Supplementary Information for:

Ancient genomes from southern Africa pushes modern human divergence beyond 260,000 years ago

Carina M. Schlebusch[1,4]*, Helena Malmström[1,4]*, Torsten Günther[1], Per Sjödin[1], Alexandra Coutinho[1], Hanna Edlund[1], Arielle R. Munters[1], Maryna Steyn[2], Himla Soodyall[3], Marlize Lombard[4,5#], Mattias Jakobsson[1,4,6#]

1 Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18C, SE-752 36 Uppsala, Sweden

2 Human Variation and Identification Research Unit, School of Anatomical Sciences, Faculty of Health Sciences, University of the Witwatersrand, South Africa.

3 Division of Human Genetics, Faculty of Health Sciences, University of the Witwatersrand and National Health Laboratory Service, South Africa.

4 Centre for Anthropological Research & Department of Anthropology and Development Studies, University of Johannesburg, P.O. Box 524, Auckland Park, 2006, South Africa

5 Stellenbosch Institute for Advanced Study, (STIAS) Wallenberg Research Centre at Stellenbosch University, Marais Street, Stellenbosch, 7600, South Africa

6 SciLife Lab Uppsala, Sweden

* These authors contributed equally

Correspondence: Mattias Jakobsson (Mattias.jakobsson@ebc.uu.se) and Marlize Lombard (mlombard@uj.ac.za)

Supplement Table of contents

Supplementary Information section 1 - Samples

Overview and geography

We investigate the skeletal material from seven individuals excavated in the KwaZulu-Natal Province along the east coast of South Africa (Figure S1.1). The Ballito Bay A, Ballito Bay B and Doonside individuals were retrieved from the shoreline near the towns of Ballito Bay and Doonside. The skeletal material from Eland Cave and Champagne Castle were excavated from caves in the uKhahlamba-Drakensberg mountain range, and the Mfongosi and Newcastle individuals are from inland KwaZulu-Natal. Five of the individuals were AMS dated in this study and three were dated previously (Ballito Bay B dated twice, see text below) (Ribot et al. 2010). All conventional radiocarbon dates were modelled using OxCal v.4.2 and SHCal13 calibration curves (Ramsey 2009; Hogg et al. 2013).

Ballito Bay A (BBayA): The skeletal remains of Ballito Bay A belongs to a juvenile individual dated to 1986-1831 cal BP (95% probability) (1980+/-20 BP, Pta-5796) (Ribot et al. 2010). The remains were excavated by Schoute-Vanneck and Walsh during the 1960s (KwaZulu-Natal Museum 1960b), first curated at the Durban Museum, and then transferred to the KwaZulu-Natal Museum where it is now curated (accession no. 2009/007). The site from which it was retrieved is said to have been a mound formed by a shell midden overlooking the beach, about 46 m from the high water mark. The skeletal material cannot be directly associated with archaeological material from the site, as clear stratigraphic context is unknown, but the unpublished archaeology includes Early Iron Age pottery (KwaZulu-Natal Museum 1960b). Ribot et al. (Ribot et al. 2010) performed stable isotope analyses and AMS radiocarbon dating. Together with other individuals, they placed this specimen in the cultural context of Later Stone Age populations with δ13C and δ15N values that indicate a diet that included a sizable sea food intake (-14.4‰ and 11.8‰ respectively) (Ribot et al. 2010).

