A catalogue of early diverged contemporary human genome variation reveals distinct Khoe-San populations

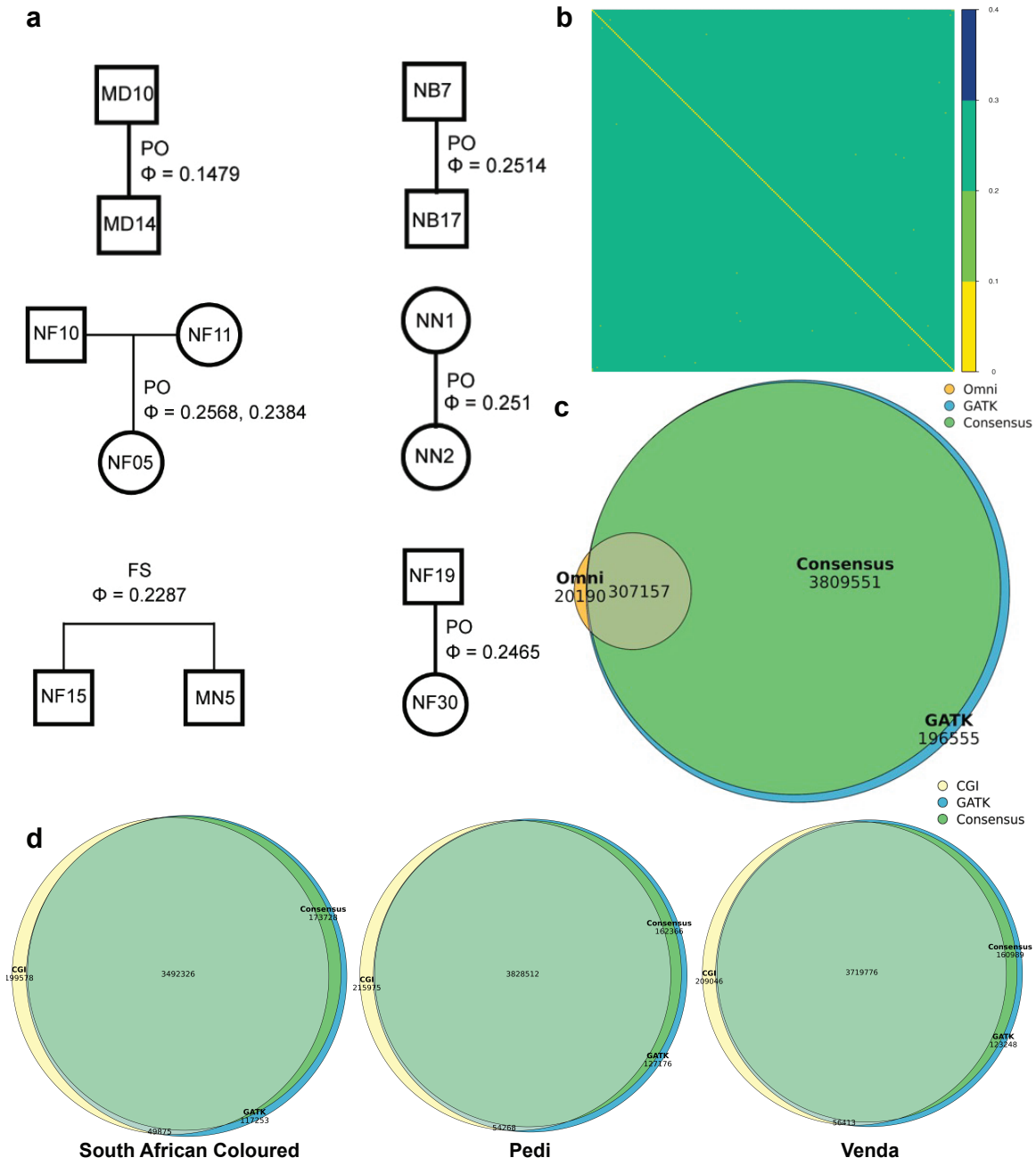
Weerachai Jaratlerdsiri[1,2,*], Pamela X.Y. Soh[1], Tingting Gong[1,3], Jue Jiang[1], Zolani Simayi[4], Desiree C. Petersen[5], Errol Holland[6], Eva K.F. Chan[7], Kathrine E. Theron[4], Wilfrid H.G. Haacke[8], Hagen E.A. Förtsch[9], M.S. Riana Bornman[10,11], David M. Thomas[12], Jeffrey Mphahlele[13,14], & Vanessa M. Hayes[1,9,15,16,*]

[1]Ancestry and Health Genomics Laboratory, Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Camperdown, New South Wales, Australia; [2]Computational Genomics Group, Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Camperdown, New South Wales, Australia; [3]Human Phenome Institute, Fudan University, Shanghai, China; [4]Faculty of Health Sciences, University of Limpopo, Mankweng, South Africa; [5]South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa; *[6]Formerly from Faculty of Medicine, Sefako Makgatho Health Sciences University, Pretoria, South Africa;* [7]Statewide Genomics, NSW Health Pathology, Newcastle, New South Wales, Australia; *[8]Formerly from the Department of Language and Literature Studies, University of Namibia, Windhoek, Namibia;* [9]Windhoek Central Hospital, *University of Namibia,* Windhoek Khomas, Namibia; [10]School of Health Systems and Public Health, University of Pretoria, Pretoria, Gauteng, South Africa; [11]Department of Biological Sciences, Faculty of Science, Engineering and Agriculture, University of Venda, Thohoyandou, South Africa. [12]Centre for Molecular Oncology, School of Biomedical Sciences, University of New South Wales Sydney, Randwick, New South Wales, Australia; *[13]*Department of Virology, National Health Laboratory Service and Sefako Makgatho Health Sciences University, Pretoria, South Africa; *[14]Office of the Deputy Vice Chancellor and Innovation, North-West University, Potchefstroom, South Africa;* [15]Manchester Cancer Research Centre, University of Manchester, Manchester M20 4GJ, United Kingdom; [16]Norwich Medical School, University of East Anglia, Norwich, United Kingdom

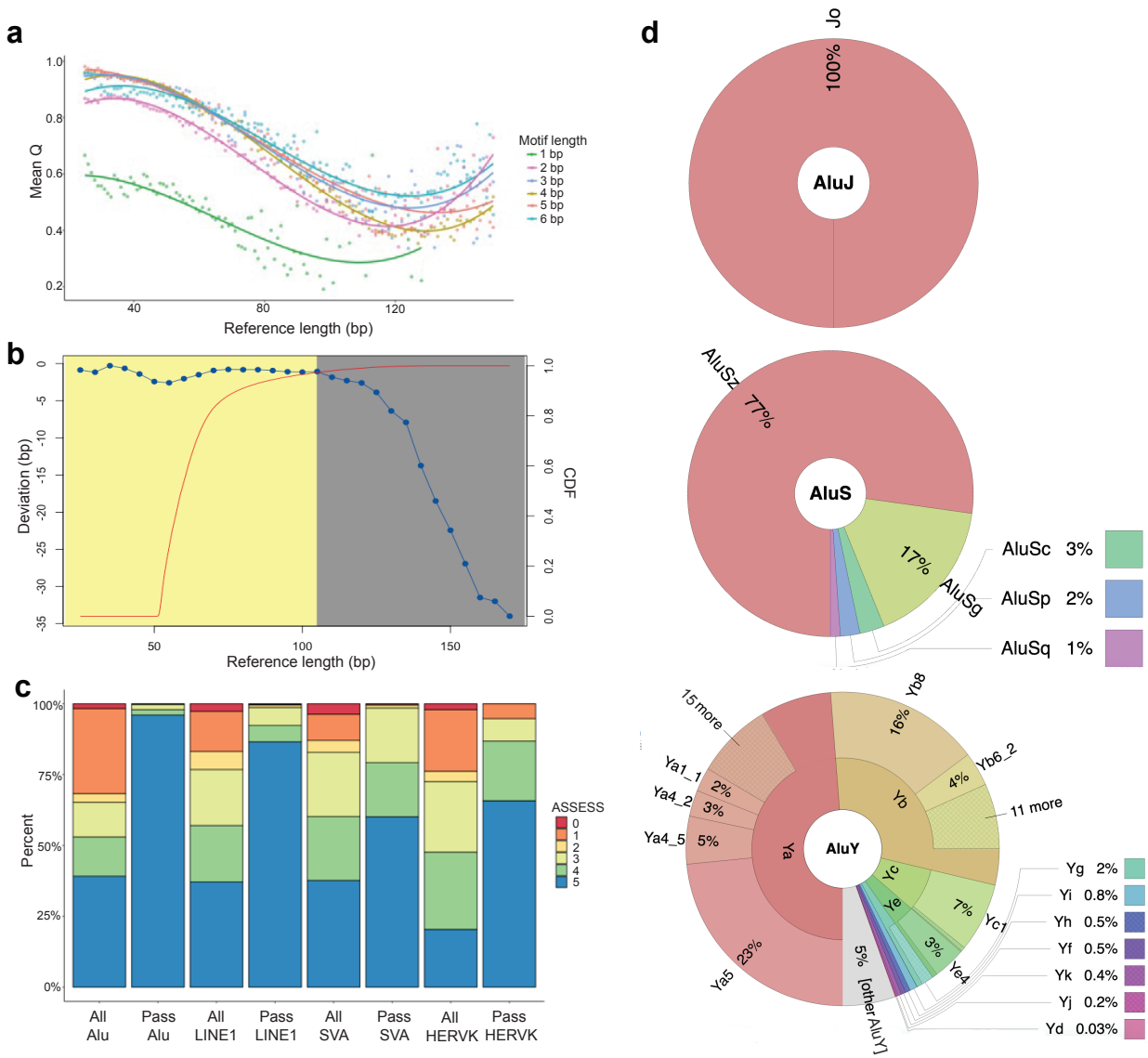
*Correspondence: vanessa.hayes@sydney.edu.au; weerachai.jaratlerdsiri@sydney.edu.au

SUPPLEMENTARY DATA

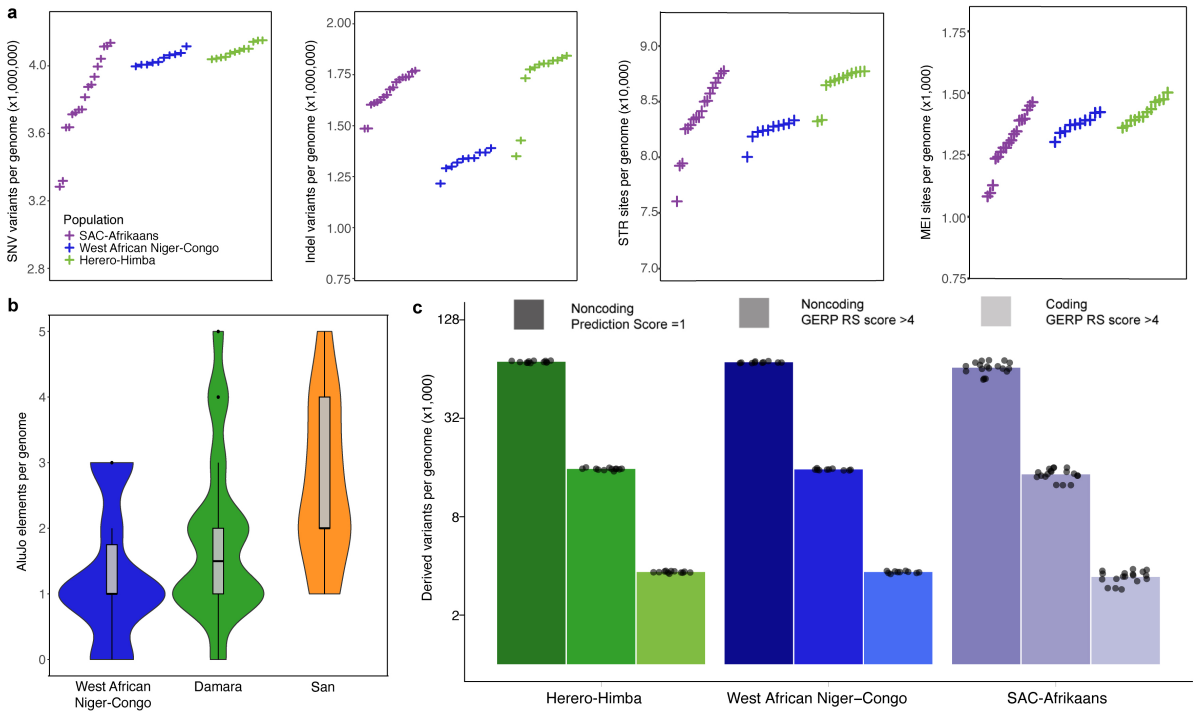
Supplementary Figures



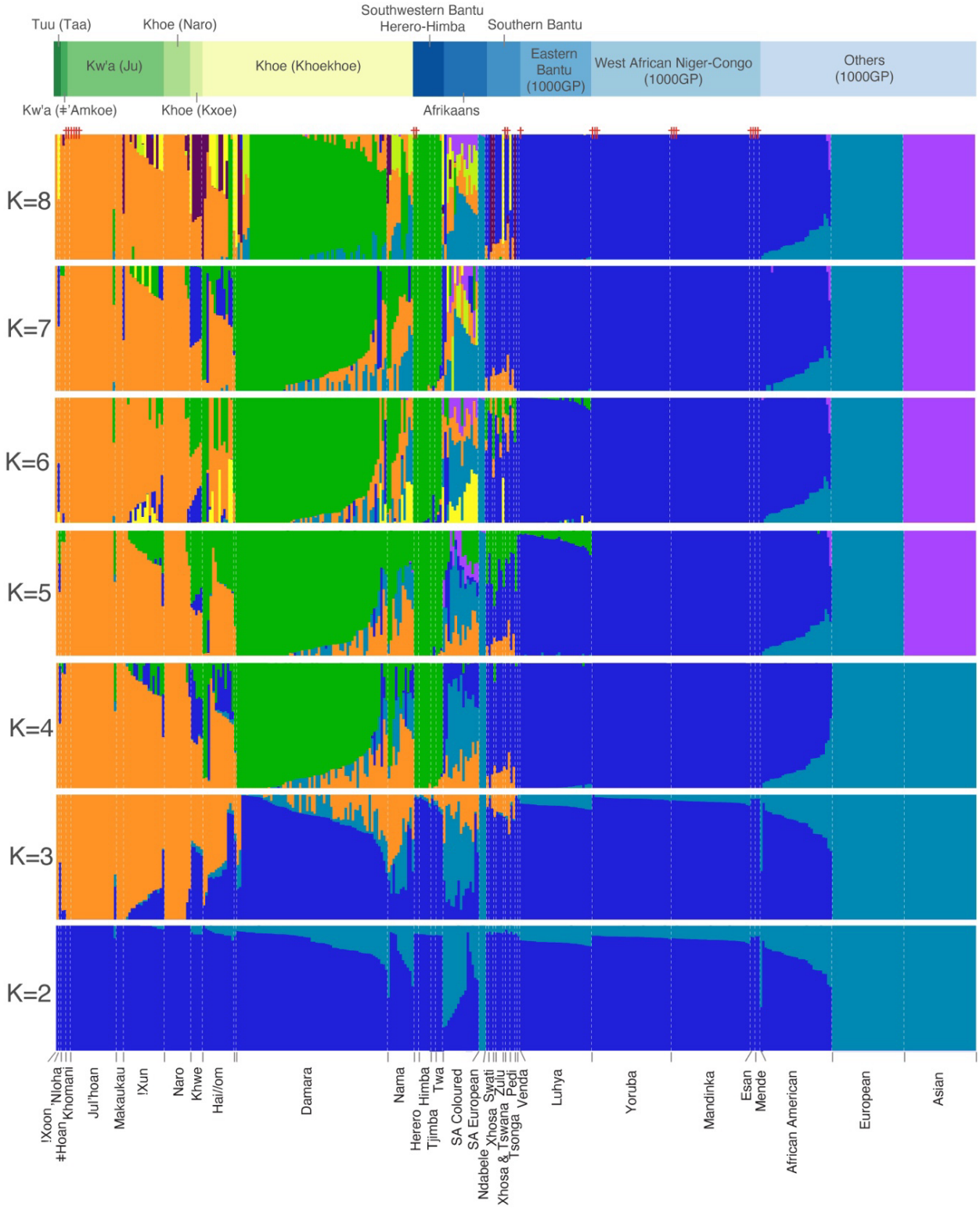
Supplementary Fig. 1 **WGS quality control. a** Revised pedigree information using WGS data. Degrees of relatedness (FS, full-sib and PO, parent-offspring) and kinship coefficients (Φ) are labelled on the side. A negative kinship coefficient indicates an unrelated relationship. Duplicate/monozygotic twin, 1st-degree, 2nd-degree, and 3rd-degree relationships correspond to the following coefficient ranges: >0.354, 0.177-0.354, 0.088-0.177 and 0.044-0.088. **b** Pairwise comparison scores across 226 individuals. A cut-off of more than 0.2 indicates no similarity between a pair of samples. **c** Single nucleotide variant detection comparing three methods, Omni-quad genotyping array, GATK and consensus call sets. Numbers indicate mean values for each comparison across 17 San individuals. The Euler diagram is plotted using eulerr package in R. **d** Single nucleotide variant comparisons between Complete Genomics Inc (CGI), GATK and consensus call sets using three genomes representing South African Coloured (SAC), Pedi and Venda individuals.



