Wide-scale Geographical Analysis of Genetic Ancestry in the South African Coloured Population

Imke Lankheet[1*], Rickard Hammarén[1*], Lucía Ximena Alva Caballero[1*], Maximilian Larena[1], Helena Malmström[1,5], Cecile Jolly[1], Himla Soodyall[2,3], Michael de Jongh[4], and Carina Schlebusch[1,5,6]

[*]These authors contributed equally
[1]Human Evolution, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Sweden
[2]Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[3]Academy of Science of South Africa, Pretoria, South Africa
[4]Department of Anthropology and Archaeology, University of South Africa, Pretoria, South Africa
[5]Palaeo-Research Institute, University of Johannesburg, Johannesburg, South Africa
[6]SciLife Lab, Uppsala, Sweden

Corresponding authors:

Carina Schlebusch (carina.schlebusch@ebc.uu.se)

Imke Lankheet (imke.lankheet@ebc.uu.se)

Wide-scale Geographical Analysis of Genetic Ancestry in the South African Coloured Population

Abstract

The South African Coloured (SAC) population, a prominent admixed population in South Africa, reflects centuries of migration, admixture, and historical segregation. Descendants of local Khoe-San and Bantu-speaking populations, European settlers, and enslaved individuals from Africa and Asia, SAC individuals embody diverse ancestries. This study investigates the genetic makeup of SAC individuals, utilizing autosomal genotypes, mitochondrial DNA and Y-chromosome data. We analyze new genotype data for 125 SAC individuals from seven locations. Our analysis, based on a dataset comprising 356 SAC individuals from 22 geographic locations, revealed significant regional variations in ancestry. Khoe-San ancestry predominates in 14 locations, highlighting its lasting influence. Inland regions exhibit higher proportions of Khoe-San ancestry, eastern regions show more Bantu-speaker/West African ancestry, and western/coastal areas, particularly around Cape Town, display increased Asian ancestry. These patterns reflect historical migrations and settlement patterns. Additionally, sex-biased admixture ratios show male-biased admixture from East Africans and Europeans, and female-biased admixture from Khoe-San populations, which is supported by mitochondrial and Y-chromosome data. This research underscores the importance of studying the SAC population to understand South Africa's historical migrations, providing insights into the complex genetic heritage of South Africans.

Keywords

South African Coloured population, genetic admixture, Khoe-San ancestry, sex-biased admixture

Introduction

The South African Coloured (SAC) population is among the most admixed populations in the world, and SAC individuals trace their genetic roots to local Khoe-San and Bantu-speaking groups, European colonists, and enslaved people from other regions in Africa as well as from Asia. Genetically, Khoe-San populations represent one of the two branches of the earliest population divergence of the human population tree and therefore show high genetic diversity [1, 2, 3]. They also host early diverging mitochondrial and Y-chromosome lineages [4, 5, 6, 7]. Until approximately 2000 years ago, the San ancestors were the only inhabitants of Southern Africa and they practiced hunter-gathering [8]. Around 2000 years ago, East-African pastoralists arrived in Southern Africa, and admixed with the local San populations [1, 2, 9, 10, 11], which gave rise to the Khoekhoe herding groups. Today, Khoe-San is the term used to refer to both populations collectively; the hunter-gatherer San and the herder Khoekhoe [12, 13]. The arrival of East-African pastoralists was followed by the arrival of Bantu-speaking groups practicing agriculture and carrying West African ancestry around 1800 years ago as part of the Bantu expansion [14, 15, 16, 17, 18]. The colonial times introduced both European and Asian ancestries into Southern Africa [19]. In 1652, the Dutch East India company founded a small refueling station that gradually grew over the decades into what became known as the Cape Colony and later on as Cape Town. The Dutch settlers interacted heavily with the local Khoekhoe communities from the very foundation of the colony. They traded for cattle and, as time went by, some Khoekhoe would work on settler farmsteads [20]. There were disproportionately few women among the settlers in the colony which led to formal and informal unions between European men and Khoekhoe women [20].

Over time, non-Europeans in the colony became less accepted, leading to the formation of a distinct community. Sometime after 1700, the term “Cape coloureds” emerged to refer to people of mixed ancestry [21]. The Cape coloureds were descendants of Khoe-San, Bantu-speaking populations, European settlers and enslaved people from the West and East Coast of Africa, the Indian subcontinent, Madagascar, and Indonesia, brought to South Africa during the slave trade period (1658-1806) [20]. The apartheid regime, the institutionalised racial segregation in place from 1948 to the early 1990s, enhanced the unity of the South African Coloured group identity [22, 20].

Currently, the SAC population is the largest admixed population in the country [22]. They constitute more than half of the population of the Western Cape Province today, with large presences in the Northern and Eastern Cape provinces as well, see Figure 2F. The majority of the SAC speak Afrikaans as their first language, 75.8% according to the 2011 South Africa (SA) census, and most SAC individuals identify as Christian. Religion serves as an essential characteristic that differentiates SAC from the Cape Malay population, who practice Islam [21] and are also the result of admixture events between Africans and Asians [23]. Despite the term ”Coloured” originating as a construct during the apartheid regime, its usage persists in contemporary South Africa, albeit with varied acceptance.

A number of studies has investigated the genetics of the SAC individuals [24, 25, 26, 11, 27]. They confirm the inferences drawn from historical records: the six main demographic groups that contributed to the genetic pool of the SAC were the Khoe-San, Bantu-speakers/West Africans, East Africans, South Asians/Indians, Southeast Asians and Europeans. A mitochondrial DNA study revealed that the Khoe-San had a large maternal contribution to the SAC (60.0%), while the West Eurasian/European maternal contribution was very limited (4.6%) [25]. However, most of these studies focused on single locations and the majority of locations were close to Cape Town.

In this study, we analyzed new genome-wide data for 125 individuals self-identifying as SAC. The studied individuals are from a wide range of locations, spanning the broad geographic region inhabited by the SAC community, thereby providing a more comprehensive representation of this population. Together with previously published genetic data from SAC individuals, as well as comparative groups, we shed light on the ancestral genetic components present in various SAC populations and identify geographical differences among these components. We also investigate the difference in paternal and maternal contributions for various admixture events through mitogenomes and Y chromosomes, as well as through X-to-autosomal comparisons.

Materials and Methods

Sampling and genome-wide SNP typing

Saliva samples were obtained from 152 SAC individuals from seven different sites in South Africa; two in the Eastern Cape Province (Graaff-Reinet (N=45) and Nieu-Bethesda(N=20)), and five in the Western Cape Province (Genadendal (N=29), Greyton (N=16), Kranshoek (N=11), Oudtshoorn (N=17), and Prince Albert (N=14)). Participants donated saliva samples with written informed consent. Sample collection of SAC, Khoe-San and Khoe-San descendent groups were approved by the University of the Witwatersrand Human Research Ethics board, clearance numbers M980553, with renewals M050902, M090576, M1604104. This specific project was approved by the University of the Witwatersrand Human Research Ethics board, clearance number M180655 and the National Ethics review board of Sweden, clearance number Dnr 2021-01448.

The samples were obtained using an Oragene DNA OG-500 kit. DNA was extracted using the prepIT L2P extraction protocol. The extraction of the biological samples and genotyping followed the procedure described in [11]. The data were generated in four genotyping runs on the Illumina Infinium^{TM} H3Africa Consortium Array by the SNP&SEQ Technology Platform in Uppsala, Sweden. Datasets were analyzed using GenomeStudio 2.0.3 and aligned to the Human Genome build version 37 (hg19). A total of 2,267,346 SNP markers were collected in genotyping run 1, 2, and 3, and 2,271,503 SNP markers were collected in genotyping run 4.

Quality filtering and autosomal dataset merging

The genotype data from 152 SAC individuals was merged with the same dataset as used in [11] [11, 28, 29, 30, 31, 32, 33, 34] as well as with data from additional sources [1, 26, 35, 19]. For further information about the populations included in this study, see Supplementary Table 1. The Petersen dataset was converted to hg37 positions with the LiftOver tool from the Univeristy of California Santa Cruz (UCSC) (https://genome.ucsc.edu/cgi-bin/hgLiftOver). PLINK v1.90b4.9 [36] was used to carry out data processing and quality filtering. Before merging the datasets, duplicate SNPs were removed, only overlapping SNPs between datasets were kept, and C/G and A/T SNPs were eliminated to prevent strand flipping errors. Moreover, 5 individuals with genotyping missingness higher than 15% were excluded (plink –mind 0.15) and SNPs with less than 10% genotyping rate (plink –geno 0.1) were also excluded. Hardy–Weinberg Equilibrium (HWE) was set to 0.00001 (plink –hwe 0.00001) to avoid potential genotyping errors. Once the merging was done, analyses were performed to filter out one individual within each pair of relatives (second-degree or closer) using KING [37]. In total, 22 individuals were removed due to relatedness. Also, SNPs with less than 10% genotyping rate (plink –geno 0.1) were excluded again. To prevent ADMIXTURE and PCA analysis from being negatively affected by linkage disequilibrium (LD) bias, SNPs in LD were removed (plink –indep-pairwise 200 25 0.4). Each of the comparative populations was randomly sub-sampled to 30 individuals per population to avoid a sample-size bias in further analysis. The final dataset comprised 162 382 SNPs and 1203 individuals, of which 356 were SAC individuals and 125 were newly typed SAC individuals. Geographic information of the SAC individuals from previously published data was obtained from their respective publications. Sampling locations are displayed in Supplementary Figure 1.