Ballito Bay B (BBayB): The well-preserved Ballito Bay B remains belong to an adult male that were retrieved from a shell midden context about 46 m from the shore near Ballito Bay. The remains that were discovered by Shoute-Vanneck and Walsch during the 1960s (KwaZulu-Natal Museum 1960a), were first curated at the Durban Museum, and then transferred to the KwaZulu-Natal Museum where it is now curated (accession no. 2009/008.001 and 2009/008.002). The skeletal material cannot be directly associated with the archaeological material at the site as clear stratigraphic context is unknown. However, the unpublished archaeology includes some stone artefacts and Early Iron Age pottery. Ribot et al. (Ribot et al. 2010) performed craniometric analyses, stable isotope analyses and AMS radiocarbon dating. They placed this individual in the cultural context of Later Stone Age populations because of morphological similarities to pre-2 kya foragers and to modern groups with Khoe-San ancestry, and because of his large sea-food intake (δ13C -13.6‰ and δ15N 13.3‰) (Ribot et al. 2010). More detailed analyses of the well-preserved skull showed that this individual had a non-lethal cranial injury inflicted by a sharp stone flake (Pfeiffer 2012). We AMS dated a humerus from this individual to 2149-1932 cal BP (95% probability) (2110+/-30 BP, Beta-398217) and obtained a similar stable carbon isotope value (δ13C -13.7‰) to Ribot et al. (2010). However, there is a previous radiocarbon date available of 3209-2880 cal BP (95% probability) (2940 ± 50 BP, Pta-5803) (Ribot et al. 2010). As the genomic data we have produced indicate that the bones and teeth that we have analyzed belong to the same individual (see section 4.1), we use the radiocarbon date obtained by us.

Doonside (DOO): The Doonside individual was morphologically determined to be an adult female (but lacked sufficient genetic data to call the biological sex), is dated to 2296-1910 cal BP (95% probability) (2110+/-50 BP, Pta-5800), and based on isotope values consumed much sea-food (δ13C -13.8‰ and δ15N 14.2‰) (Ribot et al. 2010). Cranio-morphologically, the individual shows extreme flatness and shortness of face characteristic of Khoe-San populations, but falls outside the 90% confidence ellipse of the Khoe-San variation although remaining closest to Khoe-San according to squared Euclidean distances (Ribot et al. 2010). The remains were initially curated at the Durban Museum, then transferred to the KwaZulu-Natal Museum where it is now curated (accession no. 2009/010). Apart from having been found near Doonside on the KwaZulu-Natal coast, there is no additional information about its discovery, and no associated archaeology was recorded.

Eland Cave (ELA): The remains of the Eland Cave individual comprise a complete left first rib, one right rib that had been fractured postmortem, left first metatarsal and the distal half of the left tibia. Its sex could not be determined morphologically due to the absence of the cranium, pelvis and long bones, but she was determined to be female using genetics (this study). The distal epiphysis of the tibia was completely closed, suggesting that she was older than 20 years when she died. Other morphological criteria are consistent with an age estimate of 50+ years. The individual was discovered by Lombard (the then landowner) in 1926 at Eland Cave (previously known as Lombard Cave) in the uKhahlamba-Drakensberg Mountains together with what appeared to be hunter-gatherer artefacts, which were acquired and curated by the KwaZulu-Natal Museum (accession no. 1925/037). The skeletal material cannot be directly associated with archaeological material from the site, as its stratigraphic context is unknown, but the published (Vinnicombe 1971) and unpublished (KwaZulu-Natal Museum 1969) archaeology include a complete hunter-gatherer bow-and-arrow kit, a bone/ivory arm ring, stone artefacts associated with the Smithfield variation of the final Later Stone Age, and a few pottery sherds associated with Iron Age farmers. The site is also associated with extensive rock paintings, which today form part of the uKhahlamba/Drakensberg UNESCO World Heritage Site (Swart 2004; Mazel 2009). We have directly AMS dated the specimen to 533-453 cal BP (95% probability) (480+/-30 BP, Beta-398219) using a tibia, and obtained a δ13C value of -11.7‰.