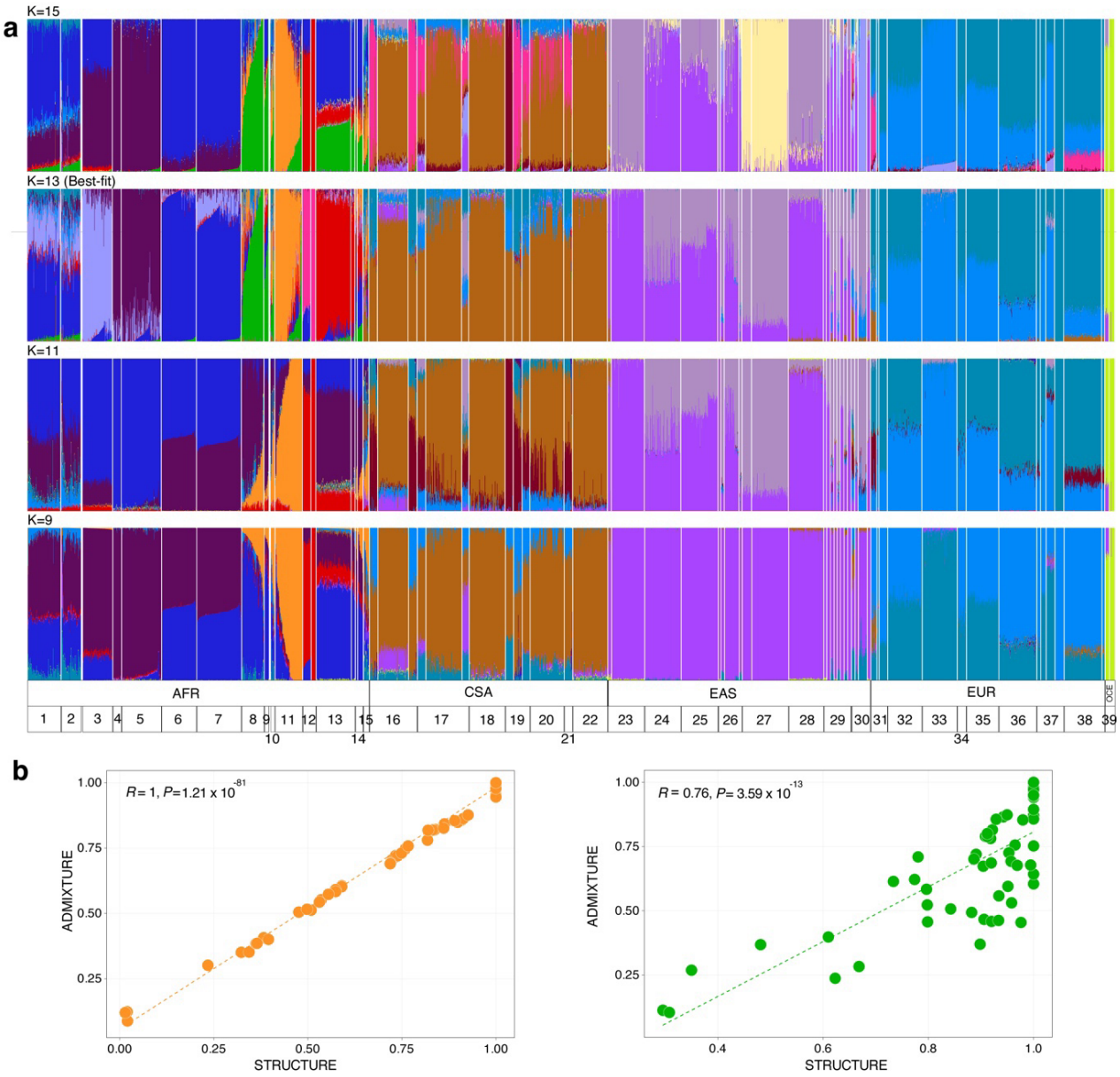
Supplementary Fig. 2 **Genomic rearrangement quality control. a** A decrease in mean genotype quality scores (Q) per sample across STR length. Each line represents different repeat motif lengths from chromosomes 1-Y (green = homopolymers; pink = dinucleotides; blue = trinucleotides; yellow = tetranucleotides; red = pentanucleotides; and aqua = hexanucleotides). No-call sites have no Q values specified. STR calls for African-specific SGDP genomes are excluded due to different lengths of sequencing reads. **b** Mean length deviation of raw STR data as a function of the hg38 reference allele length (blue curve). Negative deviation indicates lobSTR preference towards shorter alleles. STRs with reference alleles of up to 105 bp show minimal deviations (yellow region) and are expected to display unbiased frequency spectra for further analyses. These STR loci comprise close to 90% of the total genotyped STRs in our catalogue (red curve). **c** Validation of MEI feature discovery. MELT classifies passed breakpoints into one of the six following tranches on a scale from 0-5: 0, No overlapping reads; 1, Imprecise breakpoint; 2, Discordant pair evidence only; 3, Left-side target site duplication (TSD) evidence only; 4, Right-side TSD evidence only; and 5, TSD decided with split reads at both sides. **d** Krona plots of *Alu* families and subfamilies within this study



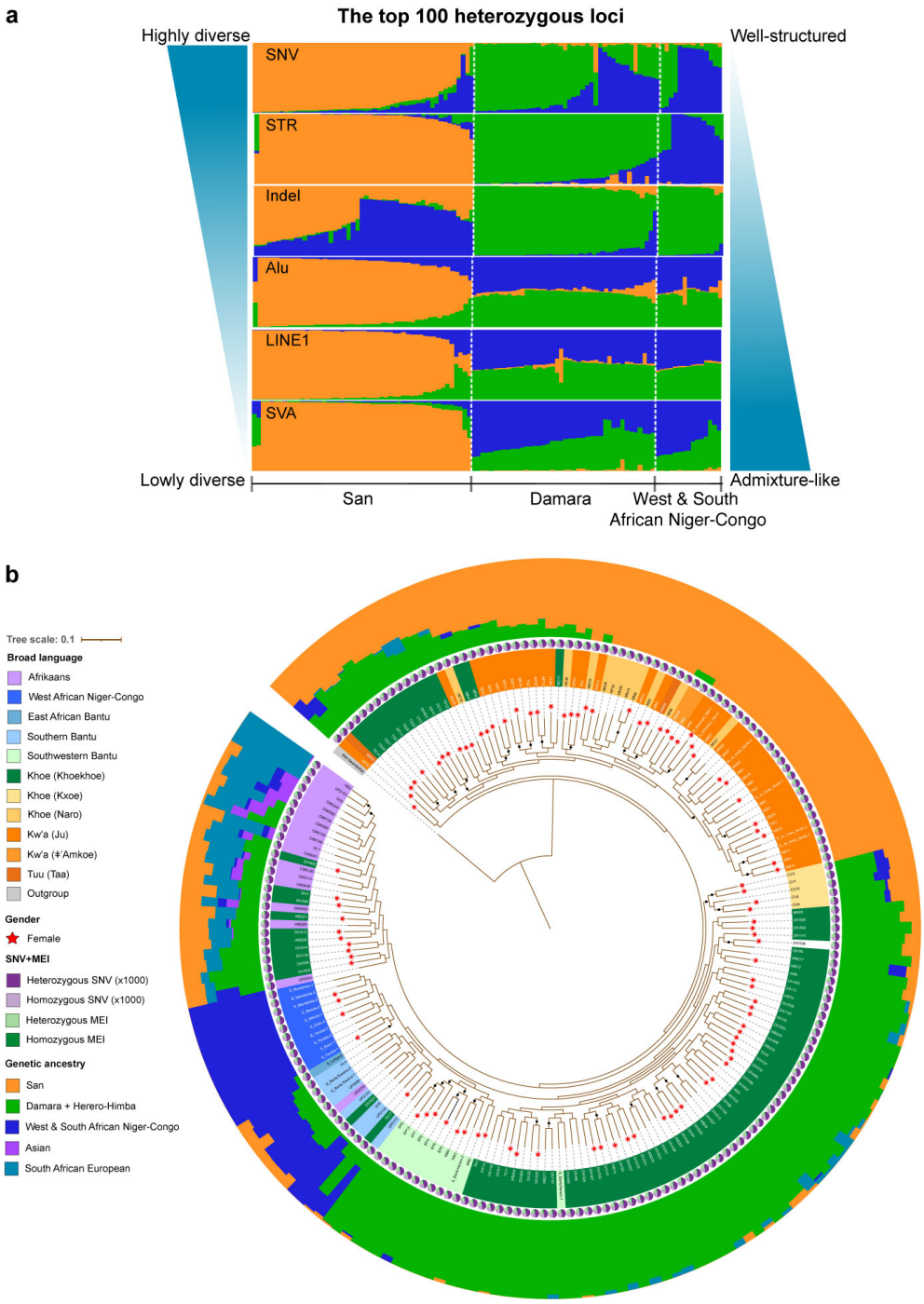
Supplementary Fig. 3 **Overall genomic variation of three broad language groups neighboring Namibian Khoe-San. a** The number of SNV, indel, STR and MEI markers per genome observed in Herero-Himba (Southwestern Bantu), West African Niger-Congo and SAC-Afrikaans populations. SAC, South African Coloured. **b** The number of *Alu*Jo elements per genome observed in San (n = 50), Damara (n = 42) and West African Niger-Congo (n = 10). The box plots show the median (centre line), the 25th and 75th percentiles (box limits), and ±1.5× the interquartile range (whiskers). **c** The average number of accumulated/derived variants per genome. Data are presented as mean values. GERP scores of RS > 4 suggest deleterious single nucleotide variants; GenoCanyon prediction probability equal to one shows the functional potential of noncoding regions in a human genome.



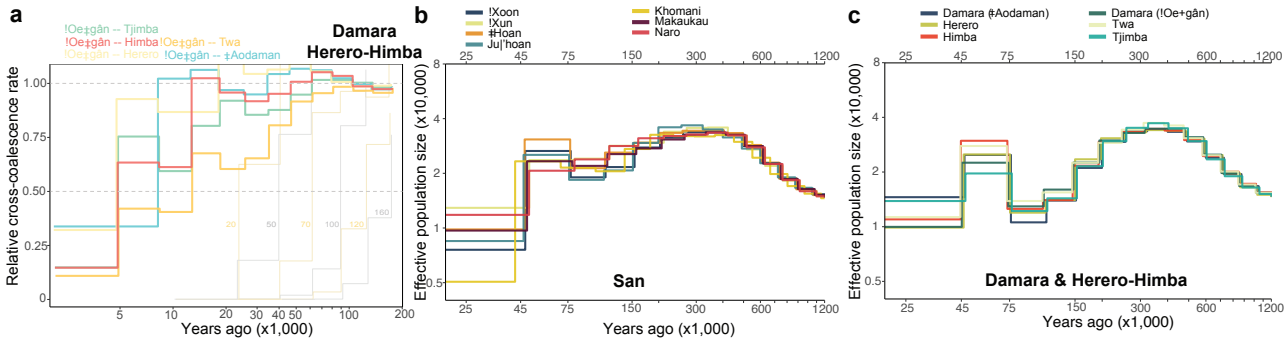
Supplementary Fig. 4 **STRUCTURE analysis of a full SNV set with the logistic prior model.** The analysis was run with five replicates with randomly chosen initial seeds, varying the number of ancestral populations from K = 2 and K = 8. The full range of K is shown to illustrate the finer-scale population structure of the data. SA, South African.



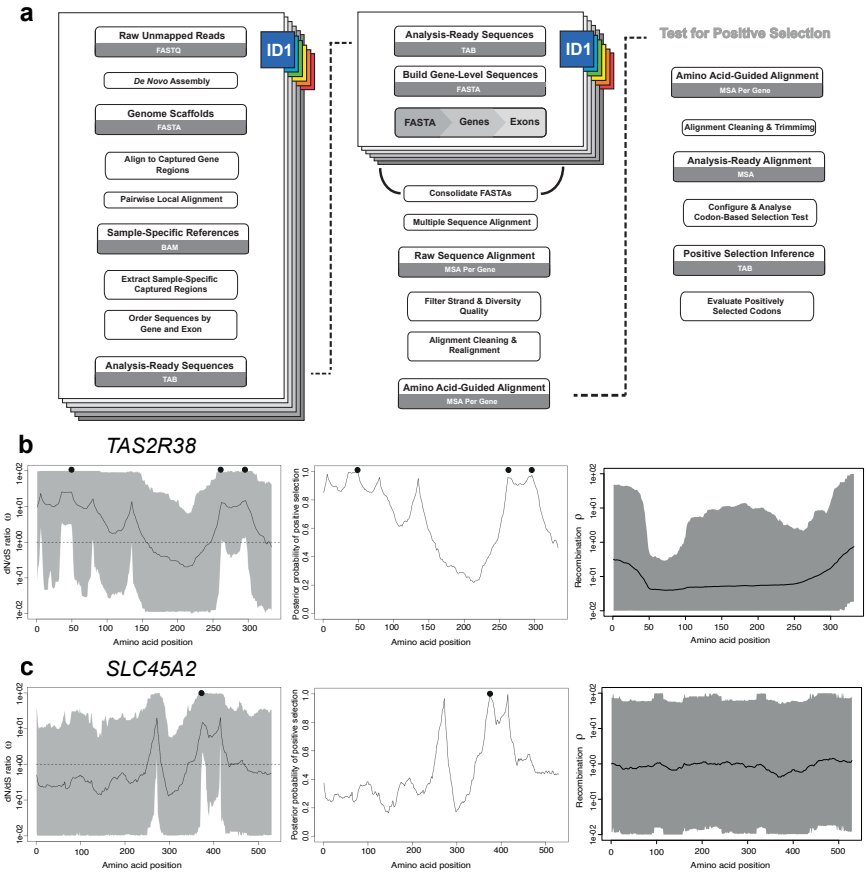
Supplementary Fig. 5 **Admixture clustering analysis (K = 9 – 15) of 2,964 whole genomes from our data (n = 183) and previous studies**. **a** K = 13 produced the lowest mean cross-validation error at 0.297514 ± 0.0001. All populations analyzed and references are shown in Supplementary Data 7. Broad population labels are Africa (AF R), Central South Asia (CSA), East Asia (EAS), Europe (EUR) and Oceania (OCE), with 'San' (orange, populati on #11) and 'Damara' (green, population #8) clearly differentiated. Notably, at K = 15 an unconfirmed 'Damara'-like ancestral fraction appears in Kenyan Bantu (population #13). **b** Pearson correlation coefficients and their significance comparing STRUCTURE and ADMIXTURE analyses for 'San' (orange) and 'Damara' (green) whole genome data, respectively.



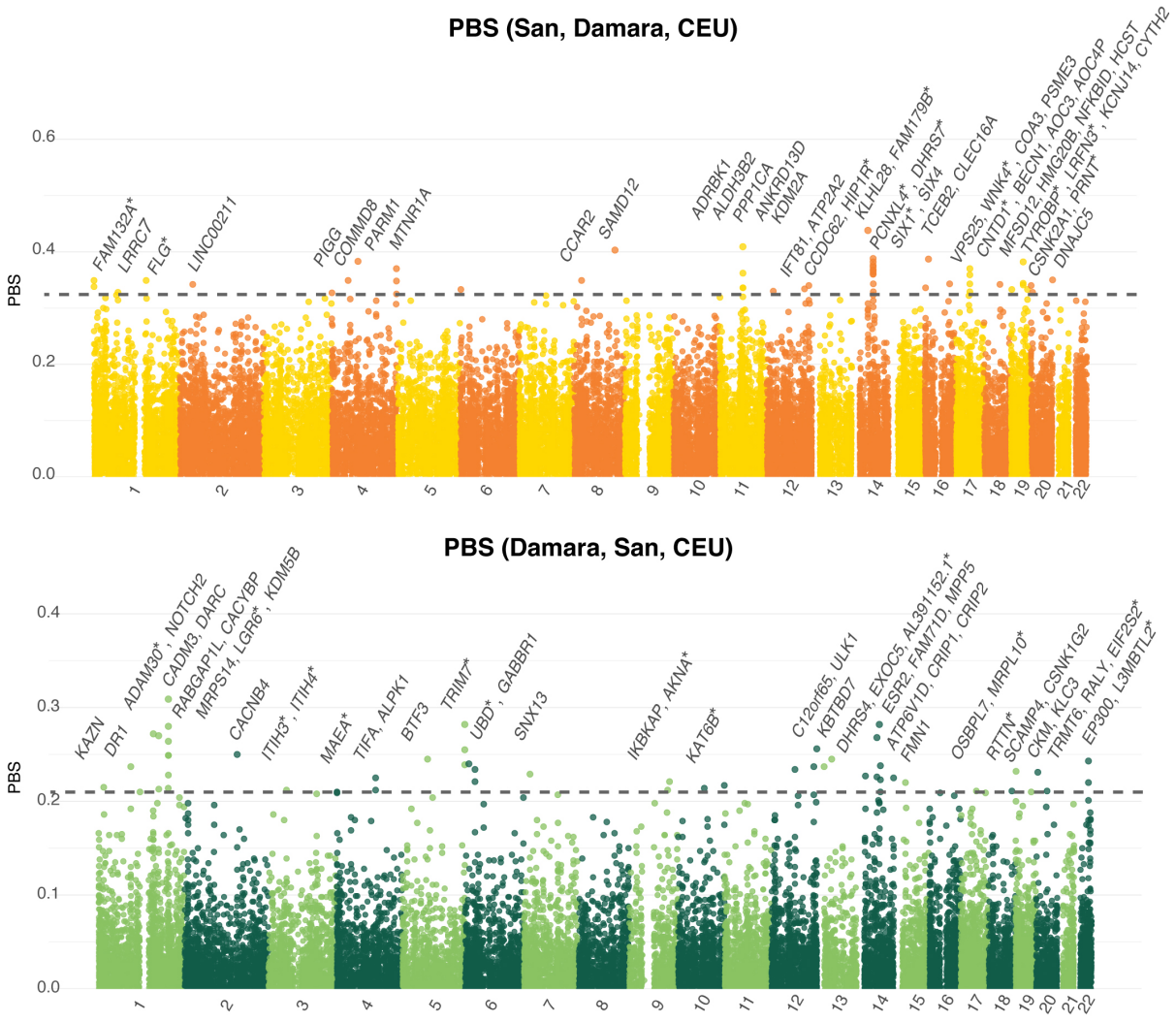
Supplementary Fig. 6 **Full-spectrum genetic analysis of variant types. a** Classical STRUCTURE analysis of the top 100 heterozygous loci per variant type among 50 'San', 42 'Damara', and 15 West and South African Niger-Congo individuals with little admixture detected in Figure 2. The analysis was run with five replicates for K = 3, with randomly chosen initial seeds. There are none of reference populations from the 1000 Genomes Project used in this analysis. With mean heterozygosity ranging from 0.32-0.50, SNV and STR markers show to be the top two heterozygous loci, regardless of total number of available variants. Indels have the mean heterozygosity of 0.27, with the mean value of MEI markers (*Alu*, SVA and LINE1) ranging from 0.07-0.08. The result shows a better population subdivision using SNV binary markers and STR multiallelic markers. **b** Maximum Likelihood tree of the full SNV dataset present among unrelated individuals (n=8,491,748 markers) using the Altai Neanderthal as an outgroup[51]. Highlighted colors indicate different language groups assigned in Supplementary Table 1. A percentage of bootstrap values (BV) greater than 90 is provided in each branch.