Population structure inferences

Unsupervised population structure inference analysis for K = 2 to K = 12 was performed with ADMIXTURE [38] version 1.3.0 using a random seed each time, and repeated 50 times. PONG version 1.5 [39] was used to visualize the results and find the major mode and pairwise similarity. Principal component analysis (PCA) was performed using the program `smartpca`, from the Eigensoft package (version 7.2.1) [40, 41]. To capture more of the global variation, Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) was performed on the genotypes directly using the `umap-learn` python library version 0.5.3.

Phasing, local ancestry estimation and admixture dating

Phasing was carried out out using `SHAPEIT` version 2.r837 [42] using the 1000 genomes phase 3 reference genomes [31] and options `--states 500 --main 20 --burn 10 --prune 10`. Any misaligned sites between the reference dataset and the panel were excluded. Local ancestry estimation was performed using `MOSAIC` version 1.5.0 compiled and ran under `R` version 4.3.2 [43], setting source populations to five (-a 5). Admixture

dates were gathered from the reported dates from the co-ancestry curves. The origin of each reconstructed ancestry was determined through F_{st} to the reference populations using `MOSAIC`.

Formal tests of admixture

The dataset was merged with a chimpanzee genome and f4-statistics were computed using popstats [44, 45] in the format f4(Chimp,SAC,Pop1,Pop2). It was used to test whether SAC individuals were more admixed with Pop1 or Pop2. If the f4-value is significantly negative, it implies gene flow between either SAC and Pop1 (or Chimp and Pop2). If it is significantly positive, it implies gene flow between SAC and Pop2 (or Chimp and Pop1).

Sex-biased admixture

To test if the admixture was sex-biased, X-chromosome/Autosomal ratios were computed for the SAC individuals. The genotyping data from the individuals from these seven newly sampled sites were merged with a comparative dataset consisting of 20 Central Europeans (CEU), 20 Sri Lankan Tamil (STU), 20 Nigerian Yoruba (YRI), 20 Ethiopian Amhara and 17 Namibian Ju/'hoansi. The data were filtered as described in the section *Quality filtering and autosomal dataset merging*. To avoid differences in chromosome size affecting the admixture proportions, chromosome 1 to 6 were cut to the length of the X-chromosome (180 centiMorgan), and chromosome 7, 10 and 12 were selected as they roughly have the same length (in centiMorgan) as the X-chromosome. For the autosomes, the number of SNPs was downsampled to the number of SNPs found on the X-chromosome (7452 SNPs). Supervised ADMIXTURE (K = 5) was run separately for each of the autosomes and the X-chromosome with 50 iterations each [38]. The results were visualized with Pong [39]. The ADMIXTURE results provided the ancestry proportions on the X-chromosome per individual and per ancestry. Average autosomal proportions were calculated from the ADMIXTURE runs of each of the autosomes, for each individual and each ancestry. Female X-chromosomal proportions were weighed twice, as females have two X-chromosomes and males only one [46]. One individual was removed because we lacked information to determine whether it was male or female, both from the informed consent form as well as plink sex analysis (plink –check-sex). Corrected X and autosomal proportions were bootstrapped (10 000 times) and average X-to-autosomal difference ratios were calculated as in [47] for each of the five ancestries as follows:

$$\overline{\Delta Admix} = F_{anc,total} * (F_{anc,X} - F_{anc,auto})/(F_{anc,X} + F_{anc,auto})$$

where $F_{anc,total}$ is the genome-wide admixture proportion for a given ancestry, $F_{anc,X}$ is the X chromosome admixture proportion for a given ancestry and $F_{anc,auto}$ is the autosomal admixture proportion for a given ancestry. Negative X-to-autosomal difference ratios are indicative of male-biased admixture for that ancestry, positive X-to-autosomal difference ratios are indicative of a female-biased admixture for that ancestry.

Uniparental markers

Barcoded primers [48](in preparation) were used to amplify the full mitochondrial sequences from 72 SAC individuals. Using a uniquely barcoded primer combination for every sample, we performed a PCR to amplify the whole mitochondrial genome (30x (98 °C, 10 sec; 67 °C, 15 min); 4 °C ∞) (300 ng DNA, 2.4 nM primers, 200 µM of each dNTP, 1x PCR buffer and 1.25 U Takara LA Taq polymerase in 25 µl reaction). Specificity of PCR products was confirmed on a 1% agarose gel and purified with AMPure PB beads. Concentrations of the cleaned PCR products were measured (Qubit). Samples were pooled (100 ng/sample). The pool was purified with 0.5x volumes AMPure PB beads. Elution was performed in 10 mM Tris-HCl, pH 8.5. Concentration of the cleaned pool was measured on the Qubit. The full mitochondrial genomes were sequenced on the PacBio Sequel II. Demultiplexing of the sequencing data was performed by Uppsala Genome Centre (UGC) at NGI-SciLifeLab using the SMRT analysis pipeline (www.pacb.com/products-and-services/analytical-software/smrt-analysis/). The full mtDNA sequence reads were mapped to the Revised Cambridge Reference Sequence (rCRS, NCBI accession number: NC_012920.1) to create BAM files. These were converted to FASTA files using DeepVariant (version 1.3.0, settings: –model_type=PACBIO) and bcftools consensus (version 1.12). Mitochondrial haplogroups were assigned using HaploGrep3 [49]. All haplogroups were associated with an ancestry, according to literature (see Supplementary Table 4).

Y chromosomal haplogroups were assigned for all 119 males using SNAPPY [50] on the genotyping array data and all haplogroups were associated with an ancestry, according to literature (see Supplementary Table 5).

Results

In this study, we aim to provide a comprehensive analysis of the genetic ancestry of the South African Coloured (SAC) population by investigating genome-wide data from 356 (125 new) individuals self-identifying as SAC, coming from 22 (7 new) locations in South Africa (Supplementary Figure 1). Building upon previous genetic research, our investigation encompasses a thorough examination of ancestral genetic components within the SAC, for the first time focusing on geographically dispersed SAC groups. Employing a combination of genomic techniques, including analysis of mitogenomes, Y chromosomes, and X-to-autosomal comparisons, we investigate the complexities of admixture events and explore geographic variations in ancestral contributions. Through these approaches, we seek to elucidate the complex genetic make up of SAC and shed light on the historical and demographic factors that have shaped this diverse population.

Autosomal ancestry contribution in geographically dispersed SAC groups

We created a database consisting of 356 SAC individuals and 847 reference individuals. To capture the major genetic variation between continental groups and to investigate the affiliations of SAC individuals in this genetic space, we applied principal component analysis (PCA) to our dataset. The first principal component separates the out-of-Africa populations from the Khoe-San and West-African populations, while the second principal component represents the variation between Khoe-San and West African-related ancestries (Figure 1A). The SAC individuals are observed scattered in-between these extremes, with some individuals associating more with either Khoe-San, non-African or West African groups. Moreover, PC3 separates East Asian and European ancestry, with South Asians grouping between these two extremes (Supplementary Figure 2). Certain SAC individuals are off-set towards the Asian extreme, suggesting increased ancestry contributions from Asians. Analyzing the average PC values per population (Supplementary Figure 3) reveals a noticeable west-to-east pattern in the PCA. The western locations District Six, Wellington, Genadendal, and Greyton tend to cluster nearer to European populations, while the eastern locations Graaff-Reinet and Nieu-Bethesda show closer proximity to Khoe-San and West-African/Bantu-speaking groups. The three other new locations, Kranshoek, Oudtshoorn, and Prince Albert, which are located geographically in between the previously mentioned groups, also occupy the space in the PCA plot between these groups. Among the three, Kranshoek is closest to Genadendal and Greyton in the PCA plot.

Uniform manifold approximation and projection for dimension reduction (UMAP) identifies the major variation in the data and reduces it down to only two dimensions, thus allowing a graphical overview of the variation [51]. Unlike PCA, UMAP aims to capture more of the global variation [51]. The UMAP analysis recapitulates the major continental ancestries within the dataset, with the more drifted out-of-Africa populations forming tightly clustered groups away from each other (Supplementary Figure 4). Khoe-San, SAC, and Bantu-speaker related ancestry populations form a larger group in the center of the UMAP. Most of the SAC are positioned close to the Khoe-San populations but are drawn towards either the European or Bantu-speaker related ancestry. Some SAC individuals cluster firmly with other populations, rather than with the other SAC individuals. Three individuals from District Six, Northern Cape, and Genadendal are closely associated with the European populations. Three other individuals are associated with South Asians, two from Wellington and one from District Six. Additionally, five SAC individuals from various locations group with the South African Bantu-speaking populations.

To further investigate the population structure and ancestral contributions to the SAC populations, we performed unsupervised ADMIXTURE analysis for K = 2 to K = 12 (Supplementary Figure 5). At K = 6 (Figure 1B), we identified components corresponding to major continental and regional groups: Khoe-San, European, West African/Bantu-speakers, East African, East Asian, and South Asian. Compared to K = 6, K = 10 revealed additional clusters: one associated with the East African Hadza, another associated with the East African Sabue, a cluster separating northern Khoe-San from southern Khoe-San populations, and a cluster separating Bantu-speakers from West African non-Bantu Niger-Congo speakers. K = 10 (Figure 1C) had the lowest cross-validation error, see Supplementary Figure 6. Average admixture fractions at K = 6 are shown in Supplementary Table 2. From the 22 locations with SAC individuals, Khoe-San ancestry is predominant at 14 locations, including the new locations of Graaff-Reinet, Nieu-Bethesda, Kranshoek, Oudtshoorn and Prince Albert. Khoe-San ancestry ranges from 12.0% (District Six) to 69.0% (Askham) across all sites, with an average of 33.4%. Based on K = 10 ADMIXTURE results, we can conclude that this observed Khoe-San ancestry is mostly southern Khoe-San rather than northern Khoe-San (yellow vs gold respectively in Figure 1C). This was confirmed by f4-statistics in the form $f4$-(Chimp, SAC, Ju/'hoansi, Karretjie) (Supplementary Figure 7). European ancestry is predominant in seven locations,

including Genadendal and Greyton. Generally, European ancestry ranges between 9.2% (Nieu-Bethesda) and 40.5% (Northern Cape) in the studied SAC populations, with an average of 21.7%. In Railton, West-African ancestry constituted the largest proportion (32.8%) while the West-African ancestry was lowest in Northern Cape (9.4%).