Mfongosi (MFO): The human remains from the Mfongosi individual consist of the mandible, frontal bone, left and right parietal bones and disarticulated occipital bone of the cranium. The left humerus, left femur and left tibia were also present, but the proximal and distal epiphyses were damaged postmortem. Morphological features indicate a female individual, which is consistent with the DNA results. No age estimate could be made based on morphology, but the long bones appeared to be adult size and all teeth present were permanent with moderate to advanced dental wear, suggesting that this was not a young adult individual. The remains were discovered in the Tugela River valley, and excavated by Jones (the then landowner) from a grave in which the body was buried in a flexed position. It was presented to the KwaZulu-Natal Museum in 1932 where it is now curated (accession no. 1925/036.002). Material associated with the grave included some pottery sherds, the tops of two carved and perforated bone pendants and twelve bone bead fragments (KwaZulu-Natal Museum 1951), which is consistent with Iron Age farmer material culture. We have directly AMS dated the individual to 448-308 cal BP (95% probability) (360+/-30 BP, Beta-398220) using a femur and obtained a δ13C value of -9.4‰.

Newcastle (NEW): The human remains from the Newcastle individual comprised of a fragmented right parietal bone, left temporal bone, left inferior-lateral orbital rim fragment, right superior-lateral orbital rim fragment, both claviculae, both scapulae, left and right os coxae, both patellae, right fibula, four cervical vertebrae, four thoracic vertebrae, five lumbar vertebrae, manubrium, six right ribs, seven left ribs, one rib fragment, 10 foot bones along with 10 metatarsals and 10 foot phalanges, nine metacarpals and eight hand phalanges, as well as three hand bones. Morphological features suggest a female individual, which was confirmed by the DNA analysis. The presence of a well-developed preauricular sulcus suggests that this woman has had at least one child. Vertebral osteophytes were noticed on the lumbar, thoracic and cervical vertebrae, with the lumbars most severely affected, and together with other features, the remains seem to indicate a middle-aged female (probably around 40 to 60 years) of short stature. The remains were discovered by employees of the Drakensville Berg Resort at Oliviershoek near Newcastle in a disturbed grave, from which the skull and most of the long bones were removed. Van de Venter and Van Heerden (Van Heerden and Van de Venter 2002) excavated the remaining bones in 2002, and they are now curated at the KwaZulu-Natal Museum (accession no. 2007/006.001). Associated archaeological material includes two burial stones, one of which is a later Iron Age lower grinding stone. Other archaeological findings close by, but not directly associated with the grave, include stone walling, a stone cairn, some late white rock art, hunter-gatherer rock paintings and some Middle and Later Stone Age artefacts (Van Heerden and Van de Venter 2002; KwaZulu-Natal Museum 2011). We have directly AMS dated the individual using the petrous portion of a temporal bone fragment to 508-327 cal BP (95% probability) (430+/-30 BP, Beta-398221), and obtained a δ13C value of -7‰.

Champagne Castle (CHA): The Champagne Castle remains include a complete cranium and mandible, right humerus, right ulna, right radius, left pubic symphysis, right os coxa (damaged postmortem), one cervical vertebra and a few small fractured bone pieces. The presence of a preauricular sulcus and very wide greater sciatic notch indicate a female, which is confirmed by the DNA analysis. The presence of this pre-auricular sulcus indicates that this individual had most probably borne at least one child during her lifetime. Based on a range of morphological criteria, it is concluded that the remains are most likely that of a young adult female (age estimated to be 20 to 30 years) with an estimated stature of about 154 cm. She has a fracture of her right parietal bone which most probably occurred around the time of death, as is evidenced by a green bone response. This fracture most possibly reflects an episode of interpersonal violence, but it cannot be determined if this traumatic injury was the actual cause of death in this case. The skeleton was excavated by Albino in 1945 from Champagne Castle in the uKhahlamba-Drakensberg Mountains and is curated in the KwaZulu-Natal Museum (accession no. 2009/023). The archaeology of the site includes two occupation layers: an upper layer associated with the Iron Age, and a lower layer associated with the Later Stone Age with stone artefacts ascribed to both the Smithfield and Wilton Industries (Albino 1947). We obtained a direct AMS date from a femur of 448-282 cal BP (95% probability) (310+/-30 BP, Beta-398218) for the specimen, and a δ13C value of -11.7‰.