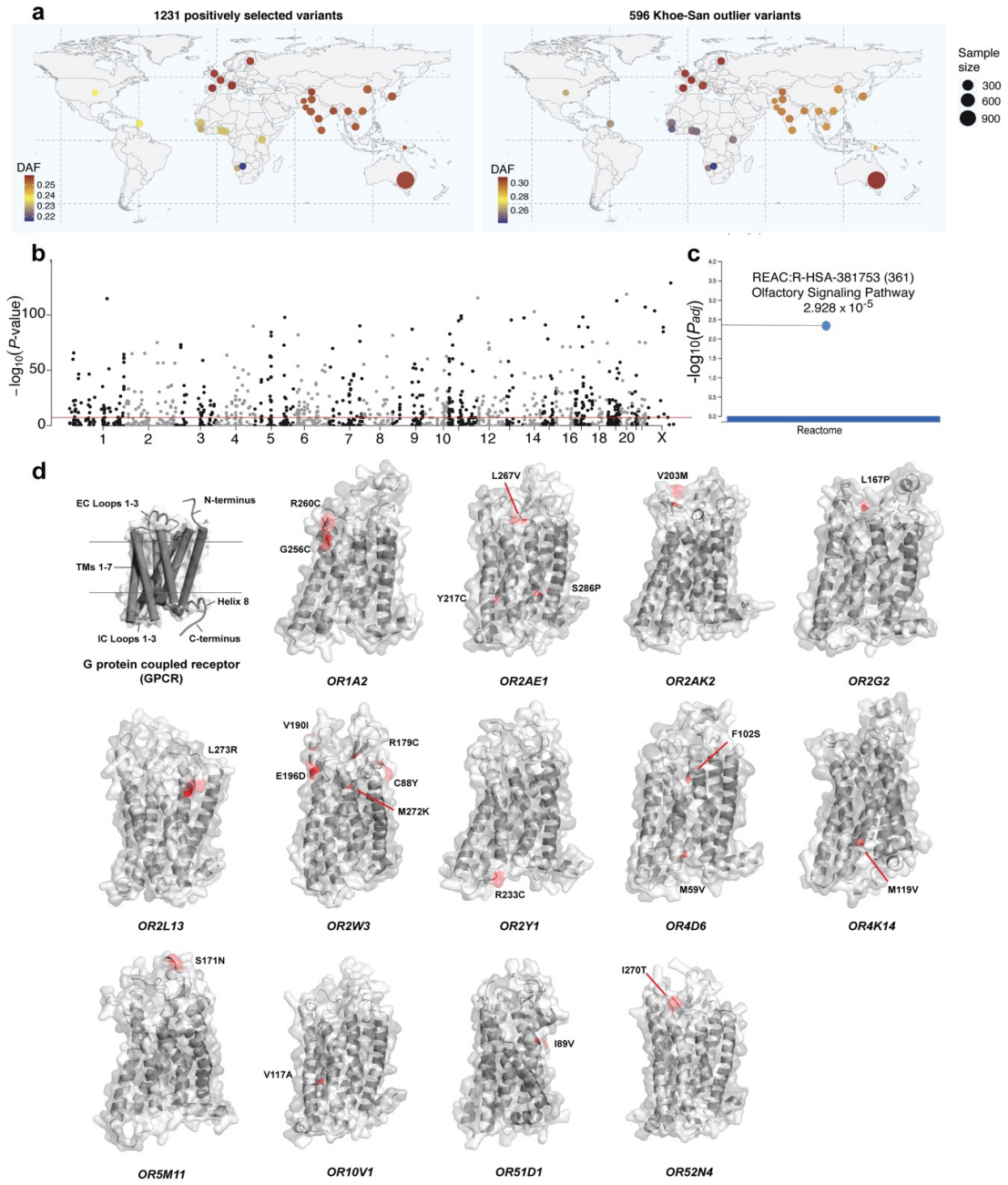
Supplementary Fig. 7 **Population separation and effective population size through time in forager populations. a** Cross-coalescence rates for pairs of Damara and Herero-Himba speaking populations. **b** Effective population sizes of seven San populations inferred using PSMC. The mutation rate of 1.25 x 10^{-8} per base pair per generation and generation time of 30 years is assumed. **c** Effective population sizes of two 'Damara' subgroups and four populations of Herero-Himba speaking peoples.



Supplementary Fig. 8 **Positive selection inference. a** Assembly-based genome analysis workflow for sequence data manipulation/processing and subsequent tests for positive selection across sequence data. ID indicates different sample labels. More details are provided in Supplementary Information 9. **b** Selection detection tests performed by Bayesian inference among 37 genome assemblies. Graphs on the left panel show change in *dN/dS* ratios (ω) across amino acid positions. Lines represent a mean of *dN/dS* value; gray areas indicate 95 % highest posterior probability dense intervals; and a dashed line shows an ω value equal to one across amino acid positions. In the middle panel, fluctuation in the posterior probability (0-1) of positive selection across amino acid positions calculated by Bayesian inference using the omegaMap program[60]. Solid circles indicate amino acids reported to be under positive selection and present to have significant ω peaks in our study. The right panel indicates a degree of recombination adjusted during selection scans.



Supplementary Fig. 9 **PBS statistic peaks between forager populations and gene annotation. a** Those of 'San'. **b** Those of 'Damara'. Dashed lines indicate outlier PBS scores greater than 0.324 for **a** and 0.210 for **b** at empirical *P*-value < 0.001. Asterisks indicate the presence of nonsynonymous variants using SNPnexus (https://www.snp-nexus.org/v4/).



Supplementary Fig. 10 **Signature of positive selection within African foragers. a** Derived allele frequency (DAF) of forager (75 'San' and 63 'Damara'), 1,144 MGRB and 2,472 GNOMAD individuals successfully merged for 1,231/2,840 positively selected sites and 596/633 Khoe-San outlier variants. Ancestral and derived alleles are inferred from the Ensembl Compara 93 database. Southwestern US samples represent African ancestry. A world map was drawn using ggplot2 and sf in R. **b** A total of 633 selected variants or 479 genes significantly differed between MGRB and forager populations ('San' and 'Damara'). A red line indicates a conservative significance cut-off at $-\log_{10}(5 \times 10^{-5})$. Only $\log_{10}(P)$ values ranging from 0-150 are shown. **c** Pathway enrichment analysis of the 479 genes, with the most significant pathway related to olfactory signalling using Benjamini-Hochberg adjusted *P*-value. **d** Structural models of GPCRs among the associated genes related to an olfactory signaling pathway. Structural segments of GPCRs are as follows[72]. C-terminus, extracellular (EC) loops, transmembrane helices (TMs 1-7), intracellular (IC) loops, Helix 8 and N-terminus. Ribbon and transparent surface representation of a homology model of the GPCR per gene based on the alignment of the query amino acid sequence, with multiple known GPCR structures. Residues in red correspond to nonsynonymous variants within GPCR regions, and those in dark gray indicate hydrophobic amino acids.

SUPPLEMENTARY DATA and TABLES GUIDE

Supplementary Data 1. Population inclusion and sequencing quality

Supplementary Data 2. qSignature sample mix-up results

Supplementary Data 3. Genotype results comparing between different methods

Supplementary Data 4. Khoe-San Genome Project genomic variation

Supplementary Data 5. STR genomic variation

Supplementary Data 6. MEI genomic variation

Supplementary Data 7. Previous studies of genetype arrays and WGS analyzed in this study

Supplementary Data 8. Admixture events and dating by fastGLOBETROTTER

Supplementary Data 9. Positively selected sites detected across 1,376 genes

Supplementary Data 10. XP-EHH selection scans in Khoe-San

Supplementary Data 11. Significant selected sites in Khoe-San forager populations

Supplementary Table 1. Whole genome sequencing references analyzed in this study

Supplementary Table 2. Composition of our custom STR reference panel (hg38; 307,149 loci in total) for autosomes and sex chromosomes (X and Y).

Supplementary Table 3. Mean heterozygosity of STR loci by chromosome

Supplementary Table 4. Common LoF alleles (hg38 reference)

Supplementary Table 5. Gene annotation of MEIs

Supplementary Table 6. High-pass whole genome data used for phylogenomic reconstruction

Supplementary Table 7. Samples used in MSMC analyses

Supplementary Information

Table of Contents

SI1 – Recruitment of study participants and community engagement 13

SI2 – Processing of sequencing data 14
2.1. Sequencing and data delivery 14
2.2. Whole genome sequence reference panels 14
2.3. Data preprocessing and alignment 15
2.4. Germline small variant discovery 16

SI3 – Data quality control 19
3.1. Test for sample mix-ups 19
3.2. Comparison of genotypes obtained by different methods 19
3.3. Relatedness 21

SI4 – Small genomic variation 22
4.1. General genomic variation within Khoe-San 22
4.2. Differences of derived allele frequency between populations 22
4.3. GERP and GenoCanyon scores 23

SI5 – Genomic rearrangements 24
5.1. Short tandem repeat 24
5.2. Mobile element discovery 30

SI6 – Population structure analysis 32

SI7 – Phylogenomic reconstruction 35
7.1. Full SNV phylogeny 35
7.2. Direct mapping consensus phylogeny 36

SI8 – PSMC and MSMC analysis 39
8.1. PSMC 39
8.2. MSMC2 39

SI9 – Assembly-based genome analysis 43

SI10 – Positive selection inference 46

SI11 – References 50

SI1 – Recruitment of study participants and community engagement

Study participants recruited within the borders of Namibia were either recruited via collaboration with the Namibian blood bank in Windhoek, which included signed consents and acknowledgement of local Namibians Sr Sinvula and previous head of the blood bank Mr Wilkinson, or within remote communities focused within the eastern corridor (bordering Botswana), north-eastern greater Kalahari Nyae Nyae region or the more northerly (Angola border) to north-central region. For the latter communities, Hayes has been visiting and engaging with the communities since her first 'work-related' visit in 2008 and most recent in 2023. Born, raised and educated in South Africa, with family citizens of Namibia, Hayes is well acquainted with the land, region and peoples, having made well over 40 visits over her lifetime. As such, she is not only well acquainted with the cultures and practices, but also in a position to communicate directly with the 'San' communities via the common language of Afrikaans (see https://www.hayeslab.net/videos). In agreement with the Namibian Ministry of Health and Social Services (MoHSS), all work-related visits to the region have required video recording to provide a record of Hayes's work and engagement within the communities. As such, Hayes travels with a videographer, see acknowledgement of Mr Bennett from Evolving Picture. It was also agreed that informed consent would take place through community discussions and verbal consent (see https://www.hayeslab.net/videos). Hayes is also a fully qualified phlebotomist and takes full responsibility for performing all recruitments for video-consented study participants. As such, Hayes has engaged extensively with local Namibians from community leaders, community advocates to ministers and local farmers to provide further language-specific translations and ensure cultural integrity, see further acknowledgement of E. Adams, A.A. Collins, R. Friederich, B. G/aq'o, N. /kun, J. /kunta, H. Mische, F. Naque, D. Naque, H. Oosthuizen, E. Oosthuizen, A. Oosthuysen, E. Oosthuysen, D. Roux, C. Swau, and T. Tsebe. A critical criterion for community participation is what the community wishes to achieve from generating and sharing their genetic information. After 15 years of engagement their goal has never changed, the older generation want their stories to be recorded, they are acutely aware that times are changing fast and they don't want the younger generation to forget the 'old ways', they want the younger generation to be proud of their past and who they represent.

Statement from Hayes. "Just because someone cannot read or write, does not mean they are not educated or understand their past and position in the world. From my experience, the Khoe-San peoples have a clear and deep understanding of their role in human history and how their story is written in the 'tracks' (DNA) of the blood that we draw. Most importantly, it is the hours around the campfire where we as academics will learn their histories, passed down verbally from generation to generation. While we bring the science and our findings back to the communities, it is the discussions that follow that are the most insightful."

SI2 – Processing of sequencing data

2.1. Sequencing and data delivery

Library kits used for Truseq Nano DNA HT Library Preparation Kit (https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-nano-dna.html) at KCCG Core Facility Sequencing Services, Garvan.

Libraries were created as per the manufacturer's instructions. One sample was loaded per flow cell lane. The flow cells were loaded onto an Illumina HiSeq X sequencer, and 2 x 150bp paired-end sequencing was performed.

The raw data from the sequencers were converted to FASTQ file format using Illumina's bcl2fastq 2.16.0. FastQC was run on the FASTQ data. Sequences were aligned to the human reference genome, human.g1k.v37 using Illumina's iSAAC aligner v01.14.11.11[1] to generate BAM files. Quality metrics were calculated using Picard WgsMetrics v1.119 (http://broadinstitute.github.io/picard/) and are shown in Supplementary Data 1.

The following metrics were estimated:

Overall performance of each sample:
Raw sequencing Yield (megabases) and the percentage of bases that had quality ≥ Q30

Poor quality reads:
PCT_EXC_BASEQ: The fraction of aligned bases of low base quality (default < 20)
PCT_EXC_CAPPED: The fraction of aligned bases that would raise coverage above the capped value (default cap = 250X)
PCT_EXC_DUPE: The fraction of aligned bases in reads marked as duplicates
PCT_EXC_MAPQ: The fraction of aligned bases in reads with low mapping quality (default <20)
PCT_EXC_OVERLAP: The fraction of aligned bases in the second observation from an insert with overlapping reads

Proportion of the genome sequenced above a certain average:
PCT_5X: The fraction of bases that attained at least 5X sequence coverage in post-filtering bases
PCT_10X: The fraction of bases that attained at least 10X sequence coverage in post-filtering bases
PCT_20X: The fraction of bases that attained at least 20X sequence coverage in post-filtering bases
PCT_30X: The fraction of bases that attained at least 30X sequence coverage in post-filtering bases

2.2. Whole genome sequence reference panels

We processed whole genome raw reads of relevant genomes from a previous publication[2] using the same alignment and variant calling pipeline described in Section 2.3. There are 22 African and two Eurasian genomes (Supplementary Table 1).

The alignment data of the Altai Neanderthal[3] was downloaded from relevant publication (http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/). The mapping quality of the alignment will be explained in relevant sections.

Supplementary Table 1 Whole genome sequencing references analyzed in this study

ID	Run ID	Reference	Sex	Population	Region	Mean coverage[a]
Altai_Neanderthal	NA	Prüfer et al. 2014	Female	Neanderthal	Siberia	50.37
B_Ju_hoan_North-4	ERR1395596	Mallick et al, 2016	Male	Ju_hoan_North	Africa	40.93
B_Mandenka-3	ERR1395593	Mallick et al, 2016	Male	Mandenka	Africa	39.35
B_Yoruba-3	ERR1395598	Mallick et al, 2016	Male	Yoruba	Africa	42.74
S_BantuHerero-1	ERR1347714	Mallick et al, 2016	Male	BantuHerero	Africa	43.05
S_BantuHerero-2	ERR1019043	Mallick et al, 2016	Male	BantuHerero	Africa	39.45
S_BantuTswana-1	ERR1347732	Mallick et al, 2016	Male	BantuTswana	Africa	41.12
S_BantuTswana-2	ERR1347723	Mallick et al, 2016	Male	BantuTswana	Africa	36.19
S_Brahmin-1	ERR1395559	Mallick et al, 2016	Male	Brahmin	SouthAsia	46.84
S_Esan-1	ERR1025621	Mallick et al, 2016	Male	Esan	Africa	32.70
S_Esan-2	ERR1025622	Mallick et al, 2016	Female	Esan	Africa	38.13
S_Han-1	ERR1019055	Mallick et al, 2016	Female	Han	EastAsia	65.76
S_Ju_hoan_North-1	ERR1347737	Mallick et al, 2016	Male	Ju_hoan_North	Africa	43.31
S_Ju_hoan_North-2	ERR1019075	Mallick et al, 2016	Male	Ju_hoan_North	Africa	31.81
S_Ju_hoan_North-3	ERR1019076	Mallick et al, 2016	Male	Ju_hoan_North	Africa	33.47
S_Khomani_San-1	ERR1395588	Mallick et al, 2016	Female	Khomani_San	Africa	42.17
S_Khomani_San-2	ERR1395572	Mallick et al, 2016	Female	Khomani_San	Africa	46.58
S_Luhya-1	ERR1347662	Mallick et al, 2016	Female	Luhya	Africa	35.52
S_Luhya-2	ERR1347655	Mallick et al, 2016	Male	Luhya	Africa	32.96
S_Mandenka-1	ERR1025640	Mallick et al, 2016	Male	Mandenka	Africa	33.73
S_Mandenka-2	ERR1025641	Mallick et al, 2016	Female	Mandenka	Africa	36.25
S_Mende-1	ERR1347671	Mallick et al, 2016	Male	Mende	Africa	38.10
S_Mende-2	ERR1347678	Mallick et al, 2016	Female	Mende	Africa	35.75
S_Yoruba-1	ERR1025657	Mallick et al, 2016	Female	Yoruba	Africa	33.15
S_Yoruba-2	ERR1025658	Mallick et al, 2016	Male	Yoruba	Africa	31.31

[a] only chr20 used for coverage estimation using 'samtools depth'

2.3. Data preprocessing and alignment

Raw FASTQ data were evaluated using FastQC. Leading and trailing stretches of Ns and bases at the 3' end with quality scores under 15 were trimmed from the reads, and only reads greater than 70 nucleotides without any Ns were kept using cutadapt v1.8.3[4]. Any Kmer content immediately at the 5' end of reads frequently found was also discarded. Filtered paired reads were aligned with the human reference genome hg38 with alternate contigs (ALT), following Genome Analysis Toolkit (GATK) Best Practices[5]. Given the GATK's hg38 resources, the reference contains: 25 assembled chromosomes (chr1-Y, plus rCRS mitochondrion); 42 unlocalized sequences (_random suffix); 127 unplaced sequences (chrUn_prefix); 261 alternate contigs (_alt suffix); 525 HLA contigs; the Epstein-Barr virus sequence; and 2,385 decoy contigs. Including this decoy prevents false mappings from genuine human sequences absent in the reference. Homologous centromeric and genomic repeat arrays on multiple autosomes and two PAR (pseudoautosomal) regions on chromosome Y are hard masked with Ns in this reference.

***Mapping to a reference genome*:**

The ALT-aware version of bwa mem was used to align FASTQ reads to take advantage of alternative loci in hg38[6]. This aligner prioritizes alignments on the primary assembly if no alignment is detected within the alternate contigs and enables variant calling on all the alternate contigs.