At K = 9, ADMIXTURE analysis separates the West African ancestral component into a cluster maximised in West African non-Bantu Niger-Congo speakers (dark-brown) and another in Bantu-speaking populations (light grey)(Supplementary Figure 5). From this K and higher, the component found among the SAC is mostly related to Bantu-speaker ancestry rather than non-Bantu Niger-Congo speakers. In addition, to directly evaluate the genetic affinity of West African/Bantu-speaker ancestry found in SAC, we conducted the test *f4*(Chimp, SAC, YRI_AFR, Zulu) (Supplementary Figure 8). All SAC groups, except the Coloured from Askham exhibit greater genetic affinity to the Yorubans relative to the South African Bantu-speaking Zulu.

F4-statistics were computed to assess the genetic affinity of the Asian component in the SAC population (*f4*(Chimp, SAC, CHB_EAS, GIH_SAS)), where CHB_EAS are the Han Chinese (East Asian) and GIH_SAS are the Gujarati Indians (South Asians) (Supplementary Figure 9). Positive values for most SAC groups imply more genetic affinity with South Asians rather than East Asians. The ADMIXTURE results also highlight an additional interesting aspect about the ancestry of the studied SAC populations, namely the presence of a genetic component shared with the Malagasy populations. From K = 4 to K = 10, the genetics of the Malagasy populations (Mikea, Temoro, Vezo) can be explained as being comprised mainly of two clusters; a West African cluster (grey) (∼60%), and a East Asian cluster (∼40%). However, from K = 11, the Malagasy populations get their own cluster (royal blue), with some minor West African and East Asian contributions. This genetic cluster can also be observed in the various SAC populations, at low percentages. The average percentage of this component across all the studied SAC locations is 5.8%. Positive values for f4-statistics in the form *f4*(Chimp, SAC, Malagasy, South Asian) (Supplementary Figure 10) for all SAC groups imply they possess greater genetic affinity with South Asians relative to Malagasy people.

As the ADMIXTURE at K = 6 captures best the diversity in ancestries in the SAC, the average ADMIXTURE derived ancestral fractions for K = 6 were plotted on a map of the southern part of South Africa, to investigate spatial patterns of the different major ancestries (Figure 2). Khoe-San ancestry is larger towards the inland regions and towards the east, while Bantu-speaker ancestry proportions are higher in the most eastern localities. The combined East and South Asian ancestry is highest close to Cape Town and decreases with increasing distance. East African ancestry is smaller than the other ancestries (0.1-2.9%), but geographical differences can be observed with higher East African ancestries along coastal regions and along the Gariep river valley (northern-most point). The European-related ancestry is highest along the coast, with the exception of the Northern Cape site.

Admixture dating

SAC individuals trace their ancestry to major ancestral groups that might have admixed during different time periods. We employed a local ancestry estimation method to discern the mosaic composition of the genomes of SAC individuals, delineating which segments most likely originated from each parental population. We employed a 5-way admixture model in MOSAIC allowing for all reference populations as parental populations, meaning that MOSAIC uses the reference populations to construct five ancestries that best describe the haplotypes observed in the target. Each constructed ancestry is compared to the source populations through F_{st}. The constructed ancestries typically reflect the major continental ancestries that we get from ADMIXTURE, see the 1-F_{st} plots in Supplementary Figures 14 to 57. However, in some cases several ancestries belong to the same continental ancestry. One such case is Oudtshoorn, where the fourth and fifth ancestries are closest to the Khomani and Karretjie respectively, both southern Khoe-San groups.

Subsequently, using this information, we retrieved the admixture dates of these parental populations from the co-ancestry curves as generated by MOSAIC (Figure 3). Most of the dates fall within less than ten generations, overlapping with the time period since European colonisation (1650 onward). Nine SAC populations display admixture dates that are above 50 generations ago (corresponding to 1450 years, assuming a generation time of 29 years [52]). Seven of these older admixture dates are associated with Khoe-San and East-African ancestry, two of them with European and South Asian ancestry.

Patterns of sex-biased admixture

Previous studies have shown that the admixture events that shaped the SAC population were sex-biased [25, 11], indicating that the extent of male and female genetic contribution from different admixing populations may have varied. Here, we investigate the sex-biased nature of the admixture events shaping the SAC populations further by performing supervised ADMIXTURE for the autosomes and X-chromosome (Supplementary Figure 11) and looking at the ΔAdmix ratios of East-African, European, Khoe-San, Asian and West-African ancestry (Figure 4A). Negative X-to-autosomal ΔAdmix ratios are indicative of male-biased admixture for that ancestry, positive ΔAdmix ratios are indicative of a female-biased admixture for that ancestry. We observe negative ΔAdmix ratio values with 95% confidence intervals not overlapping zero for East-African and European ancestries (-0.0177 and -0.0259 respectively), indicating male-biased admixture from East-Africans and Europeans. We observe a positive ΔAdmix ratio with 95% confidence intervals not overlapping zero for Khoe-San ancestry (0.0365), indicating female-biased ancestry from Khoe-San people. For both Asian and West African ancestries, 95% confidence intervals overlap zero and are therefore not significantly differing from zero, thus indicating non-significant sex-biased admixture from these ancestries. The same analysis was also performed per site (Supplementary Figure 12), and although all ΔAdmix ratios associated with European ancestry are negative and all those associated with Khoe-San ancestry are positive, the 95% confidence intervals often overlap zero, due to smaller sample sizes.

These signals of sex-biased admixture are further supported with data from the uni-parental markers of these individuals. The mitochondrial genome and the Y-chromosome allow for the study of maternal and paternal lineages separately in a population. We generated novel mitochondrial DNA sequences for 72 SAC individuals and determined the Y-chromosome haplogroups for 67 newly genotyped SAC individuals using SNAPPY [50]. We combined these data with the individuals from the reference datasets. Our results show that mitochondrial haplogroups associated with Khoe-San ancestry are more frequent in the SAC populations than the Y chromosome haplogroups associated to Khoe-San ancestry (Figure 4B). The opposite pattern is observed for West-African and European associated mitochondrial and Y chromosome haplogroups. Haplogroups associated with East-African ancestry become less frequent as we move from mitochondrial genomes to Y chromosomes, but fractions are low (less than 0.031). For both Asian ancestries, no clear pattern can be observed. Supplementary Figure 13 shows the associations of the mitochondrial genomes, autosomes and Y chromosomes to the six different ancestries for each of the separate sites. Large differences in continental distributions can be observed between sites such as Genadendal, Graaff-Reinet and Askham (numbers of indiviudals used for each analysis can be found in Supplementary Table 8). Genadendal, located in the west, generally shows more European ancestry in autosomes, and more mitochondrial and Y chromosomal haplogroups associated with Europeans. This contrasts with the locations of Graaff-Reinet and Askham, situated more to the east and north, respectively. Elevated Khoe-San ancestry can be observed at Askham for all three genetic markers (MT, autosomes and Y), whereas elevated Bantu-speaker ancestry is evident for all markers in Graaff-Reinet.

Discussion

In this study, we have analyzed genome-wide data from 125 SAC individuals, coming from seven different locations in South Africa. Combining this information with previously published population genetic data, our investigation encompasses a thorough examination of ancestral genetic components within the SAC, spanning a wide geographic distribution. We have investigated the complexities of admixture events that shaped the SAC population and explored potential geographic variations in ancestral contributions. We find evidence of geographical stratification of genetic ancestries in agreement with historical information.

Ancestry proportions in the South African Coloured population

Our analysis of the general genetic background of the SAC population through PCA and ADMIXTURE (Figure 1) supports the previously identified ancestral components: Khoe-San, West African and Bantu-speaker, European, East African, South and East Asian [24, 25, 26, 11]. We identified heterogeneity within the SAC, ancestry proportions differing substantially across individuals (Figure 1B and C). Average continental ancestry is generally comparable across sites, albeit with some regional variation.