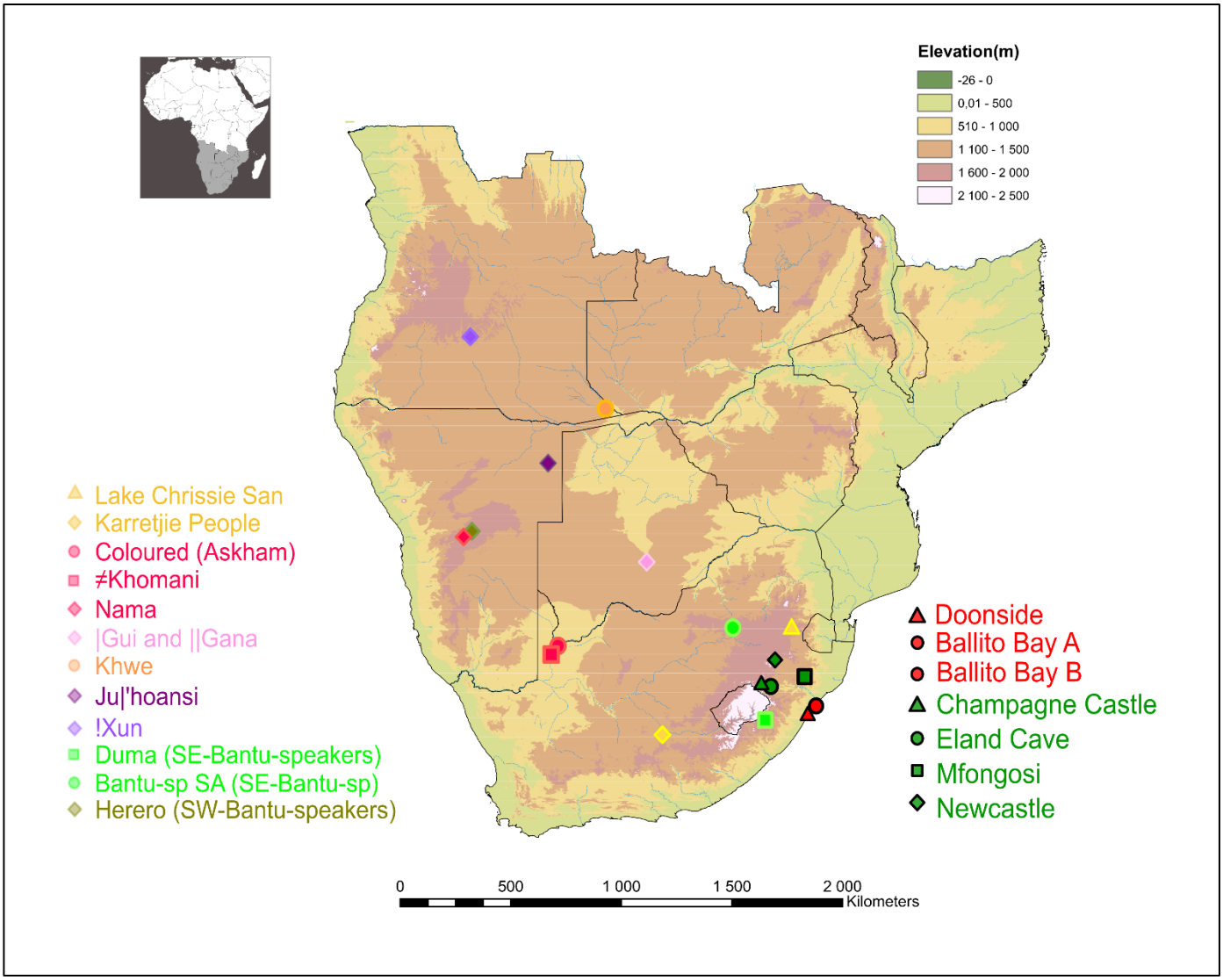

Figure S1.1: Site map of southern Africa. The map shows elevation and the geographic locations of the archaeological sites associated with the investigated ancient individuals, and comparative Khoe-San and Bantu-speaking populations from Schlebusch et al. (Schlebusch et al. 2012).