*run-bwamem -t 16 -dko READS.bam *
*-HR '@RG\tID:ReadGroup\tPL:ILLUMINA\tPU:None\tLB:1\tSM:SampleName\tCN:hcpcg' *
Homo_sapiens_assembly38.fasta R1.fastq.gz R2.fastq.gz | sh

***Marking duplicates*:**

Read alignment was sorted, and read duplicates were marked using picard v2.20.0 (https://broadinstitute.github.io/picard/).

java -jar picard.jar MarkDuplicates INPUT=$IN.bam OUTPUT=$OUT.bam METRICS_FILE=$OUT.metrics.txt

***Indel realignment*:**

For each sample, alignment was realigned at indel intervals using GATK-3.5-0[5].

java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R Homo_sapiens_assembly38.fasta -I $IN.bam -o $OUT.realign.intervals –known Homo_sapiens_assembly38.known_indels.vcf -known Mills_and_1000G_gold_standard.indels.hg38.vcf

java -jar GenomeAnalysisTK.jar -T IndelRealigner -R Homo_sapiens_assembly38.fasta -I $IN.bam -targetIntervals $OUT.realign.intervals -o $OUT.bam --filter_bases_not_stored -known Homo_sapiens_assembly38.known_indels.vcf -known Mills_and_1000G_gold_standard.indels.hg38.vcf

***Base quality score recalibration*:**

For each sample, alignment was corrected for patterns of systematic errors in the base quality scores.

java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -R Homo_sapiens_assembly38.fasta -I $IN.bam -o $OUT.recal.grp -knownSites Homo_sapiens_assembly38.dbsnp138.vcf -knownSites Mills_and_1000G_gold_standard.indels.hg38.vcf -knownSites Homo_sapiens_assembly38.known_indels.vcf --bqsrBAQGapOpenPenalty 30

java -jar GenomeAnalysisTK.jar -T PrintReads -R Homo_sapiens_assembly38.fasta -I $IN.bam -BQSR $OUT.recal.grp -o $OUT.bam

2.4. Germline small variant discovery

***Calling variants per sample*:**

To generate a robust consensus dataset, recalibrated alignment per sample was called for single nucleotide variants (SNVs) and indels using two callers, GATK HaplotypeCaller (GVCF mode) [5] and FreeBayes v1.2.0[7]. Per-sample GVCFs were used for joint genotyping across ~205 whole genomes (GATK GenotypeGVCFs). Joint-called SNVs and indels were filtered via machine learning variant quality score recalibration, and passed loci were kept. Individual FreeBayes results were filtered out if Phred-scaled quality scores were under 20. Filtered variant calls from

both callers per sample were decomposed and left-aligned using the vt 0.57 program (http://genome.sph.umich.edu/wiki/vt) and classified as SNVs and indels using vcflib v1.0.0-rc2 (https://github.com/vcflib/vcflib). Genomic coordinates of each variant type from the two tools were intersected individually using VCFtools 0.1.14[8]. Per-sample consensus call sets were counted to compare variant differences.

i) GATK

java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R Homo_sapiens_assembly38.fasta -I $IN.bam -o $OUT.g.vcf.gz --dbsnp Homo_sapiens_assembly38.dbsnp138.vcf -ERC GVCF -variant_index_type LINEAR -variant_index_parameter 128000 -pairHMM VECTOR_LOGLESS_CACHING

java -jar GenomeAnalysisTK.jar -T GenotypeGVCFs -R Homo_sapiens_assembly38.fasta --dbsnp Homo_sapiens_assembly38.dbsnp138.vcf -o $OUT.vcf.gz -nda -V $IN1.g.vcf.gz -V $IN2.g.vcf.gz

java -jar GenomeAnalysisTK.jar -T VariantRecalibrator -R Homo_sapiens_assembly38.fasta -input $IN.vcf.gz -resource:hapmap,known=false,training=true,truth=true,prior=15.0 hapmap_3.3.hg38.vcf -resource:omni,known=false,training=true,truth=true,prior=12.0 1000G_omni2.5.hg38.vcf -resource:1000G,known=false,training=true,truth=false,prior=10.0 1000G_phase1.snps.high_confidence.hg38.vcf -resource:dbsnp,known=true,training=false,truth=false,prior=2.0 Homo_sapiens_assembly38.dbsnp138.vcf -an DP -an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR -mode SNP -tranche 100.0 -tranche 99.5 -tranche 99.0 -tranche 90.0 -recalFile $OUT.recal -tranchesFile $OUT.tranches -rscriptFile $OUT.R

java -jar GenomeAnalysisTK.jar -T ApplyRecalibration -R Homo_sapiens_assembly38.fasta -input $IN.vcf.gz -mode SNP --ts_filter_level 99.5 -recalFile $IN.recal -tranchesFile $IN.tranches -o $OUT.vcf

java -jar GenomeAnalysisTK.jar -T VariantRecalibrator -R Homo_sapiens_assembly38.fasta -input $IN.vcf.gz -resource:mills,known=true,training=true,truth=true,prior=12.0 Mills_and_1000G_gold_standard.indels.hg38.vcf -resource:dbsnp,known=true,training=false,truth=false,prior=2.0 Homo_sapiens_assembly38.dbsnp138.vcf -an DP -an QD -an MQRankSum -an ReadPosRankSum -an FS -an SOR -mode INDEL -tranche 100.0 -tranche 99.5 -tranche 99.0 -tranche 90.0 --maxGaussians 4 -recalFile $OUT.recal -tranchesFile $OUT.tranches -rscriptFile $OUT.R

java -jar GenomeAnalysisTK.jar -T ApplyRecalibration -R Homo_sapiens_assembly38.fasta -input $IN.vcf.gz -mode INDEL --ts_filter_level 99.0 -recalFile $IN.recal -tranchesFile $IN.tranches -o $OUT.vcf

ii) FreeBayes

freebayes -f Homo_sapiens_assembly38.fasta $IN.bam | vcffilter -f "QUAL > 20" | gzip -c > $OUT.Q20.vcf.gz

iii) Consensus set

vcf-isec -f -n =2 IN1.vcf.gz IN2.vcf.gz | bgzip -c > $OUT.vcf.gz

***Genotype Refinement workflow*:**

Genomic coordinates of per-sample consensus call sets were merged to extract consensus calls from joint-called multi-sample variant data generated after GATK variant quality score recalibration. Only biallelic SNVs in the joint-called data were chosen for further refinement to collect a final dataset for downstream population genetic analyses discussed in relevant sections. The following refinement involves using variation data from the 1,000 Genomes project to improve genotype quality and calls and to filter out low-quality calls from the dataset.

i) Derive posterior probabilities of genotypes

java -jar GenomeAnalysisTK.jar -R Homo_sapiens_assembly38.fasta -T CalculateGenotypePosteriors --supporting 1000G_phase1.snps.high_confidence.hg38.vcf -ped $IN.ped -V $IN.vcf -o $OUT.vcf

ii) Filter low quality genotypes

java -jar GenomeAnalysisTK.jar -T VariantFiltration -R Homo_sapiens_assembly38.fasta -V $IN.vcf -G_filter "GQ < 20.0" -G_filterName lowGQ -o $OUT.vcf

java -jar GenomeAnalysisTK.jar -T SelectVariants -R Homo_sapiens_assembly38.fasta -V $IN.vcf.gz -o $OUT.vcf.gz --setFilteredGtToNocall

SI3 – Data quality control

3.1. Test for sample mix-ups

To avoid possible sample mix-ups between individuals, their recalibrated alignment data described above (BAM) were tested using qSignature-0.1, with default setting (https://sourceforge.net/projects/adamajava/). A total of 1,484,044 SNV positions were searched per genome, and the average Euclidean distance between common SNVs in a pair of samples was calculated. Comparison scores greater than 0.2 were set as a cut-off of pairwise discordance.

3.2. Comparison of genotypes obtained by different methods

***Illumina Omni-quad 1M array*:**

Genotyping data of 17 samples identified as 'San' were generated using the HumanOmni1-Quad v1.0 DNA Analysis BeadChip (July 2011). Their genotypes were filtered out with a minor allele frequency (MAF) less than 5%, a Hardy Weinberg Equilibrium (HWE) *P*-value failed at 0.0001, and the missing rate of individual and SNV greater than 10%. Concordance analysis for the array and our WGS consensus and passed GATK callsets were conducted per sample. Only autosomes were compared due to haploid calls for sex chromosomes in the array.

Concordance analyses showed that, on average, 93.88% of variant sites from genotyping array overlapped with our WGS calls (309,610/329,799; Supplementary Data 3 and Supplementary Fig. 1c). The mean overlap of genotyping array among 17 'San' genomes was 307,157 and 309,610 sites per genome, with that of consensus and GATK call sets, respectively. About 20,190 variants per sample (range 18,222-22,200) were specific to the DNA array. The consensus call set was robust, including 95.39% (4,116,708/4,315,716) of the original GATK calls per genome.

***Complete Genomics sequencing*:**

To study ethnic diversity within South Africa, ten genomes were submitted for deep-coverage whole genome sequencing (WGS) service at the Complete Genomics Inc. (CGI; http://www.completegenomics.com/), with three overlapped with the current Illumina sequencing (UL1, UL4, and UL6). CGI sequencing provides high-quality human genomes, with 45- to 87-fold coverage[9,10]. DNA samples were quality-checked for their optimum concentrations (10 μg) using the Quant-iTTM PicoGreen® dsDNA Kit and their integrity of high molecular weight and double-stranded intactness using Gel Electrophoresis. The genomic DNA was circularized and amplified to build copies of DNA nanoballs, a compact structure resistant to degradation. Combinatorial probe anchor ligation sequencing, unchained hybridization and ligation techniques were employed to produce raw base reads and associated scores.

Read alignment and variant discovery was conducted in-house at the CGI[9]. Reads were aligned with the Genome Reference Consortium human genome build 37 (GRCh37). SNVs, insertions and deletions were sought across the genome, with alignment correction using the greedy

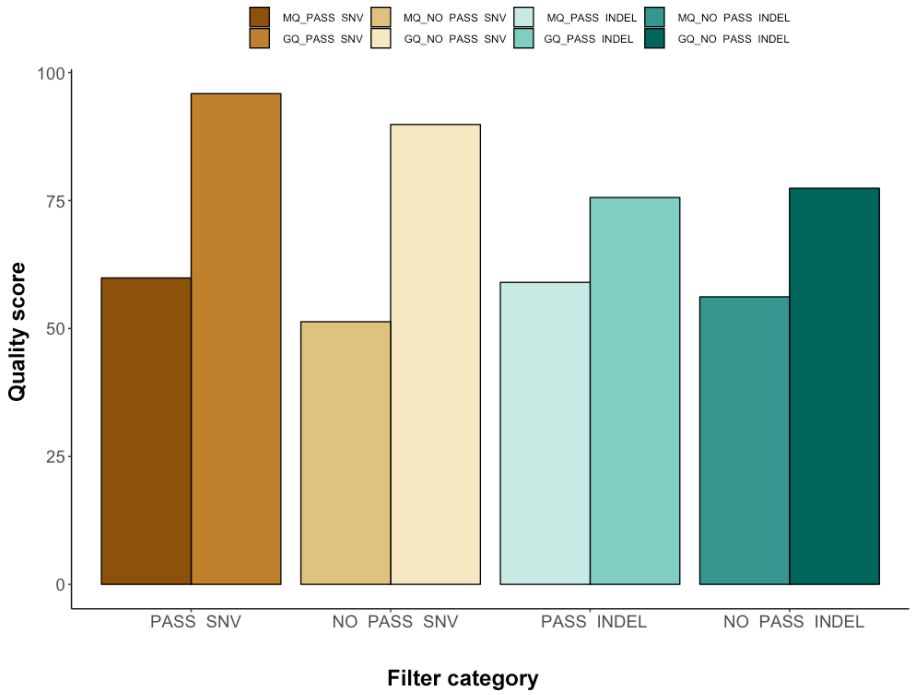
optimization process of Bayesian probability statistics supplemented with a local *de novo* assembly until convergence to an optimum diploid sequence was observed[11]. Variant calling processes assigned each genomic position to diploid references, variants, or no calls with variant quality scores. Autosomal variants with high confidence scores (VQHIGH) were chosen and compared with robust consensus and more sensitive GATK call sets from the same samples as described above.

Comparing CGI data of three high-quality whole genomes with our variant calling pipeline re vealed a positional concordance rate of around 95% for SNVs (Supplementary Data 3 and Supplementary Fig. 1d). Although the number of SNVs differed among three genomes representing different populations (range 3,741,779-4,098,755), the rate was still similar among them. On average, these three genomes had around 96% of SNVs in the original GATK present in their corresponding consensus call sets.

***Variant calling workflow for de novo assembly*:**

Genome assemblies available in 17 'San' individuals (Supplementary Information 9; Supplementary Data 9) were mapped to hg38 plus alternate contigs, and small variants were called using bcftools (http://www.htslib.org/doc/bcftools.html). Any ambiguous bases (IUPAC codes) within the assembly were replaced with an 'N' base before variant calling. For 33 other 'San' samples, filtered paired reads described in Section 2.3 were assembled into unitigs, mapped to the reference genome, and then called for small variants using FermiKit-0.13, *de novo* assembly-based variant caller[2]. To compare assembly-based and mapping-based variant calling pipelines, the raw joint-called GATK dataset, before being filtered in the variant quality score recalibration step (see 'Calling variants per sample' in Section 2.4), was used to extract all variants from each of the genomes. Each variant dataset per sample and caller was decomposed and left-aligned using the vt 0.57 program (http://genome.sph.umich.edu/wiki/vt) and classified as SNVs and indels using vcflib v1.0.0-rc2 (https://github.com/vcflib/vcflib). Genomic coordinates of each variant type from both GATK and assembly-based callers were intersected individually using VCFtools 0.1.14[8]. Per-sample consensus call sets were counted to compare the different number of variants observed to be high-confidence (PASS) or suboptimal (other than PASS but MQ >20 and GQ >20) as assigned by the GATK caller.

On average, we observed an additional 307,313 SNVs and 54,567 indels per genome compared to high-confidence or passed variants. Those small variants were initially ignored due to their suboptimal scores during machine learning variant quality score recalibration. Their mapping and genotyping quality values (MQ and GQ) were also high with, on average, 53.75 and 83.64, respectively, and relatively similar to those of the high-confidence variants (Supplementary Fig. 11). Call sets from *de novo* assembly representing a different procedure from a mapping-based calling pipeline could support suboptimal variants in other data to balance sensitivity and error rate.



Supplementary Fig. 11 **Averaged mapping and genotyping quality of high-confidence and suboptimal variants (SNVs and indels) identified using GATK HaplotypeCaller.** Suboptimal variants (NO PASS) were kept for downstream analyses if supported by any assembly-based variant callers.

3.3. Relatedness

To identify relatedness in this study, the refined joint-called SNV dataset mentioned in Section 2.4 was filtered based on a missing rate per SNV and sample, MAF and HWE at 0.1, 0.01 and 0.0001, respectively. A total of 15,603,438 remaining variants were used for relationship inference using KING version 2.1.6[12]. The analysis implemented integrated algorithms, pair-wise kinship coefficients and identical by descent (IBD) segments and could account for up to 3rd degree of relatedness, allowing the refinement of pedigree information. Any first-degree relationship (full-sib or parent-offspring) was subjected to exclusion for further genetic analyses[13]. The dataset of 205 individuals remained was reanalyzed to confirm their absence of 1st-degree relatedness.

***Biallellic SNV filtering*:**

./plink2 --chr 1-22 --vcf $IN.vcf.gz --mind 0.1 --maf 0.01 --geno 0.1 --hwe 0.0001 --out $OUT--make-bed --double-id --allow-extra-chr

***Relationship inference*:**

./king -b $IN --related --degree 3 --prefix $OUT

There were six families in this study, each of which had one or two members excluded in subsequent analyses of unrelated samples (Supplementary Fig. 1a). We chose to exclude NB7, NF05, NF19, NN2, MD14, MN5, and TS15. Five genomes, MN14, NB14, NF12, and NF27, were kept due to 2nd/3rd-degree relationships observed, although their qSignature scores were between 0.179-0.199 (Supplementary Data 2).

SI4 – Small genomic variation

4.1. General genomic variation within Khoe-San

***Variant data preparation and processing*:**

Per-sample consensus datasets of 198 genomes for SNVs and indels described above were counted for genome and exome variation on an individual basis for chromosomes 1-Y (Supplementary Data 4). Functional annotation and deleterious alterations of variant data per individual were performed using ANNOVAR and hg38 databases[14]. The Welch two-sample t-test was computed in R[15] for the number of variants between any two populations.

***Principal component analysis*:**

To assess relationships between populations studied in this study, we performed principal component analysis (PCA) using biallelic SNV consensus data after a genotype refinement step in Section 2.4. The call set of autosomes within unrelated individuals was filtered using PLINK 2.00[16], with a missing rate per SNV/sample and MAF at 0.1 and 0.01, respectively. A total of 3,281,257 variants were remained for PCA. The joint-called consensus dataset for autosomal indels generated in Section 2.4 was filtered out if less than 90% genotyping rate and subset to include biallelic sites with the expected heterozygosity greater than 0.1 (see Equation 2 for details). A total of 707,573 indels without missing genotypes were kept, and numeric values 0, 1 and 2 were assigned to homozygous references, heterozygous and homozygous alleles, respectively. The normalization step of these indels was computed following Equation 1[17], and the PCA of the normalized data was carried out and plotted in the R packages FactoMineR and factoextra[15].

Equation 1

$$M_{(i,j)} = \frac{C_{(i,j)} - \mu_{(j)}}{\sqrt{p_{(j)}\,(1 - p_{(j)})}}$$

$C_{(i,j)}$ = *Number of variant alleles for marker j, individual i (0, 1 or 2)*

$$\mu_{(j)} = \frac{\sum_{i=1}^{m} C_{(i,j)}}{m}$$

$p_{(j)} = \mu_{(j)}/2$; *An allele frequency estimate*

4.2. Differences of derived allele frequency between populations

***Ancestral allele inference*:**

The joint-called consensus call set processed after a refinement step in Section 2.4 was subjected to the analysis of derived allele variation. The analysis required biallelic, autosomal SNVs of unrelated individuals that were removed if the variant's missing rate was greater than 10% or a genome contained excess no-calls (10%). The ancestral state of the remaining variants was inferred using the ancestral hg38 genome (Ensembl compara 93 database), the consensus

alignment of 12 primate genomes released in the Ensembl Compara 93 database[18]. To identify derived alleles, the ancestral consensus sequence was used as reference alleles in PLINK 2.0[16]. Only the same reference bases supported in the ancestral genome were chosen to extract polymorphic SNVs and use them in further analyses. Derived variants outside genic regions and CpG islands were defined by Ensembl version 87, Build 37[18] and the UCSC Table Browser[19], respectively.

4.3. GERP and GenoCanyon scores

***Database retrieval and derived variant dataset*:**

Genomic Evolutionary Rate Profiling (GERP) scores were estimated to quantify the level of evolutionary constraint acting on derived SNVs, with the following mutational effects[20]: *i*) neutral (-2 < GERP RS < 2); *ii*) slightly deleterious (2 ≤ GERP RS ≤ 4); *iii*) deleterious (4 < GERP RS < 6); and *iv*) strongly deleterious (GERP RS ≥ 6). Base-wise GERP scores (hg19) quantified as rejected substitutions (RS) scores were retrieved from the UCSC genome browser[21]. The UCSC binary utility, *bigWigAverageOverBed* was implemented to obtain actual RS score data.

bigWigAverageOverBed All_hg19_RS.bw $INPUT.bed $OUTPUT

GenoCanyon prediction scores (hg19 version 1.0.3) were retrieved from Lu et al[22]. For each genomic position, these posterior probability scores of derived variants could measure the functional potential of noncoding regions in a human genome through unsupervised learning of 22 public annotation data. A prediction score equal to one indicates a strong functional signal[22].

The derived variant dataset of 205 genomes described in Section 4.2 was processed to extract autosomal SNVs, where their ancestral bases (ensembl_compara_93 database) were identical to the human reference hg38. A total of 25,219,597 derived sites remained were lifted to the UCSC hg19, using liftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) and merged with GERP and GenoCanyon databases based on their genomic coordinates. GERP and GenoCanyon prediction scores were then annotated into the derived variant dataset using the bcftools program.

