	27	11	-26.974138	20.794373
Hgdp	San	Namibia	AFRICA	2	2	-21	20
KSP	Khwe	Angola	AFRICA	17	17	-17.363921	22.950439
KSP	!Xun	Angola	AFRICA	19	13	-14.628943	17.666016
KSP	/Gui and //Gana	Botswana	AFRICA	15	7	-23.650898	24.6698
KSP	Ju/’hoansi	Namibia	AFRICA	18	17	-19.597399	20.494995
KSP	Nama	Namibia	AFRICA	20	7	-22.558559	17.072754
KSP	Karretjie	South Africa	AFRICA	20	12	-30.712638	25.101013
KSP	≠Khomani (A)	South Africa	AFRICA	39	17	-26.974138	20.794373
KB1	KB1	Namibia	AFRICA	1	1	-22.558559	17.072754
Hgdp	Biaka	Cent. Afr. Rep.	AFRICA	22	22	4	17
Hgdp	Mbuti	Congo	AFRICA	13	13	1	29
Henn	Hadza	Tanzania	AFRICA	17	7	-3.382555	36.68335
Henn	Sandawe	Tanzania	AFRICA	28	24	-6.180241	35.744019
HapMap	Luhya	Kenya	AFRICA	80	80	0.617603	34.762115
Hgdp	Bantu-speakers (Kenya)	Kenya	AFRICA	11	11	-3	37
Hgdp	Bantu-speakers (S. Africa - HGDP)	South Africa	AFRICA	6	6	-21	18.7
KSP	Herero	Namibia	AFRICA	12	8	-22.558559	17.072754
KSP	Bantu-speakers (South Africa)	South Africa	AFRICA	20	19	-26.256637	28.037109
HapMap	Yoruba (HapMap)	Nigeria	AFRICA	9	9	8	5
Hgdp	Yoruba (HGDP)	Nigeria	AFRICA	21	21	8	5
Hgdp	Mandenka	Senegal	AFRICA	22	22	12	-12
HapMap	Maasai	Kenya	AFRICA	57	57	-1.321172	36.826172
KSP	Coloured (Colesberg)	South Africa	AFRICA	20	0	-30.712638	25.101013
KSP	Coloured (Wellington)	South Africa	AFRICA	20	0	-33.643282	19.011841
HapMap	ASW [African American]	USA	NA	12	0	NA	NA
Hgdp	Mozabite	Algeria-Mzab	MIDDLE EAST	27	27	32	3
Hgdp	Druze	Israel-Carmel	MIDDLE EAST	42	42	32	35
Hgdp	Palestinian	Israel-Central	MIDDLE EAST	46	46	32	35
Hgdp	Bedouin	Israel-Negev	MIDDLE EAST	45	45	31	35
HapMap	GIH [Gujarati Indian-American]	India	C/S ASIA	84	84	23	72
Hgdp	Uygur	China	C/S ASIA	10	10	44	81
Hgdp	Xibo	China	C/S ASIA	9	9	43.5	81.5
Hgdp	Balochi	Pakistan	C/S ASIA	24	24	30.5	66.5
Hgdp	Brahui	Pakistan	C/S ASIA	25	25	30.5	66.5
Hgdp	Burusho	Pakistan	C/S ASIA	25	25	36.5	74
Hgdp	Hazara	Pakistan	C/S ASIA	22	22	33.5	70
Hgdp	Kalash	Pakistan	C/S ASIA	23	23	36	71.5
Hgdp	Makrani	Pakistan	C/S ASIA	25	25	26	64
Hgdp	Pathan	Pakistan	C/S ASIA	22	22	33.5	70.5
Hgdp	Sindhi	Pakistan	C/S ASIA	24	24	25.5	69
HapMap	CEU [European (Northwestern)]	CEPH	EUROPE	17	17	55	-3
HapMap	Tuscan (HapMap)	Italy	EUROPE	88	88	43	11
Hgdp	French	France	EUROPE	28	28	46	2
Hgdp	French Basque	France	EUROPE	24	24	43	0
Hgdp	North Italian	Italy	EUROPE	12	12	46	10
Hgdp	Sardinian	Italy	EUROPE	28	28	40	9
Hgdp	Tuscan (HGDP)	Italy	EUROPE	7	7	43	11
Hgdp	Orcadian	Orkney	EUROPE	15	15	59	-3
Hgdp	Russian	Russia	EUROPE	25	25	61	40
Hgdp	Adygei	Russia-Caucasus	EUROPE	17	17	44	39
HapMap	Han (HapMap)	China	EAST ASIA	84	84	32.5	114
HapMap	Japanese (HapMap)	Japan	EAST ASIA	82	82	38	138
Hgdp	Cambodian	Cambodia	EAST ASIA	10	10	12	105
Hgdp	Dai	China	EAST ASIA	10	10	21	100
Hgdp	Daur	China	EAST ASIA	9	9	48.5	124
Hgdp	Han	China	EAST ASIA	44	44	32.5	114
Hgdp	Hezhen	China	EAST ASIA	9	9	47.5	133.5
Hgdp	Lahu	China	EAST ASIA	8	8	22	100
Hgdp	Miaozu	China	EAST ASIA	10	10	28	109
Hgdp	Mongola	China	EAST ASIA	10	10	48.5	119
Hgdp	Naxi	China	EAST ASIA	8	8	26	100
Hgdp	Oroqen	China	EAST ASIA	9	9	50.5	126.5
Hgdp	She	China	EAST ASIA	10	10	27	119
Hgdp	Tu	China	EAST ASIA	10	10	36	101
Hgdp	Tujia	China	EAST ASIA	10	10	29	109
Hgdp	Yizu	China	EAST ASIA	10	10	28	103
Hgdp	Japanese	Japan	EAST ASIA	28	28	38	138
Hgdp	Yakut	Siberia	EAST ASIA	25	25	70	129.5
Hgdp	Melanesian	Bougainville	OCEANIA	11	11	-6	155
Hgdp	Papuan	New Guinea	OCEANIA	17	17	-4	143
Hgdp	Karitiana	Brazil	AMERICA	13	13	-10	-63
Hgdp	Surui	Brazil	AMERICA	8	8	-11	-62
Hgdp	Colombian	Colombia	AMERICA	7	7	3	-68
Hgdp	Maya	Mexico	AMERICA	21	21	19	-91
Hgdp	Pima	Mexico	AMERICA	14	14	29	-108
-	chimpanzee	-	-	1	1	NA	NA
-	Human ancestor	-	-	1	1	NA	NA
TOTAL				1745	1612		

Table S4: Description of sample groups.

Group – additional classification	Country of origin	Main language group	Language subgroup	Subsistence	Sample size ([1])
Ju/'hoansi – San	Namibia	Khoisan-Ju	Southeast	Hunter-Gatherer	18 (17)
!Xun – San	Angola	Khoisan-Ju	Northwest	Hunter-Gatherer	19 (13)
/Gui and //Gana – San	Botswana	Khoisan-Khoe	Kalahari	Hunter-Gatherer	15 (7)
Karretjie – San, Khoe	South Africa	Khoisan-Tuu, Khoisan-Khoe	!Ui and KhoeKhoe	Hunter-Gatherer and Herder	20 (12)
≠Khomani (A) – San, Khoe	South Africa	Khoisan-Tuu, Khoisan-Khoe	Taa and KhoeKhoe	Hunter-Gatherer and Herder	39 (17)
Nama – Khoe	Namibia	Khoisan-Khoe	KhoeKhoe	Herder	20 (7)
Khwe – San	Angola, Namibia	Khoisan-Khoe	Kalahari	Hunter-Gatherer	17 (17)
Coloured (Colesberg)	South Africa	Indo-European	Germanic	Mixed	20 (0)
Coloured (Wellington)	South Africa	Indo-European	Germanic	Mixed	20 (0)
Herero	Namibia	Niger-Kordofanian	Bantoid	Farmer	12 (8)
Bantu-speakers (South Africa)	South Africa	Niger-Kordofanian	Bantoid	Farmer	20 (19)
Ju/'hoansi (HGDP)	Namibia	Khoisan-Ju	Southeast	Hunter-Gatherer	2[2] (2)
KB1 – San (Schuster *et al.* 2010)	Namibia	Khoisan-Tuu	-	Hunter-Gatherer	1 (1)
≠Khomani (B) (Henn *et al.* 2011)	South Africa	Khoisan-Tuu, Khoisan-Khoe	Taa and KhoeKhoe	Hunter-Gatherer and Herder	27[2] (11)
Biaka – Pygmy (HGDP)	Central African Rep.	Niger-Kordofanian	Adamawa-Ubangi	Hunter-Gatherer	22 (22)
Mbuti – Pygmy (HGDP)	Dem. Rep. of Congo	Nilo-Saharan	Central Sudanic	Hunter-Gatherer	13 (13)
Hadza (Henn *et al.* 2011)	Tanzania	Khoisan-Hadza	Hadza	Hunter-Gatherer	17 (7)
Sandawe (Henn *et al.* 2011)	Tanzania	Khoisan-Sandawe	Sandawe	Hunter-Gatherer	28 (24)
Bantu-speakers (Kenya - HGDP)	Kenya	Niger-Kordofanian	Bantoid	Farmer	11 (11)
Luhya (HapMap)	Kenya	Niger-Kordofanian	Bantoid	Farmer	80 (80)
Bantu-speakers (S. Afr. - HGDP)	S. Africa, Namibia	Niger-Kordofanian	Bantoid	Farmer	6[2] (6)
Yoruba (HGDP)	Nigeria	Niger-Kordofanian	Defoid	Farmer	21 (21)
Yoruba (HapMap)	Nigeria	Niger-Kordofanian	Defoid	Farmer	9 (9)
Mandenka (HGDP)	Senegal	Niger-Kordofanian	Mande	Farmer	22 (22)
Maasai (HapMap)	Kenya	Nilo-Saharan	Eastern Sudanic	Herder	57 (57)