References

Albino R (1947) Note on the excavation of a rock shelter at Champagne Castle, Natal. Southern African Humanities 11:157-160

Hogg AG, Hua Q, Blackwell PG, Niu M, Buck CE, Guilderson TP, Heaton TJ, Palmer JG, Reimer PJ, Reimer RW, Turney CS (2013) SHCal13 Southern Hemisphere calibration, 0–50,000 years cal BP. Radiocarbon 55:1-15

KwaZulu-Natal Museum (1951) Unpublished site record for national site number: 2830DB 023. Archaeology Department, Pietermaritzburg

KwaZulu-Natal Museum (1960a) Unpublished site record for national site number: 2931CA 047. Archaeology Department, Pietermaritzburg

KwaZulu-Natal Museum (1960b) Unpublished site record for national site number: 2931CA 116. Archaeology Department, Pietermaritzburg

KwaZulu-Natal Museum (1969) Unpublished site record for national site number: 2929AB 022. Archaeology Department, Pietermaritzburg

KwaZulu-Natal Museum (2011) Unpublished site record for national site number: 2829CA 006. Archaeology Department, Pietermaritzburg

Mazel AD (2009) Unsettled times: shaded polychrome paintings and hunter-gatherer history in the southeastern mountains of southern Africa. Southern African Humanities 21:85-115

Pfeiffer S (2012) Field and technical report: Two disparate instances of healed cranial trauma from the Later Stone Age of South Africa. South African Archaeological Bulletin 67:256-261

Ramsey CB (2009) Bayesian analysis of radiocarbon dates. Radiocarbon 51:337-360

Ribot I, Morris AG, Sealy J, Maggs T (2010) Population history and economic change in the last 2000 years in KwaZulu-Natal, RSA. Southern African Humanities 22:89-112

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, Li S, De Jongh M, Singleton A, Blum MG, Soodyall H, Jakobsson M (2012) Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. Science 338:374-379

Swart J (2004) Rock art sequences in uKhahlamba-Drakensberg Park, South Africa. Southern African Humanities 16:13-35

Van Heerden L, Van de Venter A (2002) Unpublished report on the excavation of human skeletal remains for Drakensville Berg Resort. Amafa aKwazulu-Natali, Pietermaritzburg

Vinnicombe P (1971) A Bushman hunting kit from the Natal Drakensberg. Southern African Humanities 20:611-625

Supplementary Information section 2 - Permission and permits for sampling and export and sampling procedure

Permission was obtained for the sampling of the specimens under curatorial supervision from the Council of the KwaZulu-Natal Museum in a letter from the Assistant Director, Human Sciences, Dr. Carolyn Thorpe. A sampling permit (no 0014/06) was issued to Marlize Lombard under the KwaZulu-Natal Heritage Act No. 4 of 2008 and Section 38 (1) of the National Heritage Resources Act No. 25 of 1999. Also under the latter legislation, permits were issued by the South African Heritage Resources Agency (SAHRA) for the destructive sampling and ancient DNA analyses at Uppsala University, Sweden (permit no 1939), and for sending samples for radiocarbon dating to Beta Analytic, England (permit no 1940). Final reports on the sampling and dating have been submitted to both heritage agencies.