***Runs of homozygosity*:**

To identify a long run of homozygosity (ROH) due to recent inbreeding rather than by chance, the pruned variants generated in Section 4.1 were carried out in PLINK 2.00. Parameters included a minimum run length of 5,000 kb with a minimum number of 50 variants and runs containing no more than 1 heterozygote and up to 5 missing genotypes. The maximum gap between variants was set to 1,000 kb, with a minimum density of one variant per 50 kb in a run. The number of ROH *versus* the total sum of ROH was plotted using RStudio v1.1.463[15].

SI5 – Genomic rearrangements

5.1. Short tandem repeat

***Building a custom reference and profiling short tandem repeats*:**

We aim to generate the most comprehensive catalogue of short tandem repeat (STR) loci using deep whole genome sequencing among 198 genomes (excluding 7 CGI genomes). Recalibrated alignment data of the genomes described in Section 2.3 were used to identify STRs using lobSTR v4.0.6[23]. Therefore, a custom set of STR loci for our hg38 analysis set were generated using Tandem repeats finder[24], with the following recommended settings to achieve well-aligned tandem repeats (at least 25 bp):

trf409 Homo_sapiens_assembly38.fasta 2 7 7 80 10 50 500 -f -d -m

First, each sample's read alignment was realigned to STR-containing regions of a custom hg38 reference (excluding CODIS and Y-STR markers), using the lobSTR aligner v4.0.6 with a default setting to avoid caveats of multiple mapping of reads. Second, the consequent sorted alignment of STRs was used to specify STR alleles at each locus using the lobSTR allelotype, with default settings and version 3 of lobSTR's Illumina PCR free noise model. We jointly genotyped all the realigned genomes across 307,149 sites in our custom reference panel, with motif lengths ranging from 1-6 bp. Supplementary Table 13 provides a summary of the reference panel.

Supplementary Table 2 Composition of our custom STR reference panel (hg38; 307,149 loci in total) for a utosomes and sex chromosomes (X and Y).

Motif length	No. loci	Common repeat units in order
1 bp	69,342	T, A
2 bp	102,364	AC, TG, GT, AT, TA
3 bp	16,615	AAT, TTA, TTG, AAC, TAT
4 bp	71,960	AAAT, TTTA, AAAC, TTTG, AAAG
5 bp	29,097	AAAAC, TTTTG, AAAAT
6 bp	17,771	TTTTTG, AAAAAC, AAAAAG, AAAAAT

***Filtering STR loci*:**

To obtain confident STR calls across all study genomes for downstream analyses, multi-sample calling data generated above were subjected to the following filtering criteria. Related individuals and three outliers (CV4, DIV95 and MD2) for STR typing were excluded. Following the manual's instruction, we filtered out loci with average coverage less than 5X, average Q smaller than 0.84, and call rates (percent of samples with a genotype call) lower than 0.8. We applied a length greater than 105 bp, not 80 bp, for reference allele length because we generated longer sequencing reads (maximum 150 bp) and tested based on the lobSTR's allelic spectrum analysis (Supplementary Fig. 2b). The Welch two-sample t-test for the number of STR variants between two populations was calculated in R[15].

***Heterozygosity and principal component analyses*:**

To measure STR diversity, per-site heterozygosity (Equation 2) and its mean across all loci were computed on the filtered STR dataset of 198 genomes. The analysis compared the diversity of STRs with their genomic regions, motif length (1-6 bp) and reference allele length. The expected heterozygosity (H_E) ranges from zero to nearly one (if containing a large number of equally frequent alleles), and its equation is as follows:

Equation 2

$$H_E = 1 - \sum_{i=1}^{n} q_i^2$$

q_i^2 = *Frequency of* i^{th} *allele of n alleles at a locus*

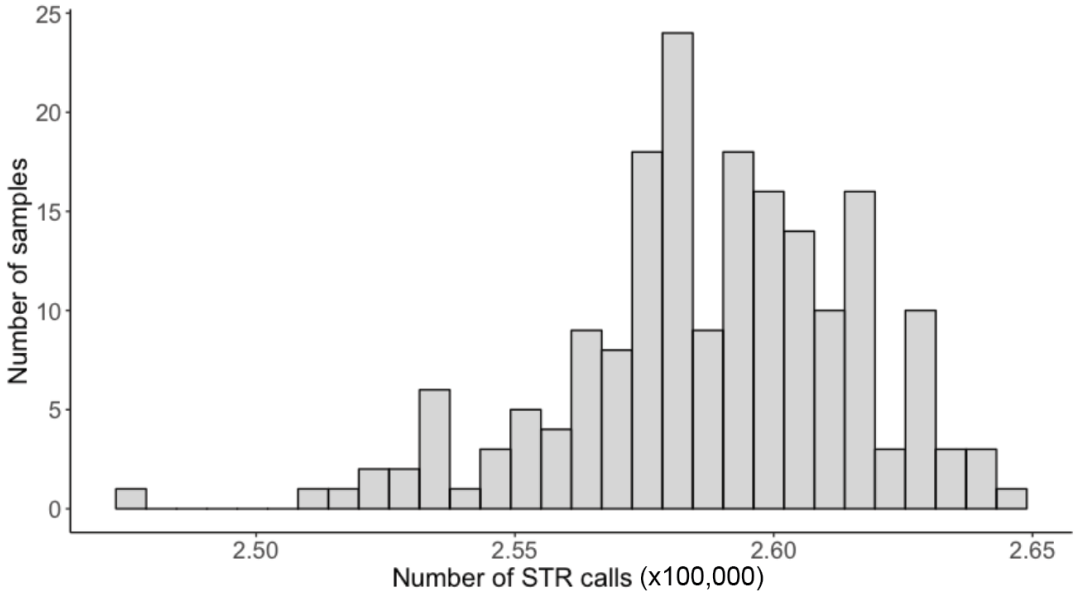
PCA and data normalization of filtered STRs were performed using the R packages FactoMineR and factoextra[15]. The STRs were chosen if per-site heterozygosity was greater than 0.1. The STR loci and their base pair length differences of any two diploid alleles across all the genomes (relative to the hg38 reference; GB field of the lobSTR variant call set) were extracted. The sum of the length of any two alleles at the same locus was calculated for normalization and PCA analyses[17]. A total of 29,856 loci were retained after ambiguous sites (the sum of both strands of heterozygous loci was equal to reference allele length) were removed.

***Annotating STR loci*:**

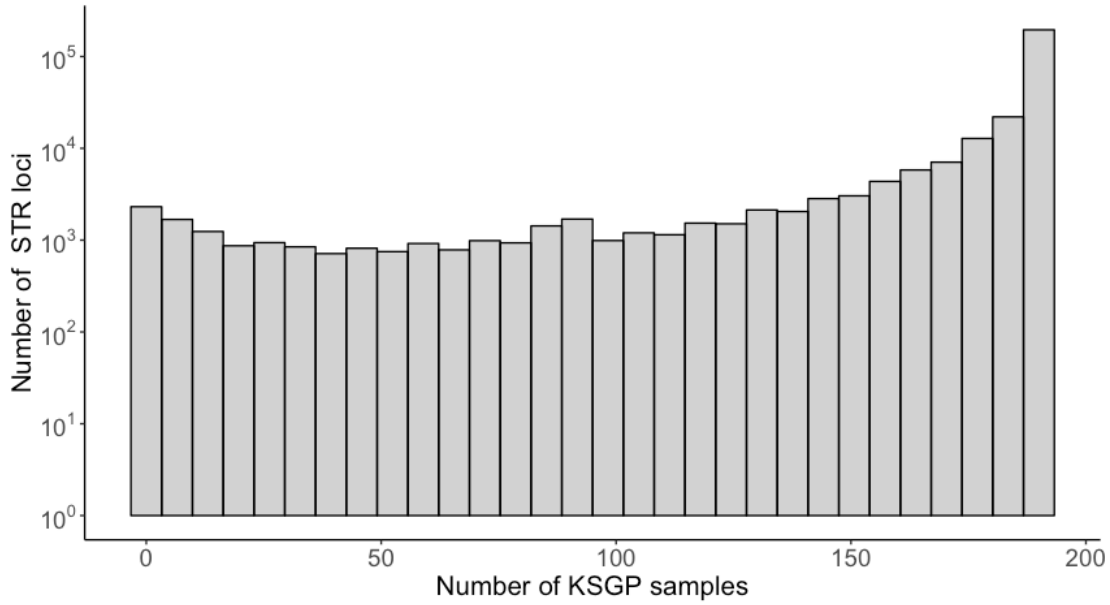
Loss of function (LoF) in our STR catalogue was identified as frameshifting nonreference variants in coding exons, each of which had common alleles found in at least 10 individuals. The annotation of the protein-coding regions was performed using ANNOVAR with hg38 databases[14].

***STR quality metrics*:**

For the raw multi-sample STR dataset generated above, a total of 280,970 loci (91.5%) of the STR reference panel were genotyped. On average, 196.8 individuals per locus (Supplementary Fig. 12) and 246,906 STR loci per individual (range 170,819-264,719; Supplementary Fig. 13) were characterized, with 10.9 informative reads (range 2-14.2; reads that span the STR loci entirely). For each sample, STR calls with coverage greater than 5X were shown in Supplementary Data 5, with mean values of 179,916 for autosomes, 9,058 for chromosome X and 806 for chromosome Y. Among 176 genomes sequenced in our lab (excluding CGI-based whole genomes), 281,187 STR loci were called; only 1,286 loci were greater than mean read length (150 bp). 26,179 STR loci of the reference panel were no-call; 15,015 were more than 100 bp, with half of the reference (13,071) greater than the mean read length. All the samples had call rates within three times the standard deviation of their mean, but three (MD2, CV4 and DIV95) were outliers, so they were removed from further analyses. The mean quality scores of STR calling among the samples aligned with the calling rates. Higher and better scores were observed in shorter reference lengths between dinucleotide and hexanucleotide (Supplementary Fig. 2a). Therefore, most undetected loci could not be physically spanned by a single read of the current sequencing length[2,25], and STRs with reference alleles smaller than 105 bp seem to be a reliable range in our study (Supplementary Fig. 2b).



Supplementary Fig. 12 **Distribution of number of samples with different STR calls (chromosomes 1-22, X and Y) observed in this study.** STR calls for African genomes from SGDP are excluded due to different read lengths.



Supplementary Fig. 13 **Distribution of the number of called samples within this study.** STR calls for African genomes from SGDP are excluded due to different read lengths. KSGP, Khoe-San Genome Project.

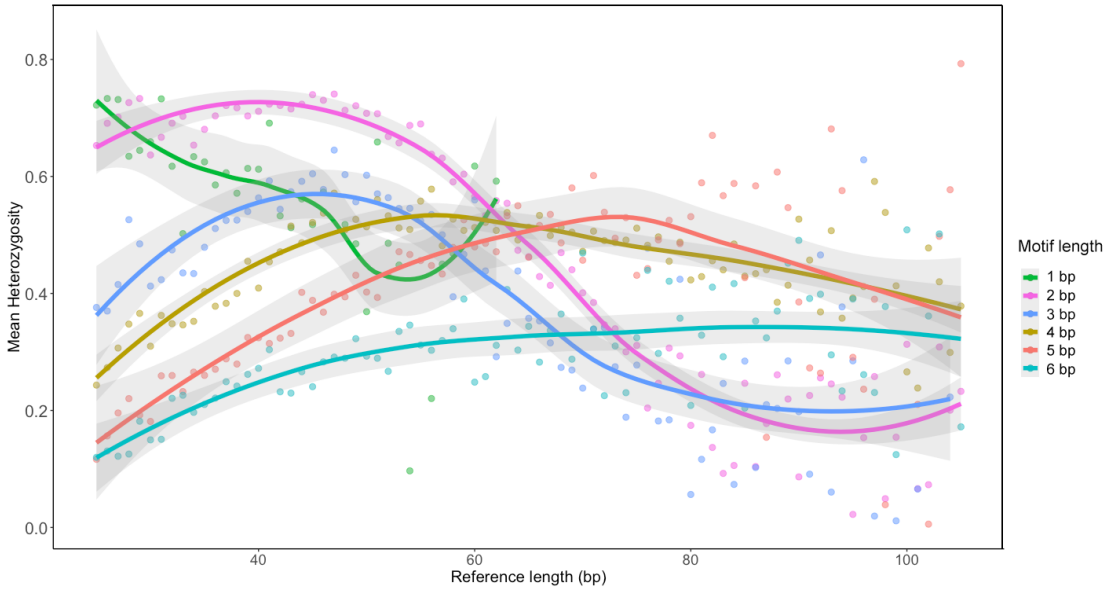
Patterns of STR variation:

After filtering the STRs described above, mean heterozygosity of these STR loci showed autosomes to be the most diverse, with the least in sex chromosomes, especially chromosome Y (Supplementary Table 3). Using autosomal STRs, their heterozygosity differed dramatically when stratifying by motif length, with less variation observed in a longer motif length (Supplementary Fig. 14). The mean heterozygosity increased in STRs with reference length from 25 bp to 40 bp, consistent with a positive correlation observed in previous publications[2,25]. However, the heterozygosity of 1-bp repeats declined in that range, and 2-bp repeats continued to considerably drop their variation after the motif length at 50 bp. The decline of STR variation

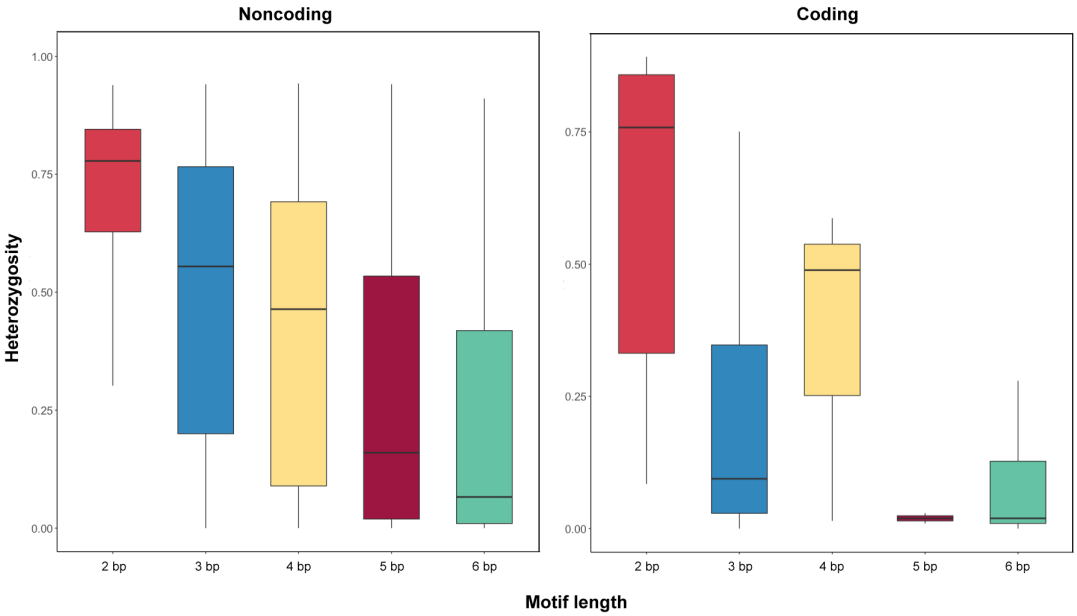
by motif length was mainly found within noncoding regions, and coding STRs were not as diverse as noncoding ones, with the most polymorphic motif length at two bp (Supplementary Fig. 15). In addition to sequence determinants, 60.3% of the STRs showed more than two common alleles, with 44 loci containing more than ten common alleles (Supplementary Fig. 16a). The variability of the common STR alleles was altered by different motif lengths (Supplementary Fig. 16b). Almost none of the homopolymers and dinucleotides were fixed with a single non-reference allele, while 3-5% of pentanucleotides and hexanucleotides were monotonically fixed with single non-reference alleles, indicating a critical role of multiallelic characteristics in STR variability.

Supplementary Table 3 Mean heterozygosity of STR loci by chromosome.

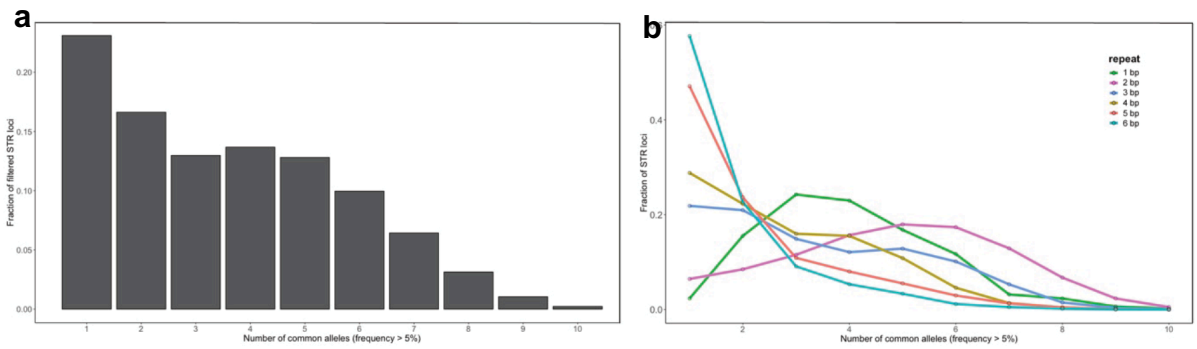
Region	Mean heterozygosity
chr1-22	0.505
chrX	0.4
chrY	0.184



Supplementary Fig. 14 **Mean heterozygosity as a function of reference length (25-105 bp) of autosomal STRs.**



Supplementary Fig. 15 **Coding characteristics affecting variation in autosomal STRs with different motif lengths.** Unrelated genomes without outliers are included in this analysis. The plot excludes 1-bp motif size due to none of the coding STRs detected. Box plots show lower (25%) to upper (75%) quartiles of per-site heterozygosity distribution. Interior lines indicate the median.



Supplementary Fig. 16 **Allele frequency spectra of autosomal STRs common in this study. a** The number of common alleles per locus. **b** Common alleles per locus stratified by motif length (1-6 bp).

***Potential loss-of-function STR variants*:**

Eight microsatellite alleles were identified to have potential common frameshifts and stop gain (Supplementary Table 4). *GID4* and *CNTNAP1* frameshifts were commonly observed in study forager populations. *LFNG* insertions, *MUC2* insertions, *FGFRL1* deletions and *DCHS2* deletions, were found across African and non-African populations. A similar loss of function within the *LFNG* and *DCHS2* genes have been previously reported[2,25].