[1] Number of individuals remaining after removing recently admixed individuals.

[2] Individuals that were identical or related to individuals in this study were removed.

Table S5: Internal classification of the Khoisan linguistic group (*51*) and population samples typed in the present study. (Only groups represented in this study are listed in "Language Name" column).

Lineages and branches	**Language name (Güldemann, 2008)**	**Group name present study**
Ju-≠Hõa		
– ≠Hõa		
– Ju (= Northern Khoisan)		
—- Northwest	!'O!Xũu, !Xũu	!Xun
—- Southeast	Ju/'hoan	Ju/'hoansi
Khoe-Kwadi		
– Kwadi		
– Khoe (= Central Khoisan)		
—- KhoeKhoe		
—— North	Nama-Damara	Nama
—— South	!Ora, Cape varieties	Coloured Populations from SA[1]
—- Kalahari		
—— East		
———— Shua	-	-
———— Tshwa	-	-
—— West		
———— Kxoe	Khwe	Khwe
———— G//ana	G//ana, G/ui	/Gui and //Gana
———— Naro	-	-
Tuu (= Southern Khoisan)		
– Taa-Lower Nossob		
—- Taa	N/amani	≠Khomani[2]
—— West		
—— East		
—- Lower Nossob	-	-
– !Ui	/Xam	Karretjie People[3]
Hadza	Hadza	Hadza
Sandawe	Sandawe	Sandawe

[1] Khoe-San ancestry from the Coloured populations included in the present study is expected to originate mainly from the !Ora, Cape Khoe and, /Xam.

[2] Most of the ≠Khomani do not speak their original language anymore but are expected to be descendants of the Taa division of Khoisan (see discussion below). They however have ancestry from the Nama (KhoeKhoe branch) as well (recorded during sampling).

[3] The Khoe-San ancestry of the Karretjie People is expected to come mostly from the /Xam (see discussion below and (*92*)), however input from the !Ora (KhoeKhoe branch) is also possible.

Table S6: Number of assignments to the most common or major mode (out of 100 cluster assignments at each K) for ADMIXTURE runs of all datasets. Number of assignments to the most common minor mode are shown in brackets. Only the major mode for each K value and dataset was plotted in DISTRUCT.

Cluster	Southern Africa Admixed Included	Southern Africa Admixed Removed	Global Admixed Included	Global Admixed Removed	African Admixed Removed	African Haplotypes Admixed Removed
K2	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
K3	100 (0)	89 (7)	100 (0)	100 (0)	100 (0)	92 (8)
K4	83 (9)	62 (18)	67 (32)	69 (30)	93 (7)	100 (0)
K5	76 (14)	22 (20)	62 (34)	51 (44)	84 (8)	22 (19)
K6	39 (13)	19 (13)	92 (6)	85 (15)	59 (28)	19 (11)
K7	19 (12)	16 (11)	96 (4)	100 (0)	56 (26)	15 (10)
K8	10 (9)	24 (9)	52 (37)	81 (12)	54 (33)	18 (15)
K9	11 (8)	9 (8)	53 (26)	29 (25)	99 (1)	19 (16)
K10	15 (14)	1 (1)	46 (15)	29 (18)	54 (37)	40 (4)
K11	-	-	37 (20)	20 (15)	55 (13)	12 (6)
K12	-	-	26 (23)	20 (18)	27 (12)	2 (1)
K13	-	-	35 (18)	33 (11)	11 (10)	1 (1)
K14	-	-	26 (19)	27 (24)	7 (6)	-
K15	-	-	29 (14)	39 (8)	6 (4)	-

Table S7: Pairwise F_{ST} values for Southern African populations based on ~2.3M SNPs.

Population	/Gui and //Gana	Ju/'hoansi	Karretjie	≠Khomani (A)	Khwe	Nama	KB1	Bantu-sp. (SA)	Herero
/Gui and //Gana	-	-	-	-	-	-	-	-	-
Ju/'hoansi	0.030669	-	-	-	-	-	-	-	-
Karretjie	0.021889	0.030668	-	-	-	-	-	-	-
≠Khomani (A)	0.016593	0.024944	0.008153	-	-	-	-	-	-
Khwe	0.032460	0.054815	0.038459	0.036431	-	-	-	-	-
Nama	0.023614	0.032110	0.012486	0.005825	0.034896	-	-	-	-
KB1	0.008443	0.030106	0.012018	0.002574	0.014715	0.010682	-	-	-
Bantu-sp. (SA)	0.043182	0.075007	0.047835	0.049021	0.013312	0.047884	0.027286	-	-
Herero	0.067025	0.103509	0.074932	0.074070	0.022786	0.071365	0.054495	0.015385	-
!Xun	0.021442	0.016284	0.023812	0.019739	0.030203	0.023578	0.011733	0.045820	0.067016

Table S8: Sixteen four-population topologies supported by concordance tests featuring both Ju/'hoansi, Mbuti, and chimpanzee as an outgroup. All tests support a common origin of San and Khoe groups (save the Khwe) separate from that of other populations. Significant C- ($Z > 2$) and D-tests ($|Z| > 2$) are shown in bold text.

A,B,C,D	$N_{(A(B(C,D))}$	$N_{(A(C(B,D))}$	$N_{(A(D(B,C))}$	C	Z_C	D	Z_D
chimp,Ju/'hoansi,Mbuti,Biaka	11,676	10,789	10,792	0.039	**6.76**	0.000	-0.02
chimp,Ju/'hoansi,Mbuti,Sandawe	12,085	11,685	10,794	0.017	**2.88**	0.040	**6.42**
chimp,Ju/'hoansi,Mbuti,Hadza	9,974	9,451	8,872	0.027	**3.93**	0.032	**4.31**
chimp,Ju/'hoansi,Mbuti,Maasai	12,519	11,462	10,957	0.044	**7.33**	0.023	**3.82**
chimp,Ju/'hoansi,Mbuti,Mandenka	11,896	11,224	10,763	0.029	**4.97**	0.021	**3.02**
chimp,Ju/'hoansi,Mbuti,Yoruba (HGDP)	12,014	11,058	10,901	0.041	**6.02**	0.007	1.02
chimp,Ju/'hoansi,Mbuti,Luhya	12,312	11,441	10,954	0.037	**6.24**	0.022	**3.99**
chimp,Ju/'hoansi,Mbuti,Bantu-speakers (Kenya - HGDP)	12,005	10,918	10,722	0.047	**7.28**	0.009	1.37
chimp,Ju/'hoansi,Mbuti,Bantu-speakers (South Africa)	11,758	11,234	10,922	0.023	**3.86**	0.014	**2.06**
chimp,Ju/'hoansi,Mbuti,Herero	11,652	10,912	10,561	0.033	**5.21**	0.016	**2.84**
chimp,Ju/'hoansi,Mbuti,Bantu-speakers (S. Africa - HGDP)	11,272	10,795	10,462	0.022	**3.26**	0.016	**2.62**
chimp,Mbuti,Ju/'hoansi,!Xun	11,568	10,771	10,597	0.036	**5.70**	0.008	1.29
chimp,Mbuti,Ju/'hoansi,/Gui and //Gana	11,180	10,561	10,398	0.028	**4.58**	0.008	1.15
chimp,Mbuti,Ju/'hoansi,Nama	11,102	10,674	10,263	0.020	**3.14**	0.020	**2.77**
chimp,Mbuti,Ju/'hoansi,Karretjie	11,333	10,745	10,603	0.027	**3.88**	0.007	1.00
chimp,Mbuti,Ju/'hoansi,≠Khomani (A + B)	11,581	10,849	10,671	0.033	**5.39**	0.008	1.27

Table S9: Estimated population divergence times.