Results from the ADMIXTURE analysis, Figure 1B & C, align with previous studies [26, 11, 1] indicating that the primary genetic ancestry found among the SAC people is Khoe-San. ADMIXTURE at K = 8 (Supplementary Figure 5) splits the northern Khoe-San from the southern Khoe-San populations, thereby revealing for the first time that the Khoe-San ancestry in the SAC is mostly southern Khoe-San-related (Nama, Karretjie, and Khomani). ADMIXTURE at K = 9 separates the West African ancestral component into a component maximised in West African non-Bantu Niger-Congo speakers (dark brown) and another in Bantu-speaking populations (light grey) (Supplementary Figure 5). The analysis at K = 10 reveals low contribution (0.4-3.3%) of the West African ancestry in the SAC. Slaves were brought to the Cape colony from the West African kingdom of Dahomey and from Angola in 1658, and they were a part of the founding population of the SAC [20]. Thus it is interesting to see that the ADMIXTURE analysis only reports low West African ancestry contributions among the SAC populations, and around 22.5% (minimum 7.6, maximum 39.5) of the Bantu-speaker-associated ancestry. According to the ADMIXTURE analysis, these initial enslaved individuals seem to have contributed a small but consistent amount of ancestry to the SAC communities. However, the f4-statistics in the form *f4*(Chimp, SAC, YRI_AFR, Zulu) indicate a greater genetic influence from the West-African Yoruba (Supplementary Figure 8). This mismatch with the ADMIXTURE results has been observed in other studies as well, and is called "neighbour repulsion" [53], where the neighbouring populations (Zulu in this case) received independent gene flow from an external source after their split from the West-Africans. In Afrikaners, the West African non-Bantu Niger-Congo speaker associated component contributes more than the Bantu-speaker associated component [19]. This difference in ancestry contributions based on ADMIXTURE likely reflects different patterns of historical admixture for the SAC and Afrikaner populations. West African admixture into Afrikaners likely occurred during the early phases of founding of the colony, with slaves of West African origin, while most of the West African component in the SAC groups was most likely contributed through continued admixture with Bantu-speakers in the contact zone towards the east.

District Six and Wellington have relatively high South Asian ancestry contributions (Supplementary Table 2), likely due to the specific social dynamics at these sites. The District Six community was formed by formerly enslaved people, merchants, and immigrants. Cape Malays, brought as part of the slave trade, composed an essential portion of the founding community, along with the Xhosa people. Afrikaners composed only a small part of the residents of District Six until apartheid laws declared it a "whites-only" area in 1966, causing many people to be forcibly relocated [54]. Today, more than 90% of its inhabitants are SAC [55]. Similarly to District Six, Wellington was founded in 1699 as an agricultural town. Until the first part of the 20th century, it was mainly composed of SAC residents, many of whom were Muslims and of Asian descent [56].

Our ADMIXTURE analysis at K = 6 also highlights the genetic contributions of Asian populations to the SAC population. With the average South Asian contribution at 12.1%, it is roughly twice as large as the contribution from East Asians. F4-values in the form *f4*(Chimp, SAC, CHB_EAS, GIH_SAS) are positive for most SAC groups, supporting more admixture from South Asians (Supplementary Figure 9). Thus, we conclude that most of the Asian slaves were brought from South Asia, and to a lesser extent from East Asia. This corresponds to the historical record stating that the Dutch East India Company imported slaves from Indonesia to South Africa [57].

Through ADMIXTURE and subsequent f4-statistics, we also elucidate for the first time the contribution

of Malagasy populations to the SAC population. At K = 11, the Malagasy populations get their own cluster (royal blue), with some minor West-African and East-Asian contributions (Supplementary Figure 5). This Malagasy population genetic cluster can also be observed in the various SAC populations, with an average percentage of across all the studied SAC locations of 5.8%. We computed f4-statistics in the form *f4*(Chimp, SAC, Vezo, GIH_SAS)(Supplementary Figure 10) and show that there is less genetic affinity of the SAC to the Malagasy, when compared to the South Asian population. This is not to say that no admixture occurred with Malagasy people, just that South Asians have made a larger contribution compared to the Malagasy contribution. The Malagasy ancestry found in the SAC population is also consistent with the historical record that the Dutch East India Company imported slaves from Madagascar to South Africa [57].

Regional differences in observed ancestry proportions

We collected data from seven new locations to further identify regional variations in SAC ancestries. The ancestry proportions at K = 6 on a map of South Africa (Figure 2) reveal various interesting trends. The Bantu-speaker ancestry shows higher contributions in the East, and lower contributions in the West. This can largely be attributed to the dominant presence of Bantu-speaking groups in the eastern regions of South Africa, which marks the historical limit of the Bantu expansion [20]. The high prevalence of Khoe-San ancestry in the SAC in the inland regions and toward the east reflects the influence of the Cape colony and the increased admixture from Europeans and enslaved people from Asia and Madagascar in the areas closer to the coast and to Cape Town. In the Northern Cape region, Khoe-San ancestry is high (33.4-69.0%), and Bantu-speaker and West African ancestry is low (9.4-11.6%). The SAC of the Northern Cape can be traced back to the Nama herder groups who resided in Namaqualand (South Africa) and Namibia, local San hunter-gatherer groups, and to European settlers who moved into these interior areas. Thus, the Nama people likely contributed to the high Khoe-San genetic ancestry in the SAC individuals in the Northern Cape. This is supported by MOSAIC analyses, which find low F_{st} values for the Nama as a source population for the SAC at Askham and Northern Cape (Supplementary Figures 14 and 22). The low Bantu-speaker (and West African ancestry) component in Askham and the Northern Cape site points to limited admixture with Bantu-speakers. The distribution pattern for combined Asian ancestries and, to a large extent, European ancestry exhibits an interesting contrast. In the Cape region, high contributions from Asian and European ancestries can be observed, gradually decreasing as one moves eastwards. This phenomenon finds its roots in the historical influx of European settlers into the Cape Colony, with its centre and entry point at the Cape of Good Hope (current-day Cape Town), accompanied by enslaved people from Asia and other parts of Africa. Although East African ancestry proportions are very low in comparison to the other ancestries, it is higher along coastal regions and along the Gariep river valley (northern-most point) correlating with the past distribution of Khoekhoe herder groups (vs. San hunter-gatherer groups) [8, 58].

Dating admixture events in the SAC population

Since the SAC individuals trace their ancestries to various continental and sub-continental sources, we set out to investigate when these populations admixed. We identify that most of the admixture dates fall within less than ten generations, aligning with the anticipated timeframe for the formation of the Cape Colony. The admixture dating using MOSAIC (Figure 3) also identified several admixture events that can be correlated with the formation of the Khoekhoe with the arrival of East African pastoralists in southern Africa [59, 60, 61, 11]. This can be seen in a few dates that are very old (longer than 50 generations ago). These exact dates should, however, be viewed with caution as they are based on small fractions of ancestry, have deep time estimates and have parental source groups that might be distant from actual source groups. This uncertainty is reflected in the co-ancestry graphs that the dates are inferred from, in which the Khoe-San vs. East African admixture estimates are the least robust of the analyses, Supplementary Figures 14 to 57. The East African ancestry contribution is the smallest according to the ADMIXTURE results (Supplementary Table 2) and minor ancestries are problematic for the proper fitting of co-ancestry curves. Two of the admixture dates older than 50 generations ago can be attributed to European and South Asian ancestries. This reflects admixture events happening outside the African continent before these ancestries were introduced during colonial times, possibly related to Eurasian trade routes such as the Silk Road (200 BCE - 1450 AD).

The admixture events with Bantu-speakers (unlabeled in Figure 3) mostly occurred during and after colonial times. This indicates that most of the admixture between Khoe-San and Bantu-speakers also occurred after colonial times, due to the disruptions and population mobility that the colonial times instigated. Even

the admixture events between Asian and West-African/Bantu-speaker ancestries are all between 3.9 and 11.7 generations ago, also corresponding to the colonial period. Malagasy populations are known to be the result of an admixture event between Austronesian and Bantu sources around 20 to 32 generations ago [35]. These sources are supported by the ADMIXTURE analysis in this study (Figure 1B&C). However, we do not observe the same generation time-frame for the admixture event between Asian and West-African/Bantu-speaker ancestries in the SAC individuals, possibly indicating that most of these ancestries came from Asian, West-Africans, and Bantu-speakers directly, and not from Malagasy populations. This observation fits with the small Malagasy contributions observed at K = 11 (Supplementary Figure 5) and is in line with what has been observed in other SAC populations [11].

Sex-biased nature of admixture events in the Coloured

From historical records, we know that there were disproportionately few women among the European settlers, especially in the period before 1688 [20]. A previous genetic study concluded that Khoekhoe women constituted the majority of the maternal contribution for the SAC groups [25]. Moreover, additional investigations into the sex-biased nature of the admixture events shaping the SAC using X-chromosome and autosomes inferred a male-biased influence from East Africans, Asians, and Europeans, and a female-bias from Khoe-San and West-African individuals [11]. In the current study, we also observe a male-biased admixture from East Africans and Europeans, and a female-biased admixture from Khoe-San. The Asian ancestry shows a very small female bias and the West-African ancestry a male bias. However, both of these trends are not statistically significant. The investigation of sex-biased patterns at individual sites (Supplementary Figure 12) highlights the heterogeneous nature of the SAC population. Although non-significant, West-African sex-biased admixture ratios are female-biased in some sites (Genadendal, Greyton, Oudtshoorn, and Kranshoek), while being male-biased in other, more northeastern sites (Graaff-Reinet, Nieu-Bethesda and Prince Albert). The sex-biased admixture in the SAC is supported by the findings from the uni-parental markers; mitochondrial genomes and Y chromosomes. Mitochondrial haplogroups associated with Khoe-San ancestry are more frequent in the SAC populations than the Y chromosome haplogroups associated to Khoe-San ancestry (Figure 4B). The opposite pattern can be observed for West-African and European associated mitochondrial and Y chromosome haplogroups. Since the mitochondrial genome is inherited through the female line and Y chromosomes completely through the male line, we observe them as the extremes when it comes to differences between the contribution of the two sexes, whereas the autosomal ancestries are observed somewhere in the middle of these two. We also note the regional variation between the male and female contributions from different populations across the sites (Supplementary Figure 13), again highlighting the genetic heterogeneity of the SAC population.