The skeletal material for this paper is curated at the KwaZulu-Natal Museum in Pietermaritzburg, South Africa. The material was provided by the Museum research technician Mudzunga Munzhedzi and the sampling strategy for each specimen was discussed with Dr. Carolyn Thorpe and Dr. Gavin Whitelaw prior to sampling. The sampling for DNA and radiocarbon dating was done by HM and AC on location in October 2014 and a portable ancient DNA laboratory was set up in a separate room at the Museum. Three samples were taken from each individual (or museum accession number), the majority of which were from different bone elements for ancient DNA analyses. The bone elements were UV irradiated (254 nm) for 30 minutes to one hour per side and stored in plastic zip-lock bags until sampled. Further handling of the specimens was done in a bleach-decontaminated (DNA Away, ThermoScientific) enclosed sampling tent with adherent gloves (Captair Pyramide portable isolation enclosure, Erlab). Teeth were wiped with 0.5% bleach (NaOH) and UV-irradiated sterile water (HPLC grade, Sigma-Aldrich). The outer surface was removed by drilling at low speed using a portable Dremel 8100, and between 60 and 200 mg of bone powder was sampled for DNA analyses from the interior of the bones and teeth. All plastics and equipment used had been decontaminated with DNA-away and/or UV irradiation prior to their use. The researchers wore full-zip suits with caps, face-masks with visors and double latex gloves and the tent was frequently cleaned with DNA-away during sampling. Five of the individuals were sampled for AMS radiocarbon dating either through cutting off a small piece of bone (1.8-4 cm) or through drilling out bone powder (600-750 mg). The sampled bone elements were directly returned to the Museum and the ancient DNA samples and radiocarbon samples were transported to the Ancient DNA Laboratory at Uppsala University, Sweden. The radiocarbon samples were sent to Beta Analytic for AMS dating.

Supplementary Information section 3 - aDNA laboratory procedures

The 1.5 ml tubes containing the bone powder samples were thoroughly wiped with DNA-away before they were taken into the dedicated ancient DNA clean room facility at Uppsala University. The laboratory is equipped with, among other things, an air-lock between the lab and corridor, positive air pressure, UV lamps in the ceiling (254nm) and HEPA-filtered laminar flow hoods. The laboratory is frequently cleaned with bleach (NaOH) and UV-irradiation and all equipment and non-biological reagents are regularly decontaminated with bleach and/or DNA-away (ThermoScientific) and UV irradiation.

DNA was extracted from between 60 and 190 mg of bone powder using silica-based protocols, either as in Yang et al. (Yang et al. 1998) with modifications as in Malmström et al. (Malmström et al. 2007) or as in Dabney et al. (Dabney et al. 2013), and were eluted in 50-110 µl Elution Buffer (Qiagen) (Table S3.1). The collected bone powder was in some cases subdivided to enable more than one extraction from each original tube. Between 3 and 6 DNA extracts were made for each individual (or accession number) and one negative extraction control was processed for every 4 to 7 samples extracted. Indexed DNA libraries were prepared from 20 µl of extract using either a blunt-end protocol and P5 and P7 adapters as in Meyer and Kircher (Meyer and Kircher 2010) and Günther et al. (Günther et al. 2015) with the shearing step omitted or with a "damage-repair" protocol that repair post-mortem deaminated sites using Uracil-DNA-glycosylase (UDG) and endonuclease VIII (endo VIII) (Briggs and Heyn 2012) (Table S3.1). Between one and five libraries were prepared from each DNA extract and one negative library control was processed for every 6-8 ancient DNA libraries.

The optimal number of PCR cycles to use for each library was determined using quantitative PCR (qPCR) in order to see at what cycle a library reached the plateau (where it is saturated) and then deducting three cycles from that value. The 25 µl qPCR reactions were set up in duplicates and contained 1 µl of DNA library, 1X Maxima SYBR Green Mastermix and 200 nM of each IS7 and IS8 primers (Meyer and Kircher 2010) and were amplified according to supplier instructions (ThermoFisher Scientific). Each library was then amplified in four or eight reactions using between 12 and 21 PCR cycles. One negative PCR control was set up for every four reactions. Blunt-end reactions were prepared and amplified as in Günther et al. (Günther et al. 2015) using IS4 and index primers from Meyer and Kircher (Meyer and Kircher 2010). Damage-repair reactions had a final volume of 25 µl and contained 4 µl DNA library and the following in final concentrations; 1X AccuPrime Pfx Reaction Mix, 1.25U AccuPrime DNA Polymerase (ThermoFisher Scientific) and 400nM of each the IS4 primer and index primer (Meyer and Kircher 2010). Thermal cycling conditions were as recommended by ThermoFisher with an annealing temperature of 60^{o}C (Meyer and Kircher 2010).