Ingroup	Outgroup	N_{conc}	N_{disc}	Fraction N_{conc}	$\hat{T}$	95% CI
Ju/'hoansi	Mbuti	15,857	25,199	0.39	0.083	0.075-0.091
Ju/'hoansi	Karretjie	13,935	25,505	0.35	0.030	0.023-0.038
Ju/'hoansi	Karretjie+≠Khomani (A)	13,966	25,701	0.35	0.029	0.021-0.035
Ju/'hoansi	≠Khomani (B)	13,507	24,968	0.35	0.027	0.020-0.034
Karretjie	Mbuti	15,778	26,707	0.37	0.059	0.051-0.066
Karretjie	Ju/'hoansi	14,022	26,432	0.35	0.020	0.013-0.027
Karretjie	≠Khomani (A)	13,812	26,890	0.34	0.009	0.002-0.016
Mbuti	Ju/'hoansi	17,014	24,995	0.41	0.114	0.106-0.122
Mbuti	Karretjie	17,262	25,502	0.40	0.111	0.104-0.119
Mbuti	Hadza	16,342	28,018	0.37	0.054	0.047-0.061

Table S10: Evidence for gene flow between sub-Saharan African hunter-gatherers. Significant C- ($Z > 2$) and D-tests ($|Z| > 3$) are shown in bold text. D_f refers to allele frequency based estimate of D rather than estimates based on single gene copies from each population. SNPs refer to the number of SNPs for which the minor allele frequency was at least 10% in each included population.

B,C,D	$N_{(B(C,D))}$	$N_{(C(B,D))}$	$N_{(D(B,C))}$	C	Z_C	D	Z_D	D_f	Z_{Df}	SNPs
Ju/’hoansi Mbuti **Hadza**	9,974	9,451	8,872	0.027	**3.93**	0.032	**4.31**	0.034	**15.69**	83,203
Ju/’hoansi Mbuti **Sandawe**	12,085	11,685	10,794	0.017	**2.88**	0.040	**6.42**	0.024	**15.30**	108,762
Ju/’hoansi Mbuti **Maasai**	12,519	11,462	10,957	0.044	**7.33**	0.023	**3.82**	0.026	**15.47**	109,637
Ju/’hoansi Mbuti **Yoruba**	12,014	11,058	10,901	0.041	**6.02**	0.007	1.02	0.013	**7.85**	104,986
Ju/’hoansi Mbuti Biaka	11,676	10,789	10,792	0.039	**6.76**	0.000	-0.02	-0.0026	-1.70	105,368
Ju/’hoansi Hadza Sandawe	11,386	9,418	9,835	0.073	**10.35**	-0.022	**-3.21**	-0.014	**-8.73**	93,004
Ju/’hoansi Hadza Maasai	11,797	9,711	9,769	0.094	**13.97**	-0.0030	-0.43	-0.011	**-6.21**	93,642
Ju/’hoansi Hadza Yoruba	11,185	9,303	9,862	0.063	**11.70**	-0.029	**-4.62**	-0.024	**-12.35**	88,785
Mbuti Hadza Sandawe	11,842	10,089	10,622	0.054	**9.66**	-0.026	**-3.85**	-0.014	**-7.88**	99,452
Mbuti Hadza Yoruba	11,590	10,114	10,230	0.062	**9.76**	-0.0057	-0.77	-0.012	**-6.77**	95,778

Table S11: Haplotype Heterozygosity at 50 kb.

Population	Haplotype Heterozygosity at 50 kb
Khwe	81.82%
Bantu-speakers (South Africa)	81.56%
Biaka	81.33%
Sandawe	81.26%
Luhya	81.26%
Maasai	81.16%
Bantu-speakers (Kenya - HGDP)	81.13%
!Xun	80.99%
/Gui and //Gana	80.95%
Yoruba (HGDP)	80.95%
Yoruba (HapMap)	80.95%
≠Khomani (A)	80.86%
Herero	80.80%
≠Khomani (B)	80.75%
Mandenka	80.71%
Nama	80.63%
Karretjie	80.59%
Mbuti	80.43%
Ju/'hoansi	80.27%
Mozabite	77.52%
Makrani	77.24%
Sindhi	77.14%
Palestinian	77.07%
Bedouin	76.92%
Balochi	76.90%
Brahui	76.79%
Pathan	76.79%
Gujarati Indian	76.70%
Tuscan (HGDP)	76.61%
Uygur	76.61%
Burusho	76.56%
Hazara	76.45%
Tuscan (HapMap)	76.33%
Adygei	76.24%
Druze	76.20%
North Italian	75.94%
French	75.93%
Russian	75.87%
French Basque	75.78%
Orcadian	75.57%
Sardinian	75.49%
Cambodian	74.94%
Tu	74.93%
Xibo	74.81%
Mongola	74.80%
Tujia	74.76%
Kalash	74.45%
Han	74.41%
Yizu	74.38%
Daur	74.35%
Hezhen	74.26%
Dai	74.24%
Oroqen	74.18%
Naxi	74.17%
Japanese	74.14%
Miaozu	74.13%
Yakut	74.09%
She	73.81%
Hadza	73.48%
Lahu	73.27%
Maya	71.45%
Melanesian	70.72%
Papuan	70.50%
Colombian	68.32%
Pima	67.71%
Karitiana	64.08%
Surui	62.28%

Table S12: Haplotype Richness at 50 kb.

Population	Haplotype Richness at 50kb
Khwe	7.75
Bantu-speakers (South Africa)	7.68
Biaka	7.63
≠Khomani (A)	7.59
/Gui and //Gana	7.59
Luhya	7.57
Bantu-speakers (Kenya - HGDP)	7.56
!Xun	7.56
≠Khomani (B)	7.55
Sandawe	7.54
Maasai	7.54
Karretjie	7.52
Yoruba (HGDP)	7.50
Mbuti	7.47
Yoruba (HapMap)	7.47
Mandenka	7.44
Herero	7.44
Nama	7.44
Ju/'hoansi	7.36
Mozabite	6.63
Makrani	6.61
Sindhi	6.58
Palestinian	6.58
Bedouin	6.54
Balochi	6.52
Pathan	6.50
Uygur	6.49
Tuscan (HGDP)	6.49
Brahui	6.48
GIH	6.48
Hazara	6.43
Burusho	6.43
Tuscan (HapMap)	6.42
Adygei	6.38
Druze	6.37
French	6.34
NorthItalian	6.33
Russian	6.30
French Basque	6.29
Sardinian	6.23
Cambodian	6.22
Orcadian	6.22
Tu	6.21
Mongola	6.19
Xibo	6.18
Tujia	6.18
Han	6.12
Yizu	6.08
Daur	6.08
Dai	6.06
Japanese	6.05
Hezhen	6.04
Oroqen	6.04
Miaozu	6.04
Naxi	6.03
Yakut	5.97
She	5.95
Kalash	5.89
Lahu	5.81
Hadza	5.67
Maya	5.56
Papuan	5.50
Melanesian	5.48
Colombian	4.93
Pima	4.77
Karitiana	4.19
Surui	3.95

Table S13: *cRoH* of the 0.5Mb-1Mb class averaged over 50 repeats of sub-sampling 7 individuals per population.

Group	*cRoH* (0.5-1Mb class)
Mandenka	4.748
Yoruba (HapMap)	5.392
Maasai	5.870
Yoruba (HGDP)	6.932
Luhya	7.502
Bantu-speakers (Kenya - HGDP)	7.614
Karretjie	7.749
Bantu-speakers (Southern Africa - HGDP)	9.955
≠Khomani (A)	10.011
Bantu-speakers (South Africa)	10.295
Khwe	11.176
Sandawe	11.649
≠Khomani (B)	12.874
Mbuti	14.718
Biaka	15.026
Nama	15.056
/Gui and //Gana	15.777
!Xun	16.059
Ju'/hoansi	17.108
Herero	17.858
Hadza	59.226

Table S14: Top SNPs in the Khoe-San vs. Bantu-speakers scan for differentiation.