Supplementary Table 4 Common LoF alleles (hg38 reference)

Chr	STR start[a]	STR stop	Base pair from hg38	Motif	Referenc length (bp)	Refseq function	No. Ind	Exon start[a]	Exon stop	Gene
chr4	1,025,266	1,025,305	-2	CA	39	frameshift deletion	14	1,025,303	1,025,305	FGFRL1
chr17	18,039,405	18,039,471	2	GT	66	frameshift insertion	26	18,039,470	18,039,471	GID4
chr4	154,323,249	154,323,280	-4	TTTG	31	frameshift deletion	138	154,323,276	154,323,280	DCHS2
chr7	2,513,247	2,513,275	4	GATG	28	stop gain	100	2,513,274	2,513,275	LFNG
chr7	2,513,247	2,513,275	-4	GATG	28	frameshift deletion	55	2,513,271	2,513,275	LFNG
chr17	42,698,811	42,698,855	1	CCAGCC	44	frameshift insertion	13	42,698,854	42,698,855	CNTNAP1
chr10	73,168,308	73,168,340	-3	GAG	32	frameshift deletion	20	73,168,337	73,168,340	FAM149B1
chr11	1,092,987	1,093,031	1	CCA	44	frameshift insertion	10	1,093,030	1,093,031	MUC2

[a] 0-based positions

5.2. Mobile element discovery

***Mobile element call generation*:**

Mobile element insertions (MEIs) were detected across 198 unrelated genomes using the Mobile Element Locator Tool (MELT v2.1.5)[26], where discordant read pairs define potential MEI sites, and split reads identify breakpoints and target site duplications (TSDs). Recalibrated read alignment for each genome generated in Section 2.3 was first run using MELT-Single, with mean coverage per sample reported in Supplementary Data 1. The reads were aligned to the three following mobile element reference sequences available in hg38: 281 bp *Alu*; 6,019 bp LINE1; and 1,316 bp SVA. In addition to the MEI call set per sample, a joint-call MEI dataset of all the genomes was created using MELT-Split. Subsequent discovery information from MELT-Single across all the genomes was merged to determine accurate breakpoints, genotyped and filtered for the generation of a final joint-call variant call set. Filtering of candidate MEIs was based on 5'- and 3'- supporting evidence of discordant pairs at both sides of the breakpoint, total number of split reads, total percentage of no-call genotypes (i.e. './.') greater than 25%, and erroneous calls within a low complexity region[26]. Only passed sites were included for downstream analyses, and MEI annotation was performed by intersecting their positions with the RefSeq hg38 database. For *Alu* or LINE1 elements detected, their core sequences were assembled to classify them into subfamilies based on diagnostic mutations shared by subfamily members[26]. The *Alu* subfamilies were annotated and visualized using Krona plots[27]. The Welch two-sample t-test was computed in R[15] for the different numbers of mobile element insertions between two populations.

***Population genetic analyses*:**

To measure the diversity of each MEI type (*Alu*, LINE1 and SVA), per-site heterozygosity described above (Equation 2) and its mean across all autosomal loci were computed using passed MEI datasets of unrelated genomes. To explore the relationships of different populations explored in this study, we analyzed genotypes of each MEI site, following the eigenanalysis of indels described in Equation 1 for data normalization[17]. All autosomal MEI sites were concatenated, duplicates and missing genotypes were removed, and the sites with five kb apart were included for the normalization step. A total of 14,085 MEIs were normalized and then used to estimate the principal components using R packages FactoMineR and factoextra[15].

***MEI characterization*:**

A total of 14,789 MEI insertions, consisting of 12,245 *Alu*, 1,772 *LINE1*, and 772 *SVA* (*SINE*/ *VNTR*/ *Alu* composite element), were identified among our study genomes. On average, their variant length reported was 264 bp (range, 16-281), 2982 bp (21-6,019) and 1,038 bp (73-1,316) for *Alu*, LINE1 and SVA, respectively. The accuracy of this MEI discovery could be assessed based on a scale of breakpoint resolution from zero to five, with zero being the least accurate and score 5 being the most accurate[26]. For this study, 96.26% of *Alu* sites, 87.02% of LINE1 sites, and 60.49% of SVA sites fell into score 5 (Supplementary Fig. 2c).

***Identification of Alu and LINE1 subfamilies*:**

Krona plots identified three main families of *Alu* elements (*Alu*J, *Alu*S and *Alu*Y); there were the single *Alu*J (Jo), five *Alu*S (Sc, Sg, Sp, Sq and Sz) and 11 *Alu*Y subfamilies (Ya, Yb, Yc,

Yd, Ye, Yf, Yg, Yh, Yi, Yj and Yk; Supplementary Fig. 2d). The active *Alu*Ya5 and *Alu*Yb8 elements were the most abundant in the *Alu*Y family[28], with 2,822 and 1,923 sites detected out of 12,245. The ancient *Alu*Jo and *Alu*S elements aged ~65 and ~30 million years old, respectively[28,29] were less variable. The *Alu*Jo was inactive, but some *Alu*S lineages contained functional *Alu* core elements allowing transposition of MEIs in a species-specific manner[30]. The youngest LINE1 elements, L1-Ta and L1-Ta1d subfamilies, accounted for 62.5% of the total number of LINE1 elements in our study (1,106/1,772) and are known to be human-specific[31,32].

***Genic and coding MEIs*:**

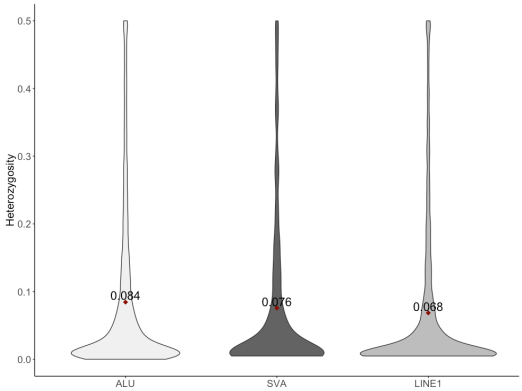
Most MEI elements detected among study genomes spanned outside coding exons (Supplementary Table 5). There were 934 *Alu*, 113 LINE1 and 123 SVA sites lying within important noncoding regions, promoters and terminators. However, two SVA elements were inserted into exonic regions of *MUC17* and *ZNF404* genes. Only modern *Alu* elements, five *Alu*Ya and a single *Alu*Yb, were found within coding exons of the six following genes: *ERICH6*, *OVCH2*, *OR6C76*, *PRAMEF4*, *SLC36A4* and *ZNF880*.

Supplementary Table 5 Gene annotation of MEIs

Gene feature	*Alu*	LINE1	SVA
Exon	6	0	2
Promoter	455	59	58
Terminator	479	54	65
Intron	5,174	668	343
3'-UTR	92	7	7
5'-UTR	17	0	1

***Population genetics analysis of MEIs*:**

Among 198 genomes analyzed, the mean heterozygosity of *Alu* elements was the highest at 0.084 (Supplementary Fig. 17). The number of *Alu* sites was 7-16 times larger than that of LINE1 and SVA. The SVA and LINE1 had lower heterozygosity at 0.076 and 0.068, respectively.



Supplementary Fig. 17 **Mean heterozygosity of *Alu*, SVA and LINE1 elements identified in this study.**

SI6 – Population structure analysis

***Variant data generation*:**

Our estimation of ancestral clusters within this study was performed using biallelic SNV consensus data after a genotype refinement step in Section 2.4. This dataset had 197 Khoe-San Genome Project and SGDP unrelated genomes representing target populations, with additional seven genomes using CGI sequencing (UL2, UL5, UL7, UL8, UL9, UL12 and ABT). Reference populations for the population structure analysis (representing possible admixture sources) consisted of 179 representatives from the 1,000 Genomes Phase 3[33]. The references were defined as Asian (CHB; n = 30), European (CEU; n = 30), African American (ASW; n = 30), East Bantu (LWK; n = 29), Proto Bantu (YRI; n = 30), and West African (GWD; n = 30). To merge all the genomes, they were lifted to the GRCh37 reference using picard v2.7.1 (https://broadinstitute.github.io/picard/) and merged using bcftools v1.3.1. The merged call set was filtered out using PLINK 2.00[16] when the variant's missing rate was greater than 10%, MAF less than 5% (singleton variants uninformative for population clustering), and *P*-values for HWE failed at 0.0001. The subsequent dataset of 7,003,530 remaining variants and 383 individuals was used for data analysis of a full SNV set.

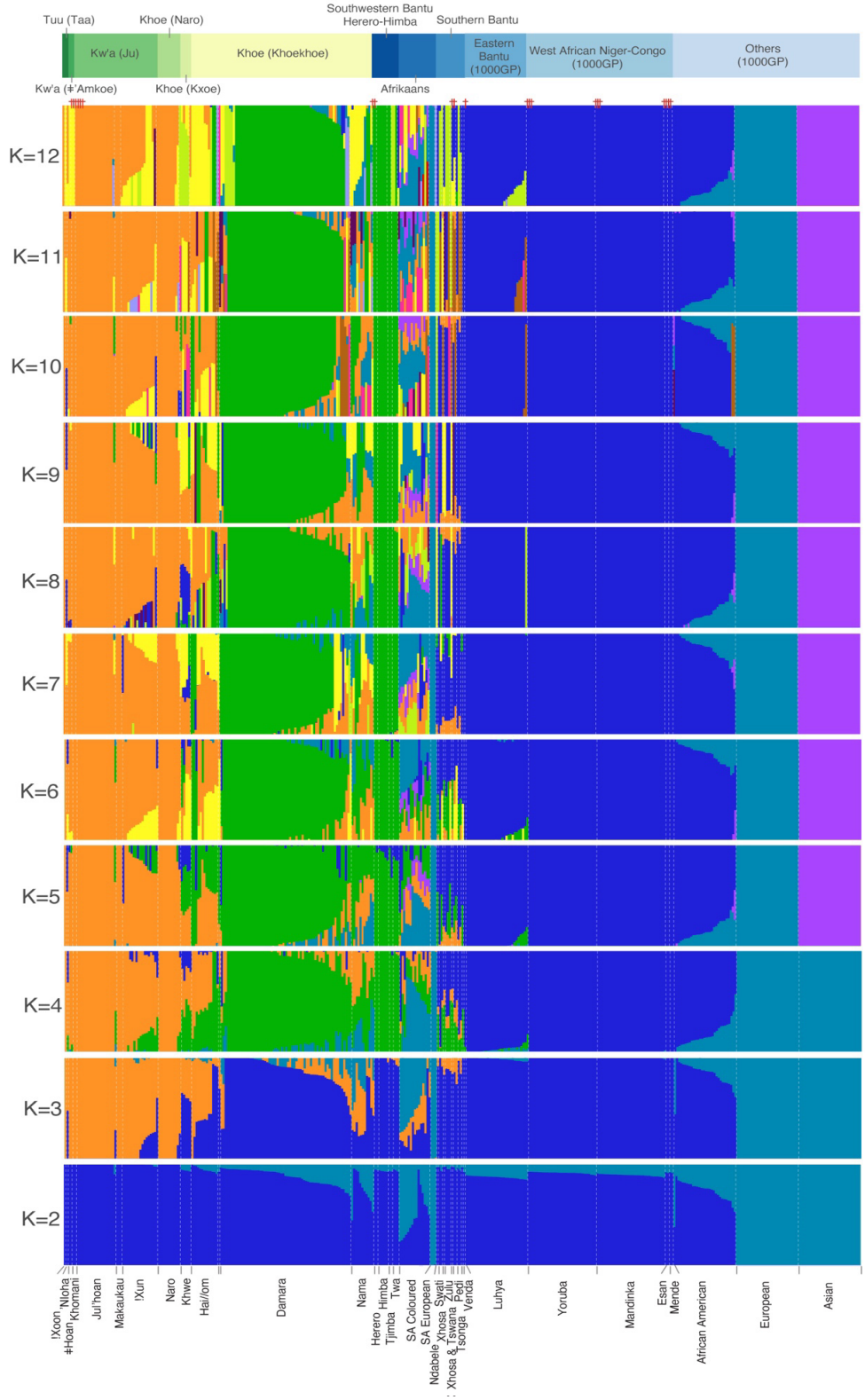
For a linkage disequilibrium (LD) pruned dataset, a further selection of autosomal markers was based on a linkage disequilibrium r^2 value > 0.2 within a 50-variant sliding window, advanced by five variants at a time. We assigned all the SNVs in the dataset with a unique identity to avoid confusion during the LD pruning step in the PLINK analysis. Given expensive computation during some downstream analyses, the pruned dataset was thinned five kb apart, with a total of 74,943 SNVs remaining.

***Population genetic analyses*:**

Ancestral clusters in this study were estimated using fastSTRUCTURE v1.0[34]. The fastSTRUCTURE program uses variational Bayesian inference for the best approximation of the marginal likelihood of a large SNV dataset. We analysed the full SNV dataset described above using fastSTRUCTURE, with randomly chosen initial seeds, varying the number of ancestral populations ranging from K = 2 to 8, and a minimum of five runs conducted per ancestral estimation. The logistic prior model was preferred, providing higher marginal likelihood values than a simple model. Our smaller LD-pruned dataset was also conducted for comparisons. The pruned data were run with random seeds and varying number of ancestral populations (range K = 2-12; Supplementary Fig. 18). Structure plots were visualized using Pophelper v2.2.7 in R[15]. The python script, chooseK.py predicted the number of ancestral populations that maximises the marginal likelihood of the fastSTRUCTURE data and the minimum number of populations with a cumulative ancestry contribution of at least 99.99%[34].

To test each variant type for the identification of three African populations studied herein ('San', 'Damara' and West & South African Niger-Congo; n = 107; Supplementary Fig. 6a), variants with the 10% most heterozygous autosomal loci (per-site heterozygosity; Equation 2) were extracted for the structure analysis using STRUCTURE v2.3.4[35], with 5,000 burn-in iterations followed by 10,000 iterations. A minimum of five runs were conducted for ancestral estimation (K) equal to three representing our three broad identifiers. Filtered and pruned SNVs and biallelic indels described in Section 4.1 were thinned five and 25 kb apart, respectively,

due to a computation burden of the tool used. Filtered STRs and passed MEIs described in Sections 5.1 and 5.2, respectively, were thinned five kb apart.



Supplementary Fig. 18 **STRUCTURE analysis of linkage disequilibrium pruned SNV data with the logistic prior model.** The analysis was run with five replicates with randomly chosen initial seeds, varying the number of ancestral populations from K = 2 to 12. More detailed methods can be found in Supplementary Information 6.

We found that log likelihood monotonically increased with K. The model components used to explain structure in the data are K = 10, and model complexity at K = 12 maximizes marginal likelihood. SA, South African.

SI7 – Phylogenomic reconstruction

7.1. Full SNV phylogeny

***Variant and sequence data preparation*:**

The biallelic SNV consensus data of unrelated individuals generated after a genotype refinement step in Section 2.4 were merged with the Altai Neanderthal variant calls downloaded from http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/. Autosomal loci without missing genotypes were included and strictly contained both homozygous reference and alternate alleles to apply an ascertainment bias correction to the phylogenetic likelihood calculation described below. The variant data were converted to multiple sequence alignment of variable sites among 199 genomes using the vcf-tab-to-fasta script (http://code.google.com/p/vcf-tab-to-fasta).

Maximum Likelihood phylogeny:

To construct a full SNV tree, we converted the multiple sequence alignment of 8,491,748 SNVs created above to PHYLIP format and ran RAxML8[36], using an Altai Neanderthal as an outgroup. The sequence alignment contained 7,054,101 patterns, and base frequencies ranged from 0.217 to 0.283. The ASC_GTRGAMMA model was used with Lewis' ascertainment bias correction, as the data analysed had no constant sites and needed to be corrected. We sought for a best-scoring Maximum Likelihood (ML) tree out of 15, and then conducted 30 bootstraps due to our extensive alignment patterns solely computed in a single partition. A final phylogeny with branch support was generated by reconciling the best ML tree with bootstrap searches.

7.2. Direct mapping consensus phylogeny

Phylogenomic inference involves the analysis of large portions of genomes to understand evolutionary relationships between samples. Direct mapping consensus methods have recently gained much preference against the phylogenomic analysis of *de novo* genome assembly due to better performance and costly and time-consuming computation of the WGS assembly[37,38]. Hence, we will generate the direct mapping consensus phylogeny in this section. Moreover, we did perform the phased allele phylogenetic analysis for multispecies coalescence mentioned in Andermann et al[37]. Results from the analysis were difficult to interpret due to human evolution that just began to develop ~200 kya (see Section 8.1), and phased sequences from the same individual might not cluster with each other due to genetic admixture (data not shown).

***Sequence data processing*:**

The recalibrated alignment data generated in Section 2.3 for 33 study genomes and the Altai Neanderthal (Supplementary Table 17) were used to construct a whole genome diploid consensus sequence per genome (only autosomes), using samtools mpileup v1.6 and bcftools v1.6.

./samtools mpileup -C50 -uf Homo_sapiens_assembly38.fasta $BAM | bcftools call -c - | vcfutils.pl vcf2fq -d $MIN -D $MAX > $FASTQ

Supplementary Table 6 High-pass whole genome data used for phylogenomic reconstruction

ID	Run ID	Reference	Sex	Population	Region	Mean coverage[a]
Altai_Neanderthal	NA	Prüfer et al. 2014	Female	Neanderthal	Siberia	50.37
S_BantuHerero-1	ERR1347714	Mallick et al, 2016	Male	BantuHerero	Africa	43.05
S_BantuHerero-2	ERR1019043	Mallick et al, 2016	Male	BantuHerero	Africa	39.45
S_Brahmin-1	ERR1395559	Mallick et al, 2016	Male	Brahmin	South Asia	46.84
S_Han-1	ERR1019055	Mallick et al, 2016	Female	Han	East Asia	65.76
S_Khomani_San-1	ERR1395588	Mallick et al, 2016	Female	Khomani_San	Africa	42.17
S_Khomani_San-2	ERR1395572	Mallick et al, 2016	Female	Khomani_San	Africa	46.58
S_Luhya-2	ERR1347655	Mallick et al, 2016	Male	Luhya	Africa	32.96
S_Yoruba-1	ERR1025657	Mallick et al, 2016	Female	Yoruba	Africa	33.15
NB4	NA	This study	Male	Ju/'hoan	Africa	46.6
NB24	NA	This study	Female	Ju/'hoan	Africa	45.76
KB5	NA	This study	Female	Naro	Africa	40.60
MN6	NA	This study	Female	Naro	Africa	41.24
MN39	NA	This study	Male	!Xoon	Africa	41.04
KB2	NA	This study	Female	ǂHoan	Africa	40.67
MN38	NA	This study	Female	ǂHoan	Africa	40.61
NF03	NA	This study	Female	!Xun	Africa	40.03
TS3	NA	This study	Female	!Xun	Africa	41.58
MN17	NA	This study	Male	Makaukau	Africa	42.09
MN18	NA	This study	Male	Makaukau	Africa	41.37
WB14	NA	This study	Male	!Oeǂgân	Africa	40.70
WB204	NA	This study	Female	!Oeǂgân	Africa	40.58
WB3	NA	This study	Male	ǂAodaman	Africa	41.61
WB207	NA	This study	Female	ǂAodaman	Africa	40.80
EP4	NA	This study	Male	Himba	Africa	42.55
KM4	NA	This study	Female	Himba	Africa	40.03
EP5	NA	This study	Male	Tjimba	Africa	41.65
EP6	NA	This study	Female	Tjimba	Africa	40.70
EP9	NA	This study	Male	Tau	Africa	40.54
EP13	NA	This study	Female	Tau	Africa	40.9
UL6	NA	This study	Male	Niger-Congo	Africa	40.31
UP2113	NA	This study	Male	Niger-Congo	Africa	57.82
KB3	NA	This study	Male	European	Africa	36.17
LAB12	NA	This study	Male	European	Africa	40.50

***Extraction of protein coding regions and sequence alignment*:**

We extracted individual consensus sequences for exon regions of 18,680 protein-coding genes well-known in the HUGO standardized database (Supplementary Table 9; https://www.genenames.org/cgi-bin/statistics). We chose the longest transcript sequence for each gene collected from the Consensus Coding Sequence (CCDS; hg38 release 20)[39] to align with the consensus sequences using the LASTZ program[40].