Conclusions

In this study, we analyzed new genotype array data for 125 South African Coloured individuals and built upon research to describe the genetics of one the most admixed populations in the world, the SAC. The Khoe-San people, especially the southern Khoe-San, played a major role in the foundation of the SAC, with their ancestry contribution ranging from 12.0-69.0% across all investigated sites. We also identified a considerable variation in ancestry contributions between different individuals. By adding genetic data from seven new geographically dispersed sites, we were able to better investigate geographical differentiation in ancestry proportions and we identified higher Khoe-San contribution in inland regions and toward the east, and higher Bantu-speaker contributions in eastern regions, whereas the Asian ancestry is higher in western regions. Near Cape Town and in the Western Cape province, the non-African ancestry is especially high, reflecting the historically greater density of European colonists and slaves in those locations. We infer that the admixture events shaping the SAC were in many ways sex-biased; mainly female-biased from Khoe-San people and male-biased from both East Africans and Europeans. Altogether, this study highlights the intricate admixture history and diverse ancestry of the SAC population.

Availability of data and materials

All data generated or analyzed during this study are included in this published article, its supplementary information files and publicly available repositories. The generated genotype data is available for academic research use through the European Genome-Phenome Archive with accession number EGAD50000000513 (152 individuals) and Data Access Committee EGAC50000000240. Scripts are available at https://github.com/imkelankheet/Sou African-Coloured-project.

Acknowledgements

We are grateful to all subjects who participated in this research. We also would like to thank Gustav and Lili Radlov, Chris and Marie Heese, Andrew and Anneke Fraser-Jones, Harry and Belinda Gordon, Dr. Morley, Monica Thomson, Michelle Moodie, Poem Mooney, Dr Judy Maguire, Maria Johnson and the family Simmers for help during field collection. We thank Michael Salter-Townshend for support with MOSAIC-related discussions. We are grateful for Cesar Fortes-lima for his help with the EGA data upload. The genotyping was performed by the SNP&SEQ Technology Platform (Uppsala, Sweden). The facility is part of the National Genomics Infrastructure supported by the Swedish Research Council for Infrastructures and Science for Life Laboratory (NGI-SciLifeLab), Sweden. The SNP&SEQ Technology Platform is also supported by the Knut and Alice Wallenberg Foundation. The computation and data handling were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppmax partially funded by the Swedish Research Council through grant agreement no. 2018-05973. This project was supported by funding to CS from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 759933), the Knut and Alice Wallenberg foundation, the Leakey foundation and the Erik Philip Sorensson foundation. An authorized NIH Data Access Committee (DAC) granted data access to Carina Schlebusch for the controlled-access genetic data analyzed in this study that were previously deposited by Scheinfeldt et al. (2019) in the NIH dbGAP repository (dbGaP accession code: phs001780.v1.p1; project approval date: 2019-05-17), as well data previously deposited by Martin et al. (2017) in the NIH dbGAP repository (dbGaP accession code: phs001753; project approval date: 2019-10-25).

Figures

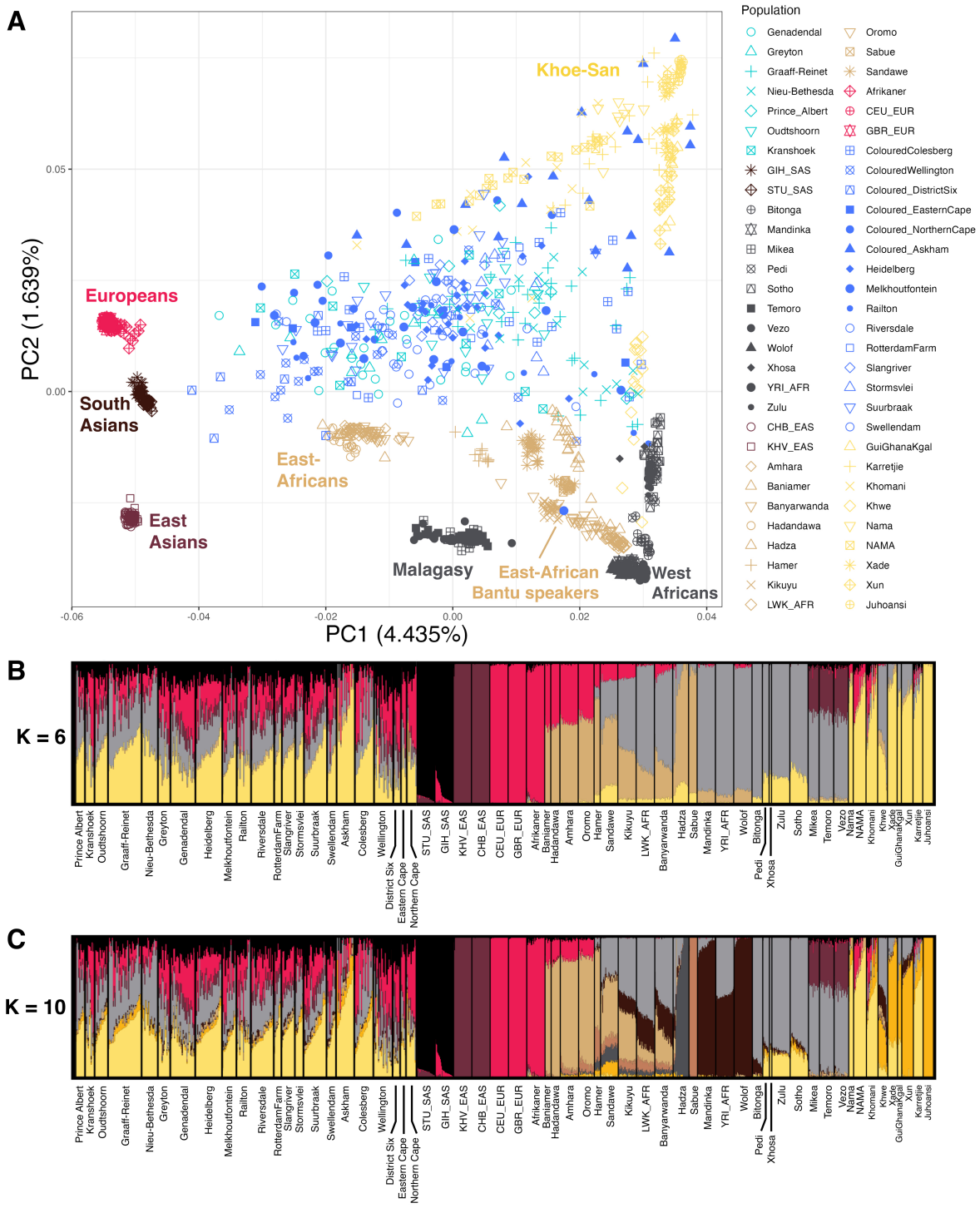

Figure 1: Population structure and genetic affinities of South African Coloured population. Principal component analysis (PCA) and ADMIXTURE results for the populations in our dataset, including 356 SAC individuals. In A, principal component analysis (PCA) results are shown, where PC1 and PC2, are plotted against each other. Labels according to continental groups were added *a posteriori* to help with legibility. The new SAC samples are shown in light blue, the previously published ones in dark blue. For geographical origins of populations, see Supplementary Figure 1. Other PCA projections can be found in Supplementary Figure 2. B and C show ADMIXTURE results, visualized using PONG for K = 6 and K = 10 respectively. ADMIXTURE results for K=2 to K=12 can be found in Supplementary Figure 5. GIH_SAS are the Gujarati Indians, STU_SAS are the Sri Lankan Tamil, YRI_AFR are the Yoruba from Nigeria, CHB_EAS are the Han from China, KHV_EAS are the Kinh from Vietnam, LWK_AFR are the Luhya from Kenya, CEU_EUR are Utah residents with Northern and Western European ancestry, and the GBR_EUR are the British in England and Scotland.