Chr	Position	SNP	F_{ST}	Gene in region
1	231928935	rs1322784	0.650320	DISC1
11	42414566	rs7127426	0.623931	none
12	76079002	rs10879959	0.622062	none
13	113110994	kgp6935620	0.616207	none
11	128382804	rs4372467	0.615883	ETS1
11	42413902	kgp12503088	0.613696	none
6	125928408	kgp4333469	0.611340	none
2	37909108	rs11683907	0.607130	none
1	231942379	kgp6919041	0.603243	DISC1
7	17725680	rs11765014	0.602502	none
11	128390012	kgp31391	0.599873	ETS1
11	128379964	kgp12532905	0.599873	ETS1
12	79523492	rs7967341	0.594879	none
3	10973088	kgp2203756	0.594185	SLC6A11
8	78846655	kgp12386959	0.593370	none
6	125954417	kgp5383057	0.592617	none
2	164370435	rs10193051	0.592617	none
8	21761443	rs11135701	0.592411	none
11	42777692	kgp8365438	0.588026	none
1	85563456	kgp1601239	0.585833	WDR63

Table S15: Top SNPs in the San vs Nama scan for differentiation.

Chr	Position	SNP	F_{ST}	Gene in region
16	13809776	kgp10343152	0.896853	none
16	13813704	kgp10516300	0.882824	none
16	13813289	rs7188713	0.882824	none
16	13658506	rs2001025	0.829805	none
16	13664726	kgp9620268	0.799884	none
11	11409067	rs11021826	0.785613	CSNK2A1P
16	13813090	kgp7916144	0.774886	none
14	78027686	kgp7290943	0.774886	SPTLC2
9	8865925	rs10977327	0.750659	PTPRD
16	13670907	kgp4664202	0.744456	none
16	13709785	kgp1477861	0.742857	none
16	13703693	rs7184684	0.742857	none
16	13666011	rs4632116	0.721744	none
6	9806230	rs17613393	0.718730	none
6	9804986	rs9476916	0.718730	none
3	70396582	kgp177491	0.718730	none
18	5186508	kgp11852213	0.718730	C18orf42
17	77007553	kgp3851511	0.718730	none
17	73035793	rs4435292	0.718730	ATP5H
11	117566167	kgp2427704	0.718730	DSCAML1

Table S16: The 7 regions that were among the top 10 according to empirical p-value, following (*43*), that also harbored top 10 $|iHS|$ values.

population	chromosome	start (Mb)	end (Mb)	position of max $\lvert iHS \rvert$
Ju/′hoansi	10	69.8	70	69.90
Karretjie	1	210.4	211	210.70
Karretjie	6	22.4	22.6	22.49
≠Khomani (A)	6	26.8	28.0	27.28
Khwe	1	219.6	222.2	220.59
Khwe	4	123.4	123.8	123.56
Bantu-speakers (South Africa)	10	27.8	28.8	28.15

Table S17: The genes on chromosome 10 from position 69.8-70 Mb corresponding to the only candidate windows in Ju/'hoansi. A single star (*) denotes non-coding RNA. See also Figure S73.

number in Figure	gene	strand	start	end	distance to position of max $\|iHS\|$
1	SIRT1	+	69644426	69678147	231655
2	HERC4	-	69681655	69835103	74699
3	MYPN	+	69865875	69971773	0
4	7SK*	+	69917835	69918165	-8033
5	ATOH7	-	69990351	69991870	-80549
6	PBLD	-	70042416	70092684	-132614
7	HNRNPH3*	+	70091767	70102953	-181965
8	RUFY2	-	70100863	70167051	-191061
9	DNA2	-	70173820	70196998	-264018

Table S18: The genes on chromosome 1 from position 210.4-211 Mb corresponding to one of two candidate windows in Karretjie. A single star (*) denotes non-coding RNA. See also Figure S74.

number in Figure	gene	strand	start	end	distance to position of max $\|iHS\|$
1	SYT14	+	210267666	210337633	363427
2	C1orf133*	-	210404803	210407466	293594
3	SERTAD4	+	210406194	210416440	284620
4	HHAT	+	210501595	210849638	0
5	KCNH1	-	210851656	211192598	-150596

Table S19: The genes on chromosome 6 from position 22.4-22.6Mb corresponding to a second candidate window in Karretjie. A single star (*) denotes non-coding RNA. See also Figure S75.

number in Figure	gene	strand	start	end	distance to position of max $\|iHS\|$
1	LINC00340*	+	22205413	22214734	276536
2	PRL	-	22287472	22303082	188188
3	HDGFL1	+	22569677	22570750	-78407

Table S20: The genes on chromosome 6 from position 26.8Mb-28.0Mb corresponding to the only candidate window in ≠Khomani (A). See also Figure 4 in the main text, which does not include any of the numerous tRNA (n=83) and histone (n=21) genes in this region. The distributions of the latter are included in Figure S77. A single star (*) denotes non-coding RNA while two stars(**) denote pseudo genes.

number in Figure	gene	strand	start	end	distance to position of max $\|iHS\|$
1	ABT1	+	26598492	26600277	675073
2	DQ786258	-	26634610	26637135	638215
3	ZNF322	-	26634610	26659980	615370
4	GUSBP2**	-	26839265	26924333	351017
5	DQ596042	+	26864472	26864826	410524
6	DQ587763	-	26864775	26864804	410546
7	DQ571494*	-	26866009	26867685	407665
8	LINC00240	+	26924771	26991753	283597
9	LOC100270746*	-	26987144	26988085	287265
10	BC014312	-	27092761	27100572	174778
11	MIR3143*	+	27115404	27115467	159883
12	PRSS16	+	27215501	27224399	50951
13	POM121L2	-	27276841	27280011	-1491
14	VN1R10P**	+	27292539	27293741	-17189
15	ZNF204P**	-	27325601	27343153	-50251
16	ZNF391	+	27356523	27369227	-81173
17	BC132797	+	27356537	27357613	-81187
18	AX747641	+	27357717	27358702	-82367
19	ZNF184	-	27418520	27440897	-143170
20	LOC100507173*	+	27661813	27678001	-386463
21	LOC100131289*	+	27729522	27730966	-454172
22	FKSG63	+	27793689	27794715	-518339
23	BC016143	-	27794098	27794504	-518748
24	OR2B2	-	27878962	27880174	-603612
25	OR2B6	+	27925018	27925960	-649668
26	ZNF165	+	28048481	28057340	-773131
27	AK309286	-	28058453	28105071	-783103
28	ZSCAN12P1**	+	28058928	28063493	-783578
29	AK127889	-	28089572	28105071	-814222
30	ZSCAN16	+	28092386	28097856	-817036
31	ZNF192	+	28109715	28125236	-834365
32	AK311106	+	28129569	28131289	-854219
33	DQ581281	+	28134728	28134756	-859378
34	TOB2P1**	-	28183115	28186707	-907765
35	ZNF193	+	28193028	28198266	-917678

Table S21: The genes on chromosome 1 from position 219.6-222.2 Mb corresponding to one of two candidate windows in Khwe. A single star (*) denotes non-coding RNA while two stars (**) denote pseudo genes. See also Figure S78.

number in Figure	gene	strand	start	end	distance to position of max $\|iHS\|$
1	SLC30A10*	-	219858768	220131989	0
2	RNU5F-1	+	220046618	220292777	0
3	EPRS	-	220141941	220220000	-63753
4	BPNT1	-	220230823	220263191	-152635
5	IARS2	+	220267454	220321383	-189266
6	MIR215*	-	220291194	220291304	-213006
7	MIR194-1*	-	220291498	220291583	-213310
8	DM119532	+	220291547	220291569	-213359
9	RAB3GAP2	-	220321609	220445843	-243421
10	SNORA36B*	-	220373883	220374018	-295695
11	AURKAPS1**	-	220439520	220441057	-361332
12	MARK1	+	220701567	220837799	-623379
13	C1orf115	+	220863627	220872499	-785439
14	MARC2	+	220921675	220957596	-843487
15	MARC1	+	220960038	220987741	-881850
16	U6atac	+	220999117	220999235	-920929
17	BC045735	+	221002596	221005770	-924408
18	HLX	+	221052742	221058400	-974554
19	C1orf140*	-	221503269	221509638	-1425081
20	DUSP10	-	221874763	221915516	-1796575

Table S22: The genes on chromosome 4 from position 123.4-123.8 Mb corresponding to a second candidate window in Khwe. Two stars (**) denotes denote pseudo genes. See also Figure S79.