*./lastz_32 $FASTA[multiple] $CCDSv20.fasta --filter=identity:90 --ambiguous=n *
--filter=coverage:90 --ambiguous=iupac --format=
general:nmismatch,name1,start1,end1,name2,start2,end2,strand1,strand2,size1,size2,cigar,identity,score --
notransition --step=20 --progress=1000 > $OUT.lastz

Sample-specific exon sequences were extracted from the consensus sequences using 'bedtools getfasta' command and genomic intervals identified by LASTZ[41] if sequences between the sample and CCDS reference had at least 90% identity and length coverage and the region aligning explicitly to a single CCDS exon that did not match somewhere else. Ambiguous bases extracted were replaced with 'N' bases, and a total of 137,937 common exons were retrieved

in the FASTA format across 34 study genomes. Among them, ten were removed due to different strand orientations in some of the genomes. For each remaining exon, accurate multiple sequence alignment was performed using the MAFFT program[42], with a local alignment option and the maximum number of iterative refinement equal to 1,000. Any gaps in the alignment were removed for phylogenomic analysis. Using the MEGA-CC analysis[43], the maximum nucleotide diversity observed among the alignment of 137,927 exons was 0.026 with a standard error of 0.009 (p-distance; 500 bootstraps).

***Phylogenomic analysis*:**

A large-scale Maximum Likelihood (ML) tree of 16,083 genes (137,927 exons; 26,039,475 bp) from 34 genomes used in this study was reconstructed using RAxML8[36]. We built the concatenated data matrix of all the above-mentioned alignments and converted it to PHYLIP format. The unpartitioned sequence alignment contained 59,371 patterns and base frequencies ranging from 0.238 to 0.262. An initial search for the best tree was run with random seeds and 20 independent runs. The GTRGAMMA model was chosen as the best tree and provided a better score than the alternative model, GTRCAT (GAMMA-based Score, -36,885,820.43). To calculate branch support for the tree, 100,000 bootstrap replicates were conducted with the same DNA substitution model. The RogueNaRok program was utilized for pruning rogue taxa that rendered ambiguous or insufficient phylogenetic signals among the 100,000 bootstrapped trees[44]. The eight following genomes were removed: KB2, KB5, KM4, MN38, WB14, S_BantuHerero-2, S_Luhya-2, and S_Yoruba-1. An extended majority rule consensus phylogeny with final bootstrap support was generated by reconciling the best ML tree with 100,000 bootstrap searches, using Altai Neanderthal as an outgroup.

SI8 – PSMC and MSMC analysis

8.1. PSMC

We used the Pairwise Sequentially Markovian Coalescent (PSMC) model to infer population sizes for selected populations. Given free parameters, including the scaled mutation rate, the recombination rate and piecewise constant ancestral population sizes, the model could infer population size history using a single diploid genome per population[45] that was scaled for the time and rate with a generation time of 30 years and a mutation rate of 1.25×10^{-8} mutations per nucleotide per generation[46]. The diploid consensus sequence of 22 autosomes from each individual was generated using samtools mpileup[47], where the sequence was transformed from the recalibrated alignment described in Section 2.3. Coverage settings, minimum and maximum read depth were adjusted based on mean coverage shown in Supplementary Table 1. The minimum scaled time was set to 10 kya (thousand years ago) because the model had a large variance if more recently than 20 kya[45]. We modelled the Altai Neanderthal genome[3] with the five following populations and representative genomes: *i*) Ju|'hoan (NB24); *ii*) 'Damara' (WB277); *iii*) West African Niger-Congo (S_Yoruba-2); *iv*) South African European (KB3); and *v*) Asian (S_Han-1). The alignment of the Neanderthal genome (chr1-22) was downloaded from http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/.

Results

The model inferred that the population ancestral to all Africans and non-Africans began to differentiate at least 200 thousand years ago (kya; *Ne* ± SD = 26,315 ± 2,123 for all populations). Assuming the same mutation rate and generation time, the model inferred the Altai Neanderthal ancestral population to have a 7.3- to 8.9-fold lower effective population size than modern humans around 200 kya and continuously decline during the early development of present-day human substructure. Note that our date estimates might be varied from previous publications using different point estimates of the human mutation rate and generation time.

8.2. MSMC2

***Phased haplotypes of multiple genome sequences*:**

The time estimate of population separation using the Multiple Sequentially Markovian Coalescent (MSMC) requires phased haplotypes[46]. The recalibrated alignment of each genome generated in Section 2.3 was used to create autosomal variant data and chromosomal masks of regions with proper coverage, following the program instruction (bamCaller.py script; https://github.com/stschiff/msmc-tools). The variant data in the hg38 reference were lifted to the GRCh37 reference (picard v2.20.0) prior to generating the required variant data and masks per autosome using bamCaller.py and its corresponding reference panel from the 1,000 Genomes Project Phase 3 (GRCh37). The mean coverage used in the analysis for each genome is shown in Supplementary Data 1 and Supplementary Table 1.

java -jar picard.jar LiftoverVcf I=$IN.hg38.vcf.gz O=$OUT.hg19.vcf.gz CHAIN=hg38ToHg19.over.chain REJECT=$OUT.rejected_variants.vcf R=ucsc.hg19.fasta

java -jar picard.jar LiftoverVcf I=$IN.hg19.vcf.gz O=$OUT.b37.vcf.gz CHAIN=hg19tob37.chain REJECT=$OUT.rejected_variants.vcf R=human_g1k_v37_decoy.fasta

*for I in {1..22}; do zcat $IN.b37.chr$i.vcf.gz | bamCaller.py *
*--legend_file 1000GP_Phase3/1000GP_Phase3_chr$i.legend.gz $Coverage *
$OUT.mask_chr$i.bed.gz | gzip -c > $OUT.chr$i.vcf.gz; done

For each genome, duplicate loci were removed from the variant dataset using bcftools norm (http://www.htslib.org/doc/bcftools.html). Variants were then phased with a reference panel of haplotypes from the 1,000 Genomes Project Phase 3, using SHAPEIT2[48]. The method is currently the most accurate for phasing sets of known genotypes, and using the reference panel improves phasing genomes closely related to populations in the reference panel. We followed the program instruction for phasing settings (MCMC iterations disabled). To perform the MSMC analysis, inputs with the specific MSMC format of the variant and mask data were generated using generate_multihetsep.py and global mappability masks. The mappability masks inform the list of autosomal regions (GRCh37), where short sequencing reads can be uniquely mapped[20]. Instead of the SHAPEIT2 reference panel, three !Xun genomes, NF10 (father), NF11 (mother) and NF05 (daughter) representing a 'San' trio were run in this step using trio information (--trio) to phase the four parental chromosomes.

*for i in {1..22}; do msmc-tools/generate_multihetsep.py --mask=$Sample1.mask_chr$i.bed.gz *
*--mask=$Sample2.mask_chr$i.bed.gz --mask=$Sample3.mask_chr$i.bed.gz *
*--mask=$Sample4.mask_chr$i.bed.gz --mask=msmc-tools/ref_masks/hs37d5_chr$i.mask.bed *
*$Sample1.chr$i.vcf.gz $Sample2.chr$i.vcf.gz $Sample3.chr$i.vcf.gz $Sample4.chr$i.vcf.gz *
> Multihetsep_chr$i.txt; done

***MSMC analysis*:**

We used MSMC2, an updated program published in Schiffels and Durbin[46], for the time course of population separation in Khoe-San. The MSMC2 runs a separate Hidden Markov Model (HMM) across every possible pair of phased haplotypes of multiple individuals, so it estimates pairwise times to a common ancestor, emphasizing the first coalescence event. In this analysis, we used four haplotypes from two diploid individuals per population in each run if applicable (Supplementary Table 18). We skipped ambiguous or unphased sites present within any genomes analyzed because they could generate severe artefacts in recent time intervals[46]. Split time estimates were combined between within- and across-population runs, following the program instruction (combineCrossCoal.py script; https://github.com/stschiff/msmc-tools). All MSMC coalescent results were scaled using a mutation rate of 1.25×10^{-8} per base pair per generation and a generation time of 30 years and plotted using R[15].

Supplementary Table 7 Samples used in MSMC analyses

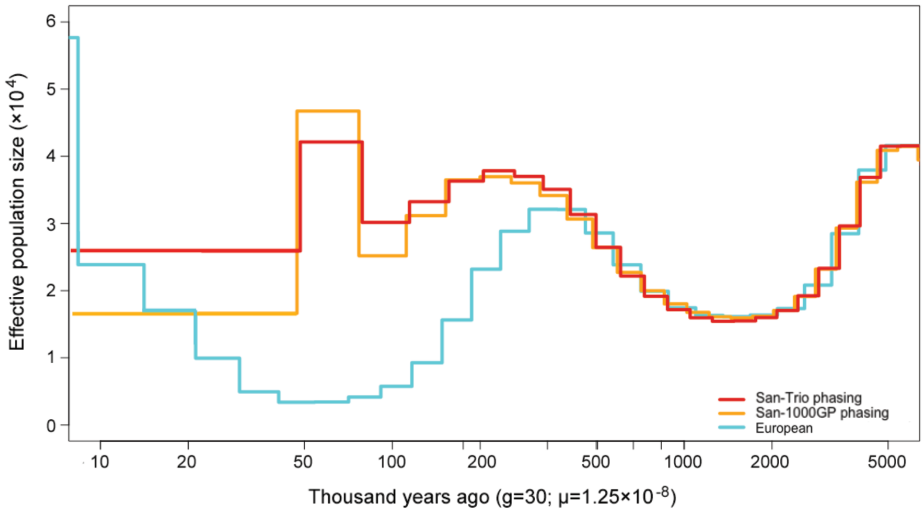
Population ID	Sample ID	Reference
Neanderthal	Altai Neanderthal	Prüfer et al. 2014
San	NB17 & NB24	This study
Damara	WB277 & DIV25	This study
Niger-Congo	S_Yoruba-1 & S_Yoruba-2	Mallick et al, 2016
Asian	S_Han-1 & S_Brahmin-1	Mallick et al, 2016
South African European	KB3 & UP2153	This study
Ju|'hoansi	NB17 & NB24	This study
Naro	KB5 & MN6	This study
Khomani	S_Khomani_San-1 & S_Khomani_San-2	Mallick et al, 2016
!Xoon	MN39	This study
ǂHoan	KB2 & MN38	This study
!Xun	NF03 & TS3	This study
Makaukau	MN17 & MN18	This study
Damara (!Oeǂgân)	WB14 & WB204	This study
Damara (ǂAodaman)	WB3 & WB207	This study
Himba	EP4 & KM4	This study
Tjimba	EP5 & EP6	This study
Twa	EP9 & EP13	This study
Herero	S_BantuHerero-1 & S_BantuHerero-2	Mallick et al, 2016
West African Niger-Congo	S_Yoruba-1 & S_Yoruba-2	Mallick et al, 2016

Note – MSMC2 limited to eight haploid sequences due to computational complexity[46]

Results

Divergence timing within study populations:

In this study, MSMC estimates of effective population size for a 'San' !Xun individual were vastly different around 50 kya between trio-phased and reference-phased data (*Ne* = 25,906 and 16,616, respectively), but the same for a mild bottleneck in the population history compared to Europeans (Supplementary Fig. 19). 'San' and 'Damara' genomes unavailable in the 1,000 Genomes Project panel for phasing leads to an underestimate of the population size in recent time[2]. Therefore, the estimation of the PSMC model is preferred for the size estimate. The MSMC would use for cross-coalescence rates between populations.



Supplementary Fig. 19 **Impact of phasing on effective population size estimates using MSMC2.** The 'San' !Xun genome, NF05 was phased using either the 1000 Genomes Project reference panel or trio information (NF10 as a father and NF11 as a mother; Supplementary Fig. 1a).

SI9 – Assembly-based genome analysis

Purpose

The current analysis pipeline for handling and processing WGS data in humans, known as the 'Genome Analysis Toolkit (GATK)', is designed to process sequencing reads (FASTQ) and produce variant calls for use in downstream analyses[5]. However, some of the analyses, namely codon-based selection tests performed in the current study, cannot handle the variant data, and contiguous sequences are needed. Here, we have developed a pipeline for the assembly-based Genome Analysis Toolkit (aGATK; Supplementary Fig. 8a) involving pre-processing raw sequence data at the individual level and then producing cohort-level multiple sequence alignment (MSA) for the downstream analyses.

Main steps

We begin by generating an individual's genome draft and extracting sample-specific target sequences. Next, we generate cohort-wise multiple sequence alignment (MSA) per gene and perform subsequent local and amino acid-guided alignment for accurate estimation of sequence quality prior to use in positive selection tests. Assembly-based genome analysis workflow consists of the following steps:

Perform de novo genome assembly:

For each genome, filtered FASTQ reads generated in Section 2.3 were assembled with pair information using ABySS v2.0.2[49]. Default paired-end sequencing parameters for human genome assemblies were set with k-mer equal to 96 and five pairs minimally required to construct a contig. Subsequent assembled scaffolds in a FASTA format were chosen if larger than one kb.

Locally align to sample-specific references:

The sequence assembly per sample is treated as a reference genome. For each individual, nucleotide sequences of targeted regions of interest, consensus coding sequences (CCDS; release 20) described in Supplementary Information 9 are locally aligned in a BAM format against the sample-specific reference using the BWA-MEM algorithm, fast and accurate local alignment[6]. The BAM alignment is then sorted and indexed.

```
./bwa mem -t 16 $REFERENCE CCDSv20.GRCh38.exon.nt.ALL.fasta \
-R '@RG\tID:CCDS.GRCh38.exon.nt\tPL:None\tPU:None\tLB:1\tSM:CCDSv20\tCN:hcpcg' | ./samtools view -h
-S -b - > $REFERENCE.CCDSv20.GRCh38.exon.bam

./samtools sort -o $REFERENCE.CCDSv20.GRCh38.exon.sort.bam \
$REFERENCE.CCDSv20.GRCh38.exon.bam

./samtools index $REFERENCE.CCDSv20.GRCh38.exon.sort.bam
```

Extract sample-specific target sequences:

At the individual level, we extract genomic coordinates and their subsequent nucleotide sequences from the BAM alignment between genome assembly and exon database using bedtools v2.26.0[41]. Targeted exon sequences aligned with each genome assembly are removed if soft-clipped.

./bedtools bamtobed -i $BAM -cigar > $BED

./bedtools getfasta -fi $REFERENCE -bed $BED -name -tab -s -fo $OUT.tab

Concatenate target sequences into a gene level:

For each individual, target sequences (exons) are ordered based on exon numbers and their 5' direction and concatenated into a gene level for 18,680 genes targeted per sample, depending on an individual's genome assembly. Multiple FASTA sequences, each of which corresponds to a single gene, are generated per sample.

Cohort-wise multiple sequence alignment per gene:

This step consists of consolidating each gene sequence across multiple samples to generate per-gene MSA within a cohort. Using MAFFT v7.402[50], this MSA per gene is locally aligned with a corresponding reference sequence from the CCDS database for quality control and protein translation purposes in the cohort analysis in the following steps. The script '*aGATK_mafft-0.1dev.py*' written in Python programming languages can be run per gene in this step as follows:

python aGATK_mafft-0.1dev.py --fasta $MSA --output $OUT

Multiple sequence alignment and orientation quality controls:

The next step is to quality check per-gene MSA datasets based on overall nucleotide diversity (mean p-distance) and sequence orientation relative to their hg38 reference sequences, using MEGA-CC v7.0.14[43] and LASTZ v1.04 programs[40], respectively. There are the two following Python scripts used for these analyses:

python aGATK_megacc-0.1dev.py --target $MSA --output $OUT

python aGATK_lastz-0.1dev.py --target $MSA --query $MSA --output $OUT --identity 90

For each gene of interest, this step could be followed by strand correction, removal of gaps and sequence alignment of genes less than 100 bp. After repeating the above-mentioned local alignment step, we chose only genes with > 90% sequence identity among individuals studied for further analyses.

Perform amino acid-guided alignment:

To facilitate amino acid-based selection tests explained in Supplementary Information 10, amino acid-guided alignment of each per-gene MSA is conducted using the transAlign program[51]. The step would provide accurate MSA of intact open reading frames within a cohort relative to reference coding sequences, and sequence alignment based on amino acids is often superior to that obtained directly from nucleotides. Overall nucleotide diversity is recalculated, with the MSA greater than 90% identity retained.

perl transAlign.pl –d$MSA -if -v -pclustalw2 -ra

Preparing alignment data for selection analysis:

Formatting data for selection tests generally involves slight differences from the steps described above. This consists of (*i*) removing reference coding sequences (CCDS release 20) and (*ii*) replacing ambiguous bases (IUPAC codes) and stop codons with gaps.

Test for positive selection:

This step is mentioned in detail in Supplementary Information 10 below.

Caveats

We perform *de novo* genome assembly as an alternative to short-read mapping implemented in the classical GATK pipeline or as a reference genome-free analysis of short-read sequencing. Despite WGS coverage of ~41X assembled in this study, target sequences of interest in some individuals are challenging to obtain, for example, 745 genes partly observed among 37 genomes analyzed in Supplementary Information 10.

SI10 – Positive selection inference

Diversifying or positive selection plays an important role in shaping human diversity to maintain advantageous variants in the gene pool of a population and, therefore, increase the population's relative fitness. As a result, there are elevated levels of polymorphism relative to neutral evolution and higher rates of nonsynonymous (*dN*) to synonymous (*dS*) nucleotide substitutions than expected under neutral evolution[52].