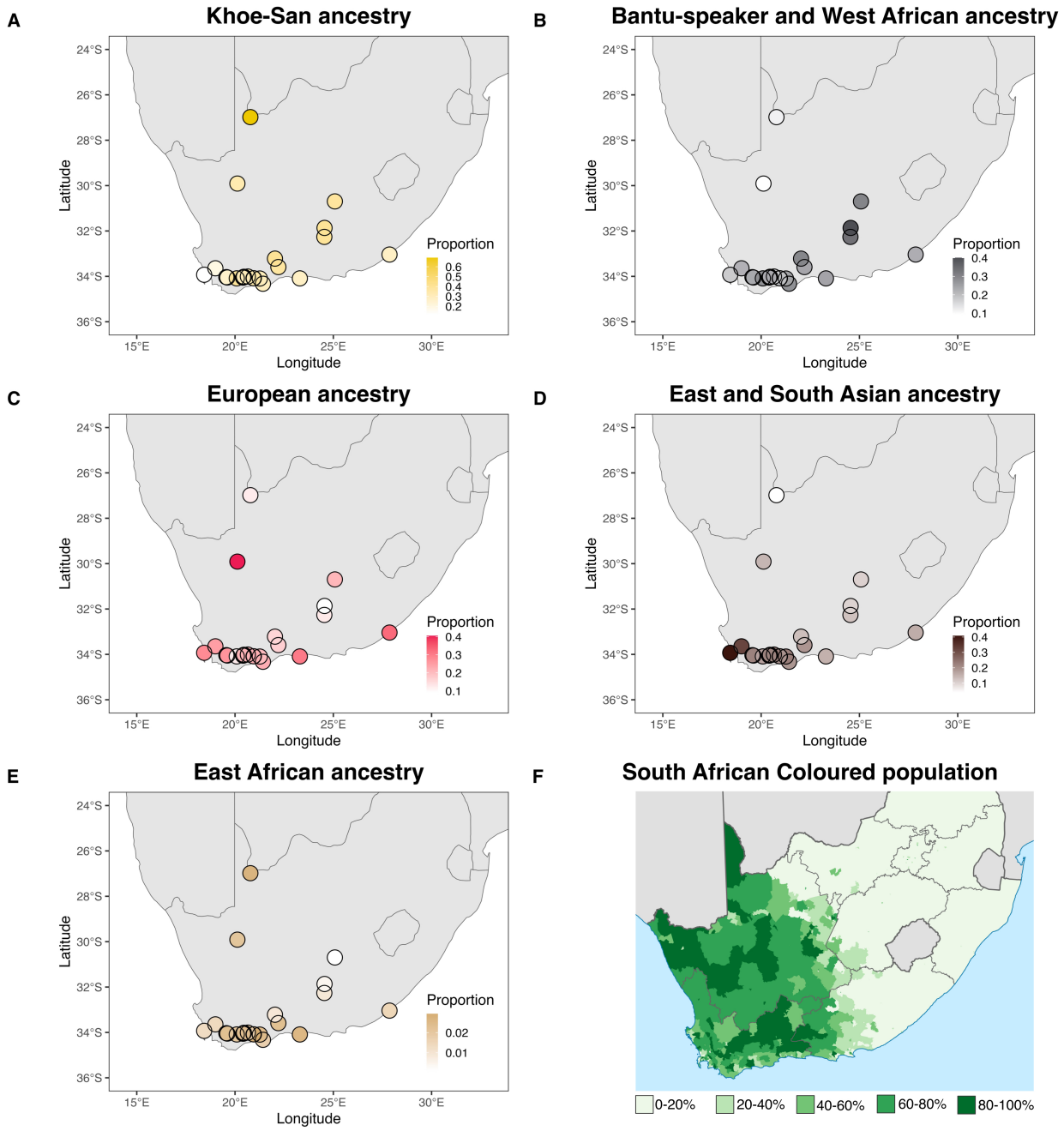

Figure 2: Visualisation of averaged ADMIXTURE derived ancestry proportions from K = 6 plotted by sampling locations. The colour scale is relative to the maximum value of each fraction of admixture. A depicts the component associated to Khoe-San ancestry, the corresponding is shown for B, West African and Bantu-speaker ancestry, C European ancestry, D South Asian and East Asian ancestry combined, and E East African ancestry. In F, the proportion of SAC people among the inhabitants is shown per region in South Africa. Based on the 2011 census. Adapted from `https://commons.wikimedia.org/wiki/File:South_Africa_2011_Coloured_population_proportion_map.svg`, Public Domain.

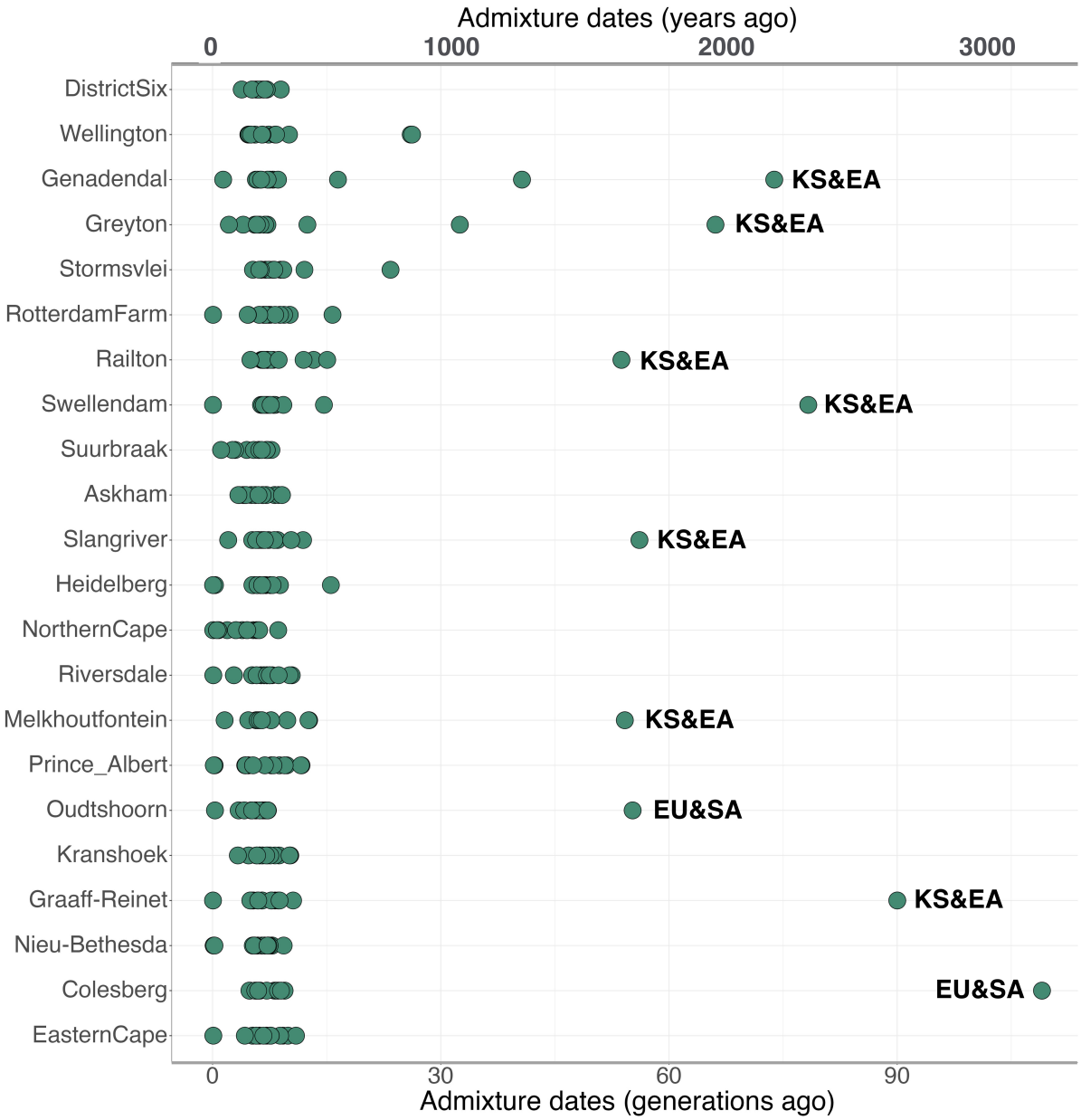

Figure 3: Inferred admixture dates for the 5-way admixture scenario for the 22 SAC populations using all reference populations as putative sources. Dots labeled with "EA & KS" indicate admixture events between Khoe-San and East African constructed ancestries, as determined by F_{st}. Dots labeled with "EU & SA" indicate admixture events between European and South Asian constructed ancestries. Sites are shown from West (high) to East (low) on the y-axis. X-axis on top shows the time in years, x-axis at the bottom shows time in generations ago.