number in Figure	gene	strand	start	end	distance to position of max $\|iHS\|$
1	KIAA1109	+	123200929	123283914	276912
2	ADAD1	+	123300120	123350947	209879
3	IL2	-	123372625	123377650	183176
4	IL21	-	123533782	123542211	18615
5	BC045668	+	123540137	123610315	0
6	CETN4P**	-	123651343	123653613	-90517
7	BBS12	+	123653856	123666098	-93030
8	FGF2	+	123747862	123819390	-187036
9	NUDT6	-	123813798	123844159	-252972
10	SPATA5	+	123844224	123978443	-283398

Table S23: The genes on chromosome 10 from position 27.8-28.8 Mb corresponding to the only candidate window in Bantu-speakers (South Africa). A single star (*) denotes non-coding RNA. See also Figure S80.

number in Figure	gene	strand	start	end	distance to position of max $\lvert iHS \rvert$
1	PTCHD3	-	27687116	27703297	259417
2	RAB18	+	27793248	27829099	133615
3	MKX	-	27961802	28034778	0
4	ARMC4	-	28101096	28287977	-138382
5	MPP7	-	28339922	28591995	-377208
6	LOC220906*	-	28808845	28821460	-846131
7	WAC	+	28821421	28912041	-858707
8	BAMBI	+	28966423	28971868	-1003709

Pop	Haplotype Count	Total N	Percentage
Maasai	46	114	40.4
Nama	5	14	35.7
≠Khomani (B)	6	22	27.3
≠Khomani (A)	5	34	14.7
Sandawe	4	48	8.3
Karretjie	2	24	8.3
Batu-speakers (South Africa)	3	38	7.9
Luhya	12	160	7.5
/Gui and //Gana	1	14	7.1
Bantu-speakers (Kenya - HGDP)	1	22	4.5
Khwe	1	34	2.9
Ju/'hoansi	1	34	2.9

Table S24: Frequencies of putative East African lactase persistence haplotype in African populations (African populations not shown have frequency of 0).

Table S25: Top 20 identified candidate regions for selection in the ancestors of modern humans. Genes discussed in the main text are bolded.

Rank	Location (hg19)	Span (bp)	SNPs	Peak aPBS	Genes
1	chr9:94682448-94939615	257,167	52	2.92	LINC00475,LOC100128076,**ROR2**,**SPTLC1**
2	chr20:46565023-46626596	61,573	31	2.60	BX648826
3	chr1:243470453-243608642	138,189	30	1.78	MIR4677,**SDCCAG8**
4	chr6:44802718-45701338	898,620	184	1.62	**RUNX2**,MIR586,SUPT3H
5	chr4:155436137-155811101	374,964	134	1.54	RBM46,PLRG1,DQ266889,**LRAT**,FGG,FGA,FGB
6	chr11:128686102-128770473	84,371	50	1.52	KCNJ1,C11orf45,KCNJ5
7	chr16:62644640-62873194	228,554	67	1.50	-
8	chr1:92205244-92312300	107,056	52	1.50	TGFBR3,Metazoa_SRP
9	chr2:207659531-207798923	139,392	48	1.40	FASTKD2
10	chr7:104368522-104514314	145,792	55	1.38	LOC645591,LHFPL3
11	chr5:30747852-30880852	133,000	61	1.37	-
12	chr17:66801331-67096691	295,360	102	1.36	ABCA6,MIR4524A,ABCA8,ABCA9
13	chr12:32097486-32156444	58,958	49	1.36	C12orf35
14	chr6:37452608-37556504	103,896	76	1.36	CCDC167,MIR4462
15	chr16:83425303-83493961	68,658	65	1.33	-
16	chr11:309127-867621	558,494	131	1.32	[1]
17	chr1:38208000-38442352	234,352	50	1.30	[2]
18	chr3:74395632-74563488	167,856	72	1.29	CNTN3
19	chr9:101332364-101460788	128,424	56	1.28	GABBR2
20	chr14:85606755-85887140	280,385	142	1.26	7SK,Mir_548

[1] B4GALNT4, PTDSS2, LOC143666, AX748330, ANO9, TALDO1, TMEM80, HRAS, PIDD, RASSF7, CD151, PKP3, RNH1, CEND1, BC031953, C11orf35, BC048998, DEAF1, PNPLA2, RPLP2, SCT, IFITM3, IFITM2, IFITM1, PHRF1, BC040735, SLC25A22LRRC56, TSPAN4, POLR2L, EPS8L2, SNORA52, AK126635, MIR210HG, IRF7, DRD4, MIR210, CDHR5, JA429539, Metazoa_SRP, SIGIRR, CHID1, EFCAB4A, PDDC1

[2] INPP5B, C1orf122, MTF1, YRDC, SF3A3, EPHA10, MANEAL, SNORA63

References and Notes

1. C. Stringer, Modern human origins: progress and prospects. *Philos. Trans. R. Soc. London Ser. B* **357**, 563 (2002). doi:10.1098/rstb.2001.1057 Medline

2. L. Barham, P. Mitchell, *The First Africans* (Cambridge Univ. Press, Cambridge, 2008).

3. M. Jakobsson *et al*., Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998 (2008). doi:10.1038/nature06742 Medline

4. J. Z. Li *et al*., Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100 (2008). doi:10.1126/science.1153717 Medline

5. D. M. Behar *et al*.; Genographic Consortium, The dawn of human matrilineal diversity. *Am. J. Hum. Genet.* **82**, 1130 (2008). doi:10.1016/j.ajhg.2008.04.002 Medline

6. J. Lachance *et al*., Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse african hunter-gatherers. *Cell* **150**, 457 (2012). doi:10.1016/j.cell.2012.07.009 Medline

7. R. E. Green *et al*., A draft sequence of the Neandertal genome. *Science* **328**, 710 (2010). doi:10.1126/science.1188021 Medline

8. S. A. Tishkoff *et al*., The genetic structure and history of Africans and African Americans. *Science* **324**, 1035 (2009). doi:10.1126/science.1172257 Medline

9. S. C. Schuster *et al*., Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943 (2010). doi:10.1038/nature08795 Medline

10. B. M. Henn *et al*., Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5154 (2011). doi:10.1073/pnas.1017511108 Medline

11. I. Gronau, M. J. Hubisz, B. Gulko, C. G. Danko, A. Siepel, Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031 (2011). doi:10.1038/ng.937 Medline

12. K. R. Veeramah *et al*., An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Mol. Biol. Evol.* **29**, 617 (2012). doi:10.1093/molbev/msr212 Medline

13. S. A. Tishkoff *et al*., History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol. Biol. Evol.* **24**, 2180 (2007). doi:10.1093/molbev/msm155 Medline

14. See supplementary materials on *Science* Online.

15. 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010). doi:10.1038/nature09534 Medline

16. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655 (2009). doi:10.1101/gr.094052.109 Medline

17. P. Skoglund *et al*., Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336**, 466 (2012). doi:10.1126/science.1216304 Medline

18. G. T. Nurse, J. S. Weiner, T. Jenkins, *The Peoples of Southern Africa and their Affinities* (Oxford Univ. Press, New York, 1985).

19. S. Ramachandran *et al*., 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.* **102**, 15942 (2005). doi:10.1073/pnas.0507611102 Medline

20. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006). doi:10.1371/journal.pbio.0040072 Medline

21. M. L. Bang *et al*., Myopalladin, a novel 145-kilodalton sarcomeric protein with multiple roles in Z-disc and I-band protein assemblies. *J. Cell Biol.* **153**, 413 (2001). doi:10.1083/jcb.153.2.413 Medline

22. N. Yang *et al*., ACTN3 genotype is associated with human elite athletic performance. *Am. J. Hum. Genet.* **73**, 627 (2003). doi:10.1086/377590 Medline

23. Y. Matsumura, C. Nishigori, T. Yagi, S. Imamura, H. Takebe, Characterization of molecular defects in xeroderma pigmentosum group F in relation to its clinically mild symptoms. *Hum. Mol. Genet.* **7**, 969 (1998). doi:10.1093/hmg/7.6.969 Medline

24. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012). doi:10.1371/journal.pgen.1002453 Medline

25. I. Kalus *et al*., Differential involvement of the extracellular 6-O-endosulfatases Sulf1 and Sulf2 in brain development and neuronal and behavioural plasticity. *J. Cell. Mol. Med.* **13**, 4505 (2009). doi:10.1111/j.1582-4934.2008.00558.x Medline