***Codon-based selection test*:**

Following Supplementary Fig. 8a, to test for the incidence of selection at specific amino acids, codon-based selection tests for 37 whole genomes randomly chosen to represent different populations in our project (Supplementary Data 9) were performed using omegaMap version 0.5[53]. This Bayesian analysis assesses an excess of nonsynonymous relative to synonymous substitution rates among nucleotide sequences of coding regions ($dN/dS > 1$ or $\omega > 1$). The list of 18,680 protein-coding genes well-known in the HUGO standardized database (Supplementary Table 9; https://www.genenames.org/cgi-bin/statistics) was retrieved for testing[54]. Consistently annotated and high-quality protein-coding sequences of the genes within the hg38 assembly were collected from the Consensus Coding Sequence (CCDS; release 20) database[39]. The longest coding transcripts for each gene were chosen. Among those genes, 15,155 well-known genes were assembled across the 37 genomes, and 270 more were observed from 19 to 36 genomes. The multiple sequence alignment of the 15,425 genes from our study genomes was chosen based on their nucleotide diversity of less than 10%.

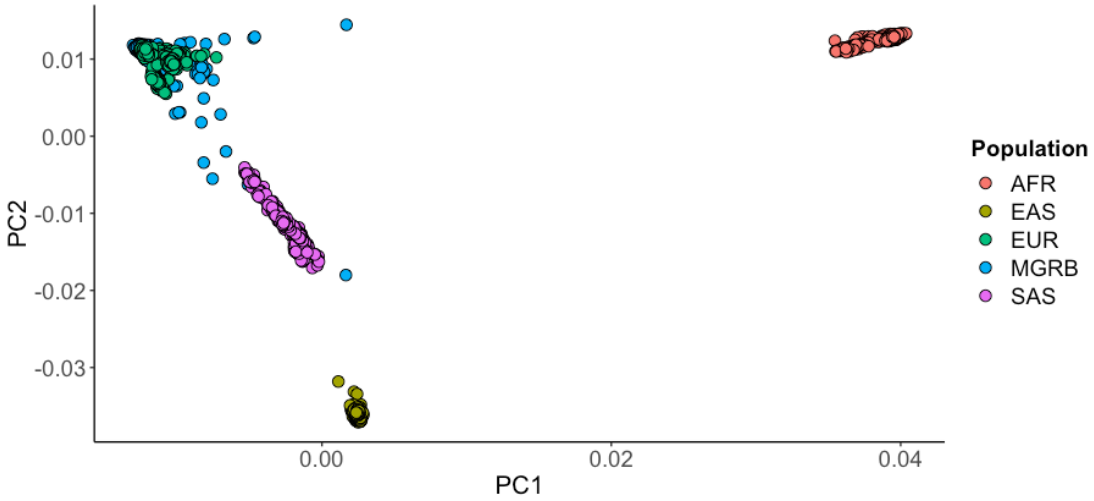
For test convergence, two runs of selection testing per gene were conducted with different sett ings, default and modified versions (Supplementary Data 9). The first analysis using the default setting, 100,000 MCMC iterations, and 10% burn-in identified 4,161 genes to have positively selected sites, with 99% posterior probability (PP). As gaps in multiple sequence alignment might be overestimated due to the misassembling of individual genomes, we removed those sites, and 2,554 positively selected genes were retained. The second analysis using the modified setting, 300,000 iterations, a burn-in of 10,000 iterations and significant PP at 95% was run among the 2,554 genes. Results from both runs were summarized to extract amino acids (aa) or genes under positive or diversifying selection. Test parameters (μ, κ and φ) and their effective sample sizes per gene were also reported. After removing sites with gaps due to the cautious signature of selection[55] and assuring that selected sites showed the same nonsynonymous changes[56] as the original GATK and assembly-based call sets described in Section 3.2, 1,376 functional genes consisting of 2,840 variants under positive selection were identified in the concordant result of the two different runs (Supplementary Data 9). Manual inspection of these selected amino acids and their quality of multiple sequence alignment was verified using Alignment Explorer in MEGA 7.0.21[57]. Gene annotation and deleterious alterations of selected variants were performed using ANNOVAR[14]. Among those selected variants, 11 (8 genes) were initially assigned in the GATK call set as 'VQSRTrancheSNP99.50to100.00' or suboptimal.

Our list of manually inspected genes with selection signature was additionally verified with the following known adaptive regions: *i*) nonsynonymous variant rs4680 G>A within the catechol-O-methyltransferase (*COMT*) gene (Val158Met associated with loss of ability to instantly react

to adverse stimuli, a requirement for forager survival)[58]; *ii*) nonsynonymous variant rs3827760 A>G within the ectodysplasin A receptor (*EDAR*) gene (Val370Ala associated with hypohidrotic ectodermal dysplasia)[59]; *iii*) four nonsynonymous variants within the arylamine N-acetyltransferase 2 (*NAT2*) gene, rs1801279 G>A, rs1801280 T>C, rs1799930 G>A and rs1799931 G>A (Arg64Gln, Ile114Thr, Arg197Gln and Gly286Glu, respectively associated with a loss in the foraging ability to rapidly eradicate dangerous dietary toxins)[60]; *iv*) nonsynonymous variant rs11150606 T>C within the serine protease 53 (*PRSS53*) gene (Gln30Arg associated with a straight hair phenotype)[61]; *v*) nonsynonymous variant rs1426654 G>A within the sodium/potassium/calcium exchanger 5 (*SLC24A5*) gene (Thr111Ala associated with lighter skin pigmentation)[62]; *vi*) nonsynonymous variant rs1871534 C>G within the zinc transporter ZIP4 (*SLC39A4*) gene (Leu372Val associated with reduced zinc transport)[63]; *vii*) nonsynonymous variant rs16891982 C>G within the solute carrier family 45, member 2 (*SLC45A2*) gene (Leu374Phe associated with skin colour); and *viii*) three nonsynonymous variants within the taste receptor, type 2, member 38 (*TAS2R38*) gene, rs713598 C>G, rs1726866 C>T, and rs10246939 C>T (Pro49Ala, Ala262Val, and Val296Ile, respectively (associated with a loss in the ancestral taster haplotype that enables forager people to detect dietary toxins)[64].

***Phenotype informative markers and Manhattan plots*:**

Medical Genome Reference Bank (MGRB) is the first catalogue of whole-genome variation across thousands of healthy elderly individuals, mostly of European descent, with no reported history of cancer, cardiovascular disease, or dementia[65]. The usefulness of this cohort also provides a high-quality resource of well-adapted individuals to non-foraging lifestyles or within modern society, therefore acting as a negative control population compared to forager populations. This could assist in variant filtering and assignment of pathogenicity for inherited or genetic traits[65]. Using the R package SNPRelate[66], the principal component analysis of the MGRB Phase 1 variation data was performed, together with the 1,000 Genomes Phase 3 database of 2,504 genomes from five continental regions[33]. After merging both data and filtered as described in Section 4.1, 3,648 samples and 1,062,513 SNVs were kept for the analysis and showed that most MGRB genomes were of European ancestry, with few being from South Asia (Supplementary Fig. 20).



Supplementary Fig. 20 **Principal component analysis of unrelated samples from MGRB (n = 1,144) and 1,000 Genomes Project Phase 3 (n = 2,504; AFR, Africa; EUR, Europe; EAS, East Asia; and SAS, South Asia).** American continental groups were excluded due to high levels of admixture.

Manhattan plot visualization of significantly selected alleles that were phenotype informative markers[67] was carried out to compare 'San' and 'Damara' (n = 111) *versus* MGRB (2-way). The comparison was performed as association analysis using Fisher's exact test to generate significance of the distribution of allele counts across each of the positively selected amino acids tested above. The identification of alleles significantly represented between the two groups was set with a conservative threshold at *P*-value $<5 \times 10^{-8}$. The association analysis between only 'San' *versus* MGRB was also calculated for significantly selected alleles and then repeated with the comparison of 'Damara' to generate a final set of forager informative markers (3-way). After merging the MGRB Phase 1 variant call set with the Khoe-San dataset of positively selected positions, 1,315 out of 2,840 variants were called in MGRB. Manhattan plots for informative markers and background noise of sites with *P*-value $<5 \times 10^{-8}$ were drawn using the R package qqman[15]. Functional impact, affected genes, gene ontology, and other relevant features of these phenotype informative markers were annotated using ANNOVAR[14]. Pathway enrichment analyses of the gene list significantly associated were searched through the Reactome pathway and gene ontology biological process using g:Profiler[68]. Visualization of significant pathways and classification of their major biological themes was carried out using Cytoscape 3.7.1[69].

***Phenotypic database searches*:**

Although high genetic differentiation between populations is an indicator of positive selection, how the variant produces a molecular effect or changes a protein and how this effect correlates with the selected phenotype provides us with a better understanding of any adaptive alleles and genes[59]. The list of positively selected variants and significantly associated variants identified above between MGRB and forager genomes ('San' and 'Damara') were sought for their potential in functional biology and significance in trait association. For the functional potential of protein changes, we assessed SIFT, Polyphen-2 and MutationAssessor scores and prediction from the ANNOVAR annotation described above. The annotation results also provided InterPro conserved domains for functional regions in known proteins. Similar protein classification using Pfam domains overlapped with the variant list, and their coordinates were retrieved from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables). Searches for trait association among our candidate variants were conducted in the following databases: *i*) PharmGKB (https://www.pharmgkb.org); *ii*) GPCRdb (http://gpcrdb.org); *iii*) GWAS Catalog[70]; and *iv*) GTEx (version 6; https://gtexportal.org). We only considered variant-trait associations with high confidence within those databases.

***Structural model of positively selected genes*:**

We generated a homology protein model of G protein-coupled receptors (GPCRs) for our candidate genes from available GPCR structures using SWISS-MODEL[71]. The modelled protein structure was recorded in the Protein Data Bank (PDB) format and visualized using the PyMOL Molecular Graphics System, Version 2.3.2 Schrödinger, LLC.

Results:

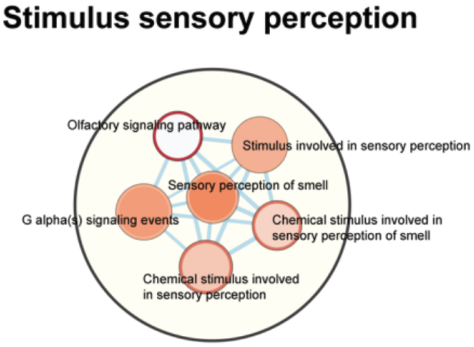
***Effect of positive selection on Khoe-San genomes*:**

On average, two amino acid sites for each gene were found to be positively selected, ranging from one to 17. Most genes (n = 1,228) were affected by one to three selected amino acids (aa). However, more than seven selected sites per gene were found among 13 genes, with several

selected sites observed in *POLQ* and *CMYA5* genes (14 and 17 aa, respectively). The 13 genes fluctuated in sequence length, ranging from 321 to 8,127 aa. Although the incidence of recombination was observed among these genes, with a ρ/θ value greater than one or an excess of the mean amount of population recombination per site over the mean number of point mutations, our positively selected sites are not expected to be erroneous due to minimal bias of high recombination rate reported using the current Bayesian inference[53].

Pathway enrichment for lifestyle association:

Pathway enrichment analysis of 479 significantly selected genes (633 significant selected variants) showed an olfactory signalling pathway of the Reactome database to be the most significant ($-\log_{10} P = 4.533$) after Benjamini-Hochberg (BH) multiple testing. The pathway consisted of the following 17 adaptive genes: *OR10Q1*, *OR10V1*, *OR1A2*, *OR1J1*, *OR2AE1*, *OR2AK2*, *OR2G2*, *OR2L13*, *OR2W3*, *OR2Y1*, *OR4D6*, *OR4K1*, *OR4K14*, *OR51D1*, *OR52K2*, *OR52N4*, and *OR5M11*. Enrichment searches of the Reactome and gene ontology (GO) biological processes showed six pathways sharing the same adaptive genes and contributing to a biological theme related to stimulus sensory perception (Supplementary Fig. 21).



Supplementary Fig. 21 **Pathway enrichment analysis of significant outlier genes.** The major biological theme involves stimulus sensory perception. Cytoscape v3.7.1 showed pathways as circles (nodes) connected with lines (edges) if the pathways share the same genes.

***Structural modelling and visualization*:**

Among the 17 genes mentioned above with strong selection signals involving the same olfactory signalling pathway, 13 displayed 21 nonsynonymous substitutions within coding regions of G protein-coupled receptors (GPCRs; Supplementary Fig. 7C). The receptor mediates most cellular responses by interacting with external stimuli[72] and has been attracted to be druggable targets for treating common diseases[73]. According to the InterPro database, the collection of functional units of proteins[74], the GPCRs in this study corresponded to well-studied family A or rhodopsin-like receptors, where most of the variants were found within their ligand–binding sites or pockets in the extracellular side of transmembrane helices (TM1-7) and extracellular loops[72]. Prominent hydrophobic residues were observed within the pockets that might play a role in conformational changes of the receptors upon ligand binding. Given that six GPCR substitutions were also significant within the GWAS catalogue and GTEx searches, these olfactory signaling genes appear to be the target of selection between forager and non-African populations.

SI11 – References

1. Raczy, C. *et al.* Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041-3 (2013).

2. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206 (2016).

3. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49 (2014).

4. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2015).

5. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **11**, 11.10.1-33 (2013).

6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754-1760 (2009).

7. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint* arXiv:1207.3907 [q-bio.GN] (2012).

8. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).

9. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).

10. Drmanac, R., Peters, B.A., Church, G.M., Reid, C.A. & Xu, X. Accurate whole genome sequencing as the ultimate genetic test. *Clin Chem* **61**, 305-306 (2015).

11. Carnevali, P. *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J Comput Biol* **19**, 279-292 (2012).

12. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-73 (2010).

13. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499-503 (2015).

14. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

15. RStudio-Team. *RStudio: Integrated Development for R*, (RStudio, Inc., Boston, MA 2015).

16. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

17. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).

18. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res* **42**, D749-55 (2014).

19. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).

20. Malaspinas, A.S. *et al.* A genomic history of Aboriginal Australia. *Nature* **538**, 207-214 (2016).

21. Davydov, E.V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).

22. Lu, Q. *et al.* A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* **5**, 10576 (2015).

23. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* **22**, 1154-1162 (2012).

24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580. (1999).

25. Willems, T. *et al.* The landscape of human STR variation. *Genome Res* **24**, 1894-1904 (2014).

26. Gardner, E.J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res* **27**, 1916-1929 (2017).

27. Ondov, B.D., Bergman, N.H. & Phillippy, A.M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011).

28. Bennett, E.A. *et al.* Active Alu retrotransposons in the human genome. *Genome Res* **18**, 1875-83 (2008).

29. Batzer, M.A. & Deininger, P.L. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**, 370-9 (2002).

30. Mills, R.E. *et al.* Recently mobilized transposons in the human and chimpanzee genomes. *Am J Hum Genet* **78**, 671-9 (2006).

31. Scott, E.C. *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res* **26**, 745-55 (2016).

32. Mills, R.E., Bennett, E.A., Iskow, R.C. & Devine, S.E. Which transposable elements are active in the human genome? *Trends Genet* **23**, 183-91 (2007).

33. 1000-Genomes-Project-Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

34. Raj, A., Stephens, M. & Pritchard, J.K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573-89 (2014).

35. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).

36. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

37. Andermann, T. *et al.* Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements. *Syst Biol* **68**, 32-46 (2018).

38. Chen, D., Braun, E.L., Forthman, M., Kimball, R.T. & Zhang, Z. A simple strategy for recovering ultraconserved elements, exons, and introns from low coverage shotgun sequencing of museum specimens: Placement of the partridge genus Tropicoperdix within the galliformes. *Mol Phylogenet Evol* **129**, 304-314 (2018).

39. Pujar, S. *et al.* Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Res* **46**, D221-D228 (2018).

40. Harris, R.S. The Pennsylvania State University (2007).

41. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

42. Nakamura, T., Yamada, K.D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490-2492 (2018).

43. Kumar, S., Stecher, G., Peterson, D. & Tamura, K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* **28**, 2685-86 (2012).

44. Aberer, A.J., Krompass, D. & Stamatakis, A. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst Biol* **62**, 162-6 (2013).

45. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-6 (2011).

46. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-25 (2014).

47. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. **25**, 2078-2079 (2009).

48. Delaneau, O., Marchini, J. & 1000-Genomes-Project-Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **5**, 3934 (2014).

49. Simpson, J.T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117-1123 (2009).

50. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511-8 (2005).

51. Bininda-Emonds, O.R. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* **22**, 156 (2005).

52. Hughes, A.L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167-170 (1988).

53. Wilson, D.J. & McVean, G. Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* **172**, 1411-25 (2006).

54. Yates, B. *et al.* Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res* **45**, D619-D625 (2017).

55. Kvikstad, E.M. & Duret, L. Strong heterogeneity in mutation rate causes misleading hallmarks of natural selection on indel mutations in the human genome. *Mol Biol Evol* **31**, 23-36 (2014).

56. Sabeti, P.C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-8 (2007).

57. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**, 1870-1874 (2016).

58. Stein, D.J., Newman, T.K., Savitz, J. & Ramesar, R. Warriors versus worriers: the role of COMT gene variants. *CNS Spectr* **11**, 745-8 (2006).

59. Szpak, M., Xue, Y., Ayub, Q. & Tyler-Smith, C. How well do we understand the basis of classic selective sweeps in humans? *FEBS Lett* **593**, 1431-1448 (2019).

60. Sabbagh, A., Darlu, P., Crouau-Roy, B. & Poloni, E.S. Arylamine N-acetyltransferase 2 (NAT2) genetic diversity and traditional subsistence: a worldwide population survey. *PLoS One* **6**, e18507 (2011).

61. Adhikari, K. *et al.* A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat Commun* **7**, 10815 (2016).

62. Lamason, R.L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310**, 1782-6. (2005).

63. Engelken, J. *et al.* Extreme population differences in the human zinc transporter ZIP4 (SLC39A4) are explained by positive selection in Sub-Saharan Africa. *PLoS Genet* **10**, e1004128 (2014).

64. Kim, U.K. *et al.* Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. *Science* **299**, 1221-5 (2003).

65. Lacaze, P. *et al.* The Medical Genome Reference Bank: a whole-genome data resource of 4000 healthy elderly individuals. Rationale and cohort design. *Eur J Hum Genet* **27**, 308-316 (2019).

66. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-8 (2012).

67. Petersen, D.C. *et al.* Complex patterns of genomic admixture within southern Africa. *PLoS Genet* **9**, e1003309 (2013).

68. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc* **14**, 482-517 (2019).

69. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).

70. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).

71. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, W296-W303 (2018).

72. Weis, W.I. & Kobilka, B.K. The Molecular Basis of G Protein-Coupled Receptor Activation. *Annu Rev Biochem* **87**, 897-919 (2018).

73. Hauser, A.S. *et al.* Pharmacogenomics of GPCR Drug Targets. *Cell* **172**, 41-54.e19 (2018).

74. Mitchell, A.L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res* **47**, D351-D360 (2019).