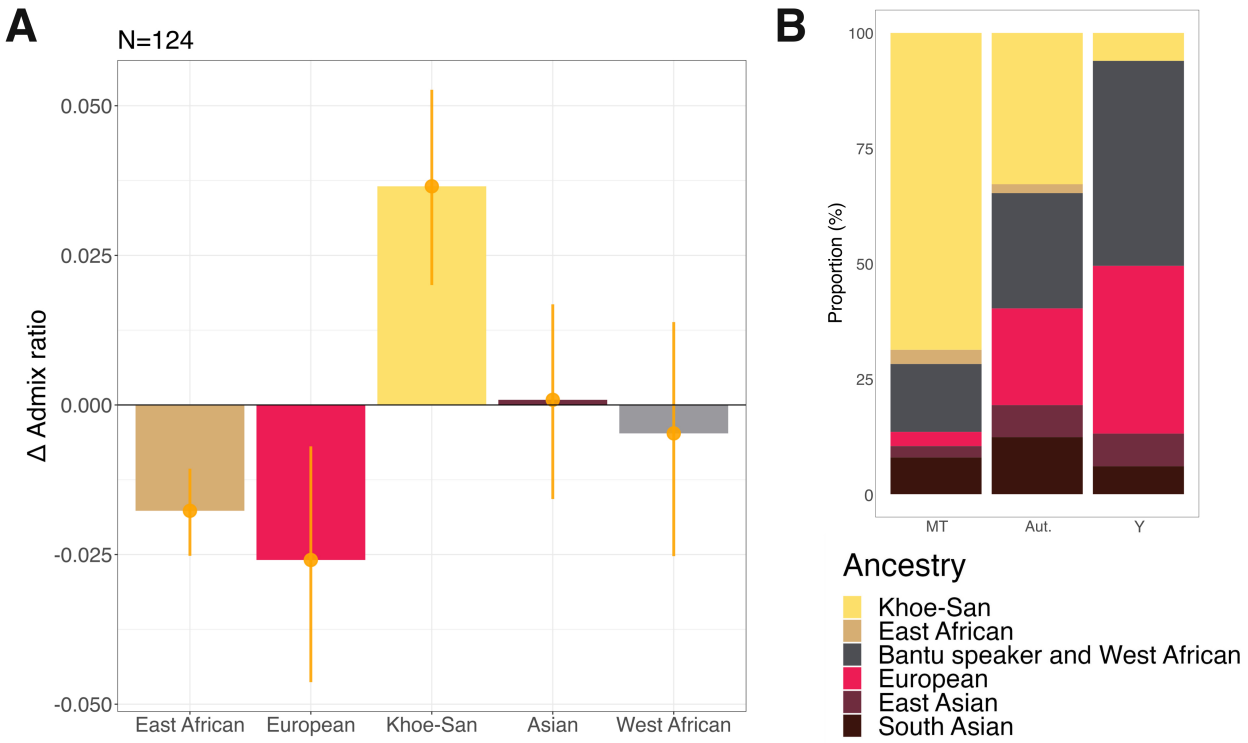

Figure 4: Average sex-biased admixture among the SAC people. In A, ΔAdmix ratios for each of the five ancestries, averaged over the seven investigated sites are shown. X and autosomal proportions were bootstrapped (10 000 times) and average X-to-autosomal difference ratios were calculated for each of the five ancestries, as well as standard deviations. The error bars indicate the 95% confidence interval. Negative X-to-autosomal difference ratios are indicative of male-biased admixture for that ancestry, positive X-to-autosomal difference ratios are indicative of a female-biased admixture for that ancestry. Results from sex-biased admixture analyses per site can be found in Supplementary Figure 12. In B) ancestries associated with the mitochondrial and Y genomes, as well as the autosomal proportions are shown for all studied SAC individuals.

References

[1] Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science (New York, NY). 2012 Oct;338(6105):374-9. doi:10.1126/science.1227721.

[2] Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science (New York, NY). 2017 Nov;358(6363):652-5. doi:10.1126/science.aao6266.

[3] Schlebusch CM, Sjödin P, Breton G, Günther T, Naidoo T, Hollfelder N, et al. Khoe-San Genomes Reveal Unique Variation and Confirm the Deepest Population Divergence in Homo sapiens. Molecular Biology and Evolution. 2020 Oct;37(10):2944-54. doi:10.1093/molbev/msaa140.

[4] Naidoo T, Schlebusch CM, Makkan H, Patel P, Mahabeer R, Erasmus JC, et al. Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. Investigative Genetics. 2010 Dec;1(1):1-11. doi:10.1186/2041-2223-1-6.

[5] Naidoo T, Xu J, Vicente M, Malmström H, Soodyall H, Jakobsson M, et al. Y-Chromosome Variation in Southern African Khoe-San Populations Based on Whole-Genome Sequences. Genome Biology and Evolution. 2020 07;12(7):1031-9. doi:10.1093/gbe/evaa098.

[6] Barbieri C, Güldemann T, Naumann C, Gerlach L, Berthold F, Nakagawa H, et al. Unraveling the complex maternal history of Southern African Khoisan populations. American Journal of Physical Anthropology. 2014 Mar;153(3):435-48. doi:10.1002/ajpa.22441.

[7] Schlebusch CM, Lombard M, Soodyall H. MtDNA control region variation affirms diversity and deep sub-structure in populations from southern Africa. BMC evolutionary biology. 2013 Dec;13(1):1-21. doi:10.1186/1471-2148-13-56.

[8] Barnard A. Hunters and Herders of Southern Africa. Cambridge University Press; 1992. doi:10.1017/CBO9781139166508.

[9] Schlebusch CM, Jakobsson M. Tales of Human Migration, Admixture, and Selection in Africa. Annual Review of Genomics and Human Genetics. 2018;19(1):405-28. doi:10.1146/annurev-genom-083117-021759.

[10] Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, et al. Reconstructing Prehistoric African Population Structure. Cell. 2017 Sep;171(1):59-71.e21. doi:10.1016/j.cell.2017.08.049.

[11] Vicente M, Lankheet I, Russell T, Hollfelder N, Coetzee V, Soodyall H, et al. Male-biased migration from East Africa introduced pastoralism into southern Africa. BMC biology. 2021 Dec;19(1):259-16. doi:10.1186/s12915-021-01193-z.

[12] Wynberg R, Schroeder D, Chennells R. Indigenous Peoples, Consent and Benefit Sharing– Learning Lessons from the San-Hoodia Case. Springer Dordrecht; 2009. doi:10.1007/978-90-481-3123-5.

[13] Schlebusch C. Issues raised by use of ethnic-group names in genome study. Nature. 2010;464(7288):487-7. doi:10.1038/464487a.

[14] De Filippo C, Bostoen K, Stoneking M, Pakendorf B. Bringing together linguistic and genetic evidence to test the Bantu expansion. Proceedings Biological sciences. 2012 Aug;279(1741):3256-63. doi:10.1098/rspb.2012.0318.

[15] Li S, Schlebusch C, Jakobsson M. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. Proceedings Biological sciences. 2014 Oct;281(1793):20141448. doi:10.1098/rspb.2014.1448.

[16] Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science (New York, NY). 2017 May;356(6337):543-6. doi:10.1126/science.aal1988.

[17] Fortes-Lima CA, Burgarella C, Hammarén R, Eriksson A, Vicente M, Jolly C, et al. The genetic legacy of the expansion of Bantu-speaking peoples in Africa. Nature. 2024;625(7995):540-7. doi:10.1038/s41586-023-06770-6.

[18] Bostoen K. The Bantu expansion. In: Oxford research encyclopedia of African history. Oxford University Press; 2018. p. 28. doi:10.1093/acrefore/9780190277734.013.191.

[19] Hollfelder N, Erasmus JC, Hammaren R, Vicente M, Jakobsson M, Greeff JM, et al. Patterns of African and Asian admixture in the Afrikaner population of South Africa. BMC biology. 2020 Feb;18(1):16-3. doi:10.1186/s12915-020-0746-1.

[20] Giliomee H. The Afrikaners: Biography of a people. 2nd ed. Charlottesville: University of Virginia Press; 2009. doi:10.3366/afr.2004.74.2.296.

[21] Nurse GT, Weiner JS, Jenkins T. Weiner JS, Jenkins T, editors. The peoples of southern Africa and their affinities.. Oxford University Press; 1985. doi:10.1017/S0021853700030632.

[22] Adhikari M. Contending Approaches to Coloured Identity and the History of the Coloured People of South Africa. History Compass. 2005 Jan;3(1):****. doi:10.1111/j.1478-0542.2005.00177.x.

[23] Isaacs S, Geduld-Ullah T, Benjeddou M. Reconstruction of major maternal and paternal lineages of the Cape Muslim population. Genet Mol Biol. 2013 Jul;36(2):167-76. doi:10.1590/S1415-47572013005000019.

[24] de Wit E, Delport W, Rugamika CE, Meintjes A, Möller M, van Helden PD, et al. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. Human Genetics. 2010 Aug;128(2):145-53. doi:10.1007/s00439-010-0836-1.

[25] Quintana-Murci L, Harmant C, Quach H, Balanovsky O, Zaporozhchenko V, Bormans C, et al. Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. American journal of human genetics. 2010 Apr;86(4):611-20. doi:10.1016/j.ajhg.2010.02.014.

[26] Petersen DC, Libiger O, Tindall EA, Hardie RA, Hannick LI, Glashoff RH, et al. Complex patterns of genomic admixture within southern Africa. PLoS Genetics. 2013;9(3):e1003309. doi:10.1371/journal.pgen.1003309.

[27] Patterson N, Petersen DC, van der Ross RE, Sudoyo H, Glashoff RH, Marzuki S, et al. Genetic structure of a unique admixed population: implications for medical research. Human molecular genetics. 2010 Feb;19(3):411-9. doi:10.1093/hmg/ddp505.

[28] Vicente M, Jakobsson M, Ebbesen P, Schlebusch CM. Genetic Affinities among Southern Africa Hunter-Gatherers and the Impact of Admixing Farmer and Herder Populations. Molecular Biology and Evolution. 2019 07;36(9):1849-61. doi:10.1093/molbev/msz089.

[29] Prendergast ME, Lipson M, Sawchuk EA, Olalde I, Ogola CA, Rohland N, et al. Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. Science (New York, NY). 2019 Jul;365(6448):eaaw6275. doi:10.1126/science.aaw6275.

[30] Semo A, Gayà-Vidal M, Fortes-Lima C, Alard B, Oliveira S, Almeida J, et al. Along the Indian Ocean Coast: Genomic Variation in Mozambique Provides New Insights into the Bantu Expansion. Mol Biol Evol. 2020 Feb;37(2):406-16. doi:10.1093/molbev/msz224.

[31] The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68 EP. doi:10.1038/nature15393.

[32] Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015 Jan;517(7534):327-32. doi:10.1038/nature13997.

[33] Martin AR, Lin M, Granka JM, Myrick JW, Liu X, Sockell A, et al. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. Cell. 2017 Nov;171(6):1340-53. doi:10.1016/j.cell.2017.11.015.

[34] Scheinfeldt LB, Soi S, Lambert C, Ko WY, Coulibaly A, Ranciaro A, et al. Genomic evidence for shared common ancestry of East African hunting-gathering populations and insights into local adaptation. Proceedings of the National Academy of Sciences of the United States of America. 2019 Mar;116(10):4166-75. doi:10.1073/pnas.1817678116.

[35] Pierron D, Razafindrazaka H, Pagani L, Ricaut FX, Antao T, Capredon M, et al. Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. Proc Natl Acad Sci U S A. 2014 Jan;111(3):936-41. doi:10.1073/pnas.1321860111.