26. S. Mundlos *et al*., Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. *Cell* **89**, 773 (1997). doi:10.1016/S0092-8674(00)80260-3 Medline

27. D. Falk, C. P. E. Zollikofer, N. Morimoto, M. S. Ponce de León, Metopic suture of Taung (Australopithecus africanus) and its implications for hominin brain evolution. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8467 (2012). doi:10.1073/pnas.1119752109 Medline

28. A. D. Hill *et al*., A 2-Mb critical region implicated in the microcephaly associated with terminal 1q deletion syndrome. *Am. J. Med. Genet. A.* **143A**, 1692 (2007). doi:10.1002/ajmg.a.31776 Medline

29. R. Abraham *et al*., A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC Med. Genomics* **1**, 44 (2008). doi:10.1186/1755-8794-1-44 Medline

30. M. G. B. Blum, M. Jakobsson, Deep divergences of human gene trees and models of human origins. *Mol. Biol. Evol.* **28**, 889 (2011). doi:10.1093/molbev/msq265 Medline

31. C. M. Schlebusch, H. Soodyall, M. Jakobsson, Genetic variation of 15 autosomal STR loci in various populations from southern Africa. *Forensic Sci. Int. Genet.* **6**, e20 (2012). doi:10.1016/j.fsigen.2010.12.013 Medline

32. S. Purcell *et al*., PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559 (2007). doi:10.1086/519795 Medline

33. S. A. Miller, D. D. Dykes, H. F. Polesky, A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* **16**, 1215 (1988). doi:10.1093/nar/16.3.1215 Medline

34. R. Pinard *et al*., Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**, 216 (2006). doi:10.1186/1471-2164-7-216 Medline

35. P. Scheet, M. Stephens, A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629 (2006). doi:10.1086/502802 Medline

36. Y. Li, C. J. Willer, S. Sanna, G. R. Abecasis, Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387 (2009). doi:10.1146/annurev.genom.9.081307.164242 Medline

37. 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010). doi:10.1038/nature09534 Medline

38. B. Paten, J. Herrero, K. Beal, S. Fitzgerald, E. Birney, Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18**, 1814 (2008). doi:10.1101/gr.076554.108 Medline

39. B. Paten *et al*., Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829 (2008). doi:10.1101/gr.076521.108 Medline

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

41. T. J. Pemberton, C. Wang, J. Z. Li, N. A. Rosenberg, Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *Am. J. Hum. Genet.* **87**, 457 (2010). doi:10.1016/j.ajhg.2010.08.014 Medline

42. H. M. Cann *et al*., A human genome diversity cell line panel. *Science* **296**, 261 (2002). doi:10.1126/science.296.5566.261b Medline

43. J. K. Pickrell *et al*., Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826 (2009). doi:10.1101/gr.087577.108 Medline

44. G. Coop *et al*., The role of geography in human adaptation. *PLoS Genet.* **5**, e1000500 (2009). doi:10.1371/journal.pgen.1000500 Medline

45. M. Jakobsson, N. A. Rosenberg, CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801 (2007). doi:10.1093/bioinformatics/btm233 Medline

46. C. M. Schlebusch, Phd thesis: Genetic variation in khoisan-speaking populations from southern africa, Ph.D. thesis, University of the Witwatersrand (2010).

47. C. Schlebusch, Issues raised by use of ethnic-group names in genome study. *Nature* **464**, 487, author reply 487 (2010). doi:10.1038/464487a Medline

48. A. Barnard, *Hunters and Herders of Southern Africa - A Ccomparative Ethnography of the Khoisan Peoples* (Cambridge Univ. Press, Cambridge, 1992).

49. A. Smith, C. Malherbe, M. Guenther, P. Berens, *The Bushmen of Southern Africa* (David Philips Publishers, Cape Town, 2000).

50. W. le Roux, A. White, Eds., *Voices of the San* (Kwela Books, Cape Town, 2004).

51. T. Güldemann, *Southern African Humanities* **20**, 93 (2008).

52. C. M. Schlebusch, M. de Jongh, H. Soodyall, Different contributions of ancient mitochondrial and Y-chromosomal lineages in 'Karretjie people' of the Great Karoo in South Africa. *J. Hum. Genet.* **56**, 623 (2011). doi:10.1038/jhg.2011.71 Medline

53. J. Marshall, C. Ritchie, *Where are the Ju/wasi of Nyae Nyae? Changes in a Bushman society: 1958-1981* (Centre for African Studies, University of Cape Town (Communications No.9), Cape Town, 1984).

54. R. Gordon, *The past and future of! Kung ethnography: critical reflections and symbolic perspectives, essays in honour of Lorna Marshall*, M. Biesele, R. Gordon, R. Lee, eds. (Helmut Buske Verlag, Hamburg, 1986), pp. 53-68.

55. R. Gordon, *Past and Present in Hunter-Gatherer Studies*, C. Schrire, Ed. (Academic Press, Orlando, FL, 1984), pp. 195-224.

56. A. Barnard, Kinship, Language and Production: A Conjectural History of Khoisan Social Structure. *Africa* **58**, 29 (1988). doi:10.2307/1159869

57. D. F. Bleek, Bushmen of Central Angola. *Bantu Studies* **3**, 105 (1927). doi:10.1080/02561751.1927.9676200

58. A. De Almeida, *Bushmen and Other Non-Bantu Peoples of Angola* (Witwatersrand Univ. Press for the Institute for the Study of Man in Africa, Johannesburg, 1965).

59. L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, A. C. Wilson, African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503 (1991). doi:10.1126/science.1840702 Medline

60. Y. S. Chen *et al*., mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. *Am. J. Hum. Genet.* **66**, 1362 (2000). doi:10.1086/302848 Medline

61. M. K. Gonder, H. M. Mortensen, F. A. Reed, A. de Sousa, S. A. Tishkoff, Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol. Biol. Evol.* **24**, 757 (2007). doi:10.1093/molbev/msl209 Medline

62. R. Scozzari *et al*., Differential structuring of human populations for homologous X and Y microsatellite loci. *Am. J. Hum. Genet.* **61**, 719 (1997). doi:10.1086/515500 Medline

63. R. Scozzari *et al*., Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am. J. Hum. Genet.* **65**, 829 (1999). doi:10.1086/302538 Medline

64. M. F. Hammer *et al*., The geographic distribution of human Y chromosome variation. *Genetics* **145**, 787 (1997). Medline

65. F. Cruciani *et al*., A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am. J. Hum. Genet.* **70**, 1197 (2002). doi:10.1086/340257 Medline

66. F. Cruciani *et al.*, Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. *Am. J. Hum. Genet.* **74**, 1014 (2004). doi:10.1086/386294 Medline

67. A. Knight *et al.*, African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr. Biol.* **13**, 464 (2003). doi:10.1016/S0960-9822(03)00130-1 Medline

68. B. M. Henn *et al.*, Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10693 (2008). doi:10.1073/pnas.0801184105 Medline

69. M. F. Hammer *et al.*, Hierarchical patterns of global human Y-chromosome diversity. *Mol. Biol. Evol.* **18**, 1189 (2001). doi:10.1093/oxfordjournals.molbev.a003906 Medline

70. L. Marshall, !Kung Bushman Bands. *Africa* **30**, 325 (1960). doi:10.2307/1157596

71. R. Lee, (Cambridge Univ. Press, Cambridge, 1979).

72. M. G. Guenther, *Contemporary Studies on Khoisan*, R. Vossen, K. Keuthmann, Eds. (Helmut Buske Verlag, Hamburg, 1986), vol. 1 of *Quellen zur Khoisan-Forschung*, pp. 347–373.

73. P. A. Underhill *et al.*, Y chromosome sequence variation and the history of human populations. *Nat. Genet.* **26**, 358 (2000). doi:10.1038/81685 Medline

74. P. A. Underhill *et al.*, The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65**, 43 (2001). doi:10.1046/j.1469-1809.2001.6510043.x Medline

75. O. Semino, A. S. Santachiara-Benerecetti, F. Falaschi, L. L. Cavalli-Sforza, P. A. Underhill, Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am. J. Hum. Genet.* **70**, 265 (2002). doi:10.1086/338306 Medline

76. H. M. Cann *et al.*, A human genome diversity cell line panel. *Science* **296**, 261 (2002). doi:10.1126/science.296.5566.261b Medline

77. G. T. Nurse, T. Jenkins, Health and the Hunter-Gatherer. Biomedical studies on the hunting and gathering populations of Southern Africa. *Monogr. Hum. Genet.* **8**, 1 (1977). Medline

78. E. Cashdan, *Contemporary Studies on Khoisan*, R. Vossen, K. Keuthmann, Eds. (Helmut Buske Verlag, Hamburg, 1986), vol. 1 of *Quellen zur Khoisan-Forschung 5.1*, pp. 145-180.

79. J. Sharp, S. Douglas, *Miscast. Negotiating the Presence of the Bushmen*, P. Skotnes, Ed. (UCT Press, Cape Town, 1996), pp. 323-329.

80. G. B. Silberbauer, *Report to the Government of Bechuanaland on the Bushman Survey* (Bechuanaland Government, Gabarone, 1965).

81. K. Broyhill, R. Hitchcock, M. Biesele, Current situations facing the san peoples of southern africa, *Tech. rep.*, Review on Current San Economic and Social Situations for the University of Free State. http://www.kalaharipeoples.org/downloads/Current (2007).

82. I. Schapera, *The Khoisan Peoples of South Africa: Bushmen and Hottentots* (George Routledge and Sons, London, 1930).

83. R. Elphick, *Khoikhoi and the Founding of White South Africa*, New history of southern Africa series (Raven Press, Johannesburg, 1985).

84. A. B. Smith, *Einiqualand: Studies of the Orange River Frontier* (Univ. of Cape Town Press, Cape Town, 1995).

85. E. O. J. Westphal, Africa. *Journal of the International African Institute* **33**, 237 (1963). doi:10.2307/1157418

86. A. W. Hoernle, ed., *The social organization of the Nama and other essays* (Witwatersrand Univ. Press, Johannesburg, 1985).

87. E. T. Wood *et al*., Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur. J. Hum. Genet.* **13**, 867 (2005). doi:10.1038/sj.ejhg.5201408 Medline

88. A. Traill, *Miscast. Negotiating the Presence of the Bushmen*, P. Skotnes, ed. (UCT Press, Cape Town, 1996), pp. 171-183.

89. J. Deacon, *Miscast. Negotiating the Presence of the Bushmen*, P. Skotnes, ed. (UCT Press, Cape Town, 1996), pp. 93-113.

90. N. Penn, *Miscast. Negotiating the Presence of the Bushmen*, P. Skotnes, ed. (UCT Press, Cape Town, 1996), pp. 81-91.

91. N. Bennun, *The Broken String - The last words of an extinct people* (Penguin Books, London, 2004).

92. M. de Jongh, No Fixed Abode: The Poorest of the Poor and Elusive Identities in Rural South Africa. *J. South. Afr. Stud.* **28**, 441 (2002). doi:10.1080/03057070220140793

93. H. P. Steyn, Southern Kalahari San Subsistence Ecology: A Reconstruction. *The South African Archaeological Bulletin* **39**, 117 (1984). doi:10.2307/3888377

94. T. Güldemann, *Anthropological Linguistics* **48**, 369 (2006).

95. N. Crawhall, *Maintaining the Links: Language Identity and the Land: Proceedings of the Seventh Foundation for Endangered Languages Conference*, J. Blythe, R. McKenna Brown, eds. (Bristol: Foundation for Endangered Languages, Broome, Western Australia, 2003), pp. 13-19.

96. B. E. Sands, A. L. Miller, J. Brugman, *Selected Proceedings of the 37th Annual Conference on African Linguistics*, D. L. Payne, J. Peña, eds. (Cascadilla Proceedings Project, Somerville, MA, 2007), pp. 55-65.

97. N. Patterson *et al*., Genetic structure of a unique admixed population: implications for medical research. *Hum. Mol. Genet.* **19**, 411 (2010). doi:10.1093/hmg/ddp505 Medline

98. L. Quintana-Murci *et al*., Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. *Am. J. Hum. Genet.* **86**, 611 (2010). doi:10.1016/j.ajhg.2010.02.014 Medline

99. E. de Wit *et al*., Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Hum. Genet.* **128**, 145 (2010). doi:10.1007/s00439-010-0836-1 Medline

100. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006). doi:10.1371/journal.pgen.0020190 Medline

101. C. Wang *et al.*, *Stat. Appl. Genet. Mol. Biol.* **9**, e13 (2010).

102. D. Reich *et al.*, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053 (2010). doi:10.1038/nature09710 Medline

103. N. A. Rosenberg, distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137 (2004). doi:10.1046/j.1471-8286.2003.00566.x

104. B. M. Henn *et al.*, Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10693 (2008). doi:10.1073/pnas.0801184105 Medline

105. N. Huffman, *The Prehistory of Africa*, Soodyall, H, Ed. (Jonathan Ball Publishers, Johannesburg & Cape Town, 2006), pp. 97-108.

106. B. W. Smith, *The Prehistory of Africa*, H. Soodyall, Ed. (Jonathan Ball Publishers, Johannesburg & Cape Town, 2006), pp. 76-96.

107. B. S. Weir, *Genetic Data Analysis II* (Sinauer, Sunderland, MA, 1996).

108. F. Jay, O. François, M. G. Blum, Predictions of native American population structure using linguistic covariates in a hidden regression framework. *PLoS ONE* **6**, e16227 (2011). doi:10.1371/journal.pone.0016227 Medline

109. N. Mantel, The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209 (1967). Medline

110. P. E. Smouse, J. C. Long, R. R. Sokal, *Syst. Biol.* **35**, 627 (1986).

111. E. M. Belle, G. Barbujani, Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. *Am. J. Phys. Anthropol.* **133**, 1137 (2007). doi:10.1002/ajpa.20622 Medline

112. J. Wakeley, *Coalescent Theory* (Roberts & Company, Greenwood Village, CO, 2008).

113. P. Skoglund, A. Götherström, M. Jakobsson, Estimation of population divergence times from non-overlapping genomic sequences: examples from dogs and wolves. *Mol. Biol. Evol.* **28**, 1505 (2011). doi:10.1093/molbev/msq342 Medline

114. R. R. Sokal, F. J. Rohlf, *Biometry* (Freeman, New York, 1995), third edn.

115. C. J. Creevey, J. O. McInerney, Trees from trees: construction of phylogenetic supertrees using clann. *Methods Mol. Biol.* **537**, 139 (2009). doi:10.1007/978-1-59745-251-9_7 Medline

116. J. Pickrell, J. Pritchard, *Nature Precedings* (2012). http://hdl.handle.net/10101/npre.2012.6956.1.

117. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337 (2002). doi:10.1093/bioinformatics/18.2.337 Medline

118. D. M. Altshuler *et al.*; International HapMap 3 Consortium, Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010). doi:10.1038/nature09298 Medline

119. A. Scally *et al*., Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169 (2012). doi:10.1038/nature10842 Medline

120. M. F. Hammer, A. E. Woerner, F. L. Mendez, J. C. Watkins, J. D. Wall, Genetic evidence for archaic admixture in Africa. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15123 (2011). doi:10.1073/pnas.1109300108 Medline

121. Y. Moodley *et al*., Age of the association between Helicobacter pylori and man. *PLoS Pathog.* **8**, e1002693 (2012). doi:10.1371/journal.ppat.1002693 Medline

122. F. d'Errico *et al*., *Proceedings of the National Academy of Sciences USA* (2012).

123. D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing Indian population history. *Nature* **461**, 489 (2009). doi:10.1038/nature08365 Medline

124. C. Batini *et al*., Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. *Mol. Biol. Evol.* **28**, 2603 (2011). doi:10.1093/molbev/msr089 Medline

125. D. F. Conrad *et al*., A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 1251 (2006). doi:10.1038/ng1911 Medline

126. A. Auton *et al*., Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.* **19**, 795 (2009). doi:10.1101/gr.088898.108 Medline

127. M. A. Abdulla *et al*.; HUGO Pan-Asian SNP Consortium; Indian Genome Variation Consortium, Mapping human genetic diversity in Asia. *Science* **326**, 1541 (2009). doi:10.1126/science.1177074 Medline

128. L. Huang *et al*., Haplotype variation and genotype imputation in African populations. *Genet. Epidemiol.* **35**, 766 (2011). doi:10.1002/gepi.20626 Medline

129. M. R. Nelson *et al*., An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100 (2012). doi:10.1126/science.1217876 Medline

130. S. T. Kalinowski, hp-rare 1.0: a computer program for performing rarefaction on measures of allelic richness. *Mol. Ecol. Notes* **5**, 187 (2005). doi:10.1111/j.1471-8286.2004.00845.x