[36] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007 Sep;81(3):559-75. doi:10.1086/519795.

[37] Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics (Oxford, England). 2010 Nov;26(22):2867-73. doi:10.1093/bioinformatics/btq559.

[38] Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC bioinformatics. 2011 Jun;12(1):246-6. doi:10.1186/1471-2105-12-246.

[39] Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics (Oxford, England). 2016 Sep;32(18):2817-23. doi:10.1093/bioinformatics/btw327.

[40] Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLOS Genetics. 2006 12;2(12):1-20. doi:10.1371/journal.pgen.0020190.

[41] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006;38(8):904-9. doi:10.1038/ng1847.

[42] O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genetics. 2014 Apr;10(4):e1004234. doi:10.1371/journal.pgen.1004234.

[43] Salter-Townshend M, Myers S. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. Genetics. 2019 Jul;212(3):869-89. doi:10.1534/genetics.119.302139.

[44] Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009 Sep;461(7263):489-94. doi:10.1038/nature08365.

[45] Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, et al. Genetic evidence for two founding populations of the Americas. Nature. 2015;525(7567):104-8. doi:10.1038/nature14895.

[46] Goldberg A, Rosenberg NA. Beyond 2/3 and 1/3: The Complex Signatures of Sex-Biased Admixture on the X Chromosome. Genetics. 2015 07;201(1):263-79. doi:10.1534/genetics.115.178509.

[47] Rishishwar L, Conley A, Wigington C, Wang L, Valderrama A, Jordan K. Ancestry, admixture and fitness in Colombian genomes. Scientific Reports. 2015 07;5. doi:10.1038/srep12376.

[48] Details omitted for double-blind reviewing.

[49] Schönherr S, Weissensteiner H, Kronenberg F, Forer L. Haplogrep 3 - an interactive haplogroup classification and analysis platform. Nucleic Acids Research. 2023 04;51(W1):W263-8. doi:10.1093/nar/gkad284.

[50] Severson AL, Shortt JA, Mendez FL, Wojcik GL, Bustamante CD, Gignoux CR. SNAPPY: Single Nucleotide Assignment of Phylogenetic Parameters on the Y chromosome. bioRxiv. 2018. doi:10.1101/454736.

[51] McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018;3(29):861. doi:10.21105/joss.00861.

[52] Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American Journal of Physical Anthropology. 2005 Oct;128(2):415-23. doi:10.1002/ajpa.20188.

[53] Atağ G, Somel M. An explanation for the neighbour repulsion phenomenon in Patterson's f-statistics. bioRxiv. 2024. doi:10.1101/2024.02.17.580509.

[54] Layne V. The District Six Museum: An Ordinary People's Place. Public Historian. 2008 02;30:53-62. doi:10.1525/tph.2008.30.1.53.

[55] Dewar N. Seeking closure: conflict resolution, land resolution, and inner city redevelopments in 'DISTRICT SIX' CAPE TOWN. South African Geographical Journal. 2001. doi:10.1080/03736245.2001.9713718.

[56] Cleophas FJ. Writing and contextualising local history. A historical narrative of the Wellington Horticultural Society (Coloured). Yesterday and Today. 2014:21-53.

[57] Byrnes RM. South Africa : a country study /;. "Research completed May 1996.". Available from: `http://lawcat.berkeley.edu/record/185025`.

[58] Vicente M, Priehodová E, Diallo I, Podgorná E, Poloni ES, Černý V, et al. Population history and genetic adaptation of the Fulani nomads: inferences from genome-wide data and the lactase persistence trait. BMC genomics. 2019 Dec;20(1):915-2. doi:10.1186/s12864-019-6296-7.

[59] Breton G, Schlebusch CM, Lombard M, Sjödin P, Soodyall H, Jakobsson M. Lactase Persistence Alleles Reveal Partial East African Ancestry of Southern African Khoe Pastoralists. Current Biology. 2014 Apr;24(8):852-8. doi:10.1016/j.cub.2014.02.041.

[60] Pickrell JK, Patterson N, Loh PR, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa. Proceedings of the National Academy of Sciences. 2014 Feb;111(7):2632-7. doi:10.1073/pnas.1313787111.

[61] Lander F, Russell T. The archaeological evidence for the appearance of pastoralism and farming in southern Africa. PLOS ONE. 2018 Jun;13(6):e0198941. doi:10.1371/journal.pone.0198941.

[62] Kutanan W, Kampuansai J, Srikummool M, Kangwanpong D, Ghirotto S, Brunelli A, et al. Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai–Kadai languages. Human Genetics. 2017 Jan;136(1):85-98. doi:10.1007/s00439-016-1742-y.

[63] Hill C, Soares P, Mormina M, Macaulay V, Clarke D, Blumbach PB, et al. A mitochondrial stratigraphy for island southeast Asia. The American Journal of Human Genetics. 2007 Jan;80(1):29-43. doi:10.1086/510412.

[64] Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, et al. Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. American journal of human genetics. 2012 May;90(5):915-24. doi:10.1016/j.ajhg.2012.04.003.

[65] Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, et al. The dawn of human matrilineal diversity. American journal of human genetics. 2008 May;82(5):1130-40. doi:10.1016/j.ajhg.2008.04.002.

[66] Barbieri C, Vicente M, Rocha J, Mpoloka SW, Stoneking M, Pakendorf B. Ancient substructure in early mtDNA lineages of southern Africa. American journal of human genetics. 2013 Feb;92(2):285-92. doi:10.1016/j.ajhg.2012.12.010.

[67] Rito T, Richards MB, Fernandes V, Alshamali F, Cerny V, Pereira L, et al. The first modern human dispersals across Africa. PLoS ONE. 2013. doi:10.1371/journal.pone.0080031.

[68] Cerezo M, Achilli A, Olivieri A, Perego UA, Gómez-Carballa A, Brisighelli F, et al. Reconstructing ancient mitochondrial DNA links between Africa and Europe. Genome research. 2012 May;22(5):821-6. doi:10.1101/gr.134452.111.

[69] Underhill PA, Kivisild T. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annual review of genetics. 2007;41:539-64. doi:10.1146/annurev.genet.41.110306.130407.

[70] Rosa A, Brehem A. African human mtDNA phylogeography at-a-glance. Journal of anthropological sciences = Rivista di antropologia : JASS. 2011;89:25-58. doi:10.4436/jass.89006.

[71] Gomez F, Hirbo J, Tishkoff SA. Genetic variation and adaptation in Africa: implications for human evolution and disease. Cold Spring Harbor perspectives in biology. 2014 Jul;6(7):a008524. doi:10.1101/cshperspect.a008524.

[72] González AM, Larruga J, Abu-Amero KK, Shi Y, Pestano J, Cabrera M. Mitochondrial lineage M1 traces an early human backflow to Africa. BMC Genomics. 2007 Jul;8:223. doi:10.1186/1471-2164-8-223.

[73] Maji S, Krithika S, Vasulu TS. Phylogeographic distribution of mitochondrial DNA macrohaplogroup M in India. Journal of genetics. 2009 Apr;88(1):127-39. doi:10.1007/s12041-009-0020-3.

[74] Kumar S, Ravuri RR, Koneru P, Urade BP, Sarkar BN, Chandrasekar A, et al. Reconstructing Indian-Australian phylogenetic link. BMC evolutionary biology. 2009 Jul;9(1):173-5. doi:10.1186/1471-2148-9-173.

[75] Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, et al. Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC genetics. 2004 Aug;5(1):26-5. doi:10.1186/1471-2156-5-26.

[76] Schlebusch CM. Genetic variation in Khoisan-speaking populations from southern Africa; 2010.

[77] Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sellitto D, Scozzari R. A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. American journal of human genetics. 2011 Jun;88(6):814-8. doi:10.1016/j.ajhg.2011.05.002.

[78] Batini C, Ferri G, Destro-Bisol G, Brisighelli F, Luiselli D, Sánchez-Diz P, et al. Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. Molecular Biology and Evolution. 2011 Sep;28(9):2603-13. doi:10.1093/molbev/msr089.

[79] Zhong H, Shi H, Qi XB, Xiao CJ, Jin L, Ma RZ, et al. Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. Journal of Human Genetics. 2010 Jul;55(7):428-35. doi:10.1038/jhg.2010.40.

[80] Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, et al. Origin, Diffusion, and Differentiation of Y-Chromosome Haplogroups E and J: Inferences on the Neolithization of Europe and Later Migratory Events in the Mediterranean Area. The American Journal of Human Genetics. 2004;74(5):1023-34. doi:10.1086/386295.

[81] Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CET, et al. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. The American Journal of Human Genetics. 2006;78(2):202-21. doi:10.1086/499411.

[82] Shah AM, Tamang R, Moorjani P, Rani DS, Govindaraj P, Kulkarni G, et al. Indian Siddis: African descendants with Indian admixture. American journal of human genetics. 2011 Jul;89(1):154-61. doi:10.1016/j.ajhg.2011.05.030.

[83] Balaresque P, Bowden GR, Adams SM, Leung HY, King TE, Rosser ZH, et al. A Predominantly Neolithic Origin for European Paternal Lineages. PLOS Biology. 2010;8(1):e1000285. doi:10.1371/journal.pbio.1000285.

[84] Mendez FL, Karafet TM, Krahn T, Ostrer H, Soodyall H, Hammer MF. Increased Resolution of Y Chromosome Haplogroup T Defines Relationships among Populations of the Near East, Europe, and Africa. Human Biology. 2011;83(1):39 53. doi:10.3378/027.083.0103.