131. B. J. Hayes, P. M. Visscher, H. C. McPartlan, M. E. Goddard, Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* **13**, 635 (2003). doi:10.1101/gr.387103 Medline

132. A. Tenesa *et al*., Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* **17**, 520 (2007). doi:10.1101/gr.6023607 Medline

133. M. Kirin *et al*., Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* **5**, e13996 (2010). doi:10.1371/journal.pone.0013996 Medline

134. S. Purcell, Plink v1.07. http://pngu.mgh.harvard.edu/purcell/plink/ (2010).

135. T. D. White *et al*., Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* **423**, 742 (2003). doi:10.1038/nature01669 Medline

136. I. McDougall, F. H. Brown, J. G. Fleagle, Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433**, 733 (2005). doi:10.1038/nature03258 Medline

137. J. D. Wall *et al.*, A novel DNA sequence database for analyzing human demographic history. *Genome Res.* **18**, 1354 (2008). doi:10.1101/gr.075630.107 Medline

138. M. F. Hammer *et al.*, The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nat. Genet.* **42**, 830 (2010). doi:10.1038/ng.651 Medline

139. K. Yamamura *et al.*, Clinical and photobiological characteristics of xeroderma pigmentosum complementation group F: a review of cases from Japan. *Br. J. Dermatol.* **121**, 471 (1989). doi:10.1111/j.1365-2133.1989.tb15514.x Medline

140. M.-S. Hung *et al.*, Functional polymorphism of the CK2alpha intronless gene plays oncogenic roles in lung cancer. *PLoS ONE* **5**, e11418 (2010). doi:10.1371/journal.pone.0011418 Medline

141. R. Davoli *et al.*, Mapping, identification of polymorphisms and analysis of allele frequencies in the porcine skeletal muscle myopalladin and titin genes. *Cytogenet. Genome Res.* **102**, 152 (2003). doi:10.1159/000075741 Medline

142. Y. Jiao, L. S. Zan, Y. F. Liu, H. B. Wang, B. L. Guo, A novel polymorphism of the MYPN gene and its association with meat quality traits in Bos taurus. *Genet. Mol. Res.* **9**, 1751 (2010). doi:10.4238/vol9-3gmr906 Medline

143. M. Sano *et al.*, Activation and function of cyclin T-Cdk9 (positive transcription elongation factor-b) in cardiac muscle-cell hypertrophy. *Nat. Med.* **8**, 1310 (2002). doi:10.1038/nm778 Medline

144. J. A. Buglino, M. D. Resh, Hhat is a palmitoylacyltransferase with specificity for N-palmitoylation of Sonic Hedgehog. *J. Biol. Chem.* **283**, 22076 (2008). doi:10.1074/jbc.M803901200 Medline

145. Y. Kawakami *et al.*, Isolation of a new melanoma antigen, MART-2, containing a mutated epitope recognized by autologous tumor-infiltrating T lymphocytes. *J. Immun.* **166**, 2871 (2001). Medline

146. M. J. Vincent *et al.*, Crimean-Congo hemorrhagic fever virus glycoprotein proteolytic processing by subtilase SKI-1. *J. Virol.* **77**, 8640 (2003). doi:10.1128/JVI.77.16.8640-8649.2003 Medline

147. C. Ling *et al.*, Identification of functional prolactin (PRL) receptor gene expression: PRL inhibits lipoprotein lipase activity in human white adipose tissue. *J. Clin. Endocrinol. Metab.* **88**, 1804 (2003). doi:10.1210/jc.2002-021137 Medline

148. I. A. Aligianis *et al.*, Mutations of the catalytic subunit of RAB3GAP cause Warburg Micro syndrome. *Nat. Genet.* **37**, 221 (2005). doi:10.1038/ng1517 Medline

149. H. Ehara *et al.*, Martsolf syndrome in Japanese siblings. *Am. J. Med. Genet.* **143A**, 973 (2007). doi:10.1002/ajmg.a.31626 Medline

150. D. A. van Heel *et al.*, A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.* **39**, 827 (2007). doi:10.1038/ng2058 Medline

151. C. Stoetzel *et al.*, Identification of a novel BBS gene (BBS12) highlights the major role of a vertebrate-specific branch of chaperonin-related proteins in Bardet-Biedl syndrome. *Am. J. Hum. Genet.* **80**, 1 (2007). doi:10.1086/510256 Medline

152. V. Marion *et al.*, Transient ciliogenesis involving Bardet-Biedl syndrome proteins is a fundamental characteristic of adipogenic differentiation. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1820 (2009). doi:10.1073/pnas.0812518106 Medline

153. D. Bem *et al.*, Loss-of-function mutations in RAB18 cause Warburg micro syndrome. *Am. J. Hum. Genet.* **88**, 499 (2011). doi:10.1016/j.ajhg.2011.03.012 Medline

154. Y. Ito *et al.*, The Mohawk homeobox gene is a critical regulator of tendon differentiation. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 10538 (2010). doi:10.1073/pnas.1000525107 Medline

155. W. G. Degen *et al.*, Expression of nma, a novel gene, inversely correlates with the metastatic potential of human melanoma cell lines and xenografts. *Int. J. Cancer* **65**, 460 (1996). doi:10.1002/(SICI)1097-0215(19960208)65:4<460::AID-IJC12>3.0.CO;2-E Medline

156. X. Luo *et al.*, Identification of BMP and activin membrane-bound inhibitor (BAMBI) as a potent negative regulator of adipogenesis and modulator of autocrine/paracrine adipogenic factors. *Diabetes* **61**, 124 (2012). doi:10.2337/db11-0998 Medline

157. N. S. Enattah *et al.*, Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233 (2002). doi:10.1038/ng826 Medline

158. S. A. Tishkoff *et al.*, Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31 (2007). doi:10.1038/ng1946 Medline

159. C. M. Schlebusch, P. Sjödin, P. Skoglund, M. Jakobsson, Stronger signal of recent selection for lactase persistence in Maasai than in Europeans. *Eur. J. Hum. Genet.* (2012). doi:10.1038/ejhg.2012.199 Medline

160. K. N. North *et al.*, A common nonsense mutation results in alpha-actinin-3 deficiency in the general population. *Nat. Genet.* **21**, 353 (1999). doi:10.1038/7675 Medline

161. A. B. Smith, Origins and Spread of Pastoralism in Africa. *Annu. Rev. Anthropol.* **21**, 125 (1992). doi:10.1146/annurev.an.21.100192.001013

162. K. Sadr, The first herders at the Cape of Good Hope. *Afr. Archaeol. Rev.* **15**, 101 (1998). doi:10.1023/A:1022158701778

163. D. Pleurdeau *et al.*, “Of sheep and men”: earliest direct evidence of caprine domestication in southern Africa at leopard cave (erongo, namibia). *PLoS ONE* **7**, e40340 (2012). doi:10.1371/journal.pone.0040340 Medline

164. X. Yi *et al.*, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75 (2010). doi:10.1126/science.1190371 Medline

165. G. Konopka *et al.*, Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462**, 213 (2009). doi:10.1038/nature08549 Medline

166. L. Feuk *et al*., Absence of a paternally inherited FOXP2 gene in developmental verbal dyspraxia. *Am. J. Hum. Genet.* **79**, 965 (2006). doi:10.1086/508902 Medline

167. W. Enard *et al*., Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**, 869 (2002). doi:10.1038/nature01025 Medline

168. S. Otsuki *et al*., Extracellular sulfatases support cartilage homeostasis by regulating BMP and FGF signaling pathways. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 10202 (2010). doi:10.1073/pnas.0913897107 Medline

169. A. Ratzka *et al*., Redundant function of the heparan sulfate 6-O-endosulfatases Sulf1 and Sulf2 during skeletal development. *Dev. Dyn.* **237**, 339 (2008). doi:10.1002/dvdy.21423 Medline

170. E. A. Otto *et al*., Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy. *Nat. Genet.* **42**, 840 (2010). doi:10.1038/ng.662 Medline

171. E. Schaefer *et al*., Molecular diagnosis reveals genetic heterogeneity for the overlapping MKKS and BBS phenotypes. *Eur. J. Med. Genet.* **54**, 157 (2011). doi:10.1016/j.ejmg.2010.10.004 Medline

172. A. Scherag *et al*., Two new loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and German study groups. *PLoS Genet.* **6**, e1000916 (2010). doi:10.1371/journal.pgen.1000916 Medline

173. R. Thomson, J. K. Pritchard, P. Shen, P. J. Oefner, M. W. Feldman, Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 7360 (2000). doi:10.1073/pnas.97.13.7360 Medline