Because the femur from Ballito Bay B yielded low amounts of endogenous human DNA, one blunt-end library was enriched using Mybait Human Whole Genome Capture Kit (MYcroarray) following the manufacturer´s instructions (Mybaits manual version 2.3.1) and amplified as above. For each library, four reactions with identical indexing primers were pooled and purified using AMPure XP Beads (Agencourt). The resulting libraries were quantified either on a TapeStation using a High Sensitivity kit (Agilent Technologies) or using a Bioanalyzer 2100 and a High Sensitivity DNA chip (Agilent Technologies). The

negative controls processed did not yield any DNA and were therefore not sequenced. The DNA libraries were sequenced at SciLife Sequencing Centre in Uppsala using either Illumina HiSeq 2500 with v2 paired-end 125 bp chemistry or HiSeq XTen with paired end 150 bp chemistry. The initial strategy was to screen the DNA extracts to evaluate the endogenous ancient human DNA content by building blunt-end libraries and sequencing each library on either a 1/10th of a HiSeq 2500 lane or on a 1/20th of a HiSeq XTen lane. Additional blunt-end or damage-repair libraries were then built and sequenced and high-quality libraries were sequenced to completion (up to 97% clonality) while libraries with low endogenous contents were sequenced to a lesser extent (average 36% clonality over all libraries).

Table S3.1. The number and types of extractions and libraries for each individual.

Individual	**Accession no**	**Bone element**	**Extract (Yang)** (Yang et al. 1998)	**Extract (Dabney)** (Dabney et al. 2013)	**Library Blunt-end**	**Library Damage-repair**	**Library Mybaits capture**
Ballito Bay A	2009/007	Petrous, left	1	-	1	4	-
	2009/007	Petrous, right	1	-	-	2	-
	2009/007	Premolar, upper left	1	-	1	-	-
Ballito Bay B	2009/008.001	Premolar, lower left	1	-	-	2	-
	2009/008.001	Premolar, lower right	1	-	-	1	-
	2009/008.001	Petrous, left	1	-	1	4	-
	2009/008.002	Femur	2	2	6	5	1
Doonside	2009/010	Humerus	1	1	2	1	-
	2009/010	Femur	-	2	3	1	-
	2009/010	Foot/handbone	-	2	3	1	-
Champaigne Castle	2009/023	Molar, lower left	1	1	4	-	-
	2009/023	Canine, lower left	1	-	2	-	-
	2009/023	Femur	-	1	-	1	-
Newcastle	2007/006.001	Incisor	1	-	-	5	-
	2007/006.001	Premolar	1	1	1	1	-
	2007/006.001	Foot/handbone	3	-	1	11	-
Mfongosi	1925/036.002	Molar	1	-	1	4	-
	1925/036.002	Incisor	1	-	1	4	-
	1925/036.002	Femur	1	1	-	7	-
Eland Cave	1925/037	Tibia	3	-	1	12	-
	1925/037	Foot/handbone	1	1	2	4	-

References

Briggs AW, Heyn P (2012) Preparation of next-generation sequencing libraries from damaged DNA. Methods Mol Biol 840:143-154

Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, Garcia N, Paabo S, Arsuaga JL, Meyer M (2013) Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci U S A 110:15758-15763

Günther T, Valdiosera C, Malmström H, Urena I, Rodriguez-Varela R, Sverrisdottir OO, Daskalaki EA, Skoglund P, Naidoo T, Svensson EM, Bermudez de Castro JM, Carbonell E, Dunn M, Stora J, Iriarte E, Arsuaga JL, Carretero JM, Gotherström A, Jakobsson M (2015) Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. Proc Natl Acad Sci U S A 112:11917-11922

Malmström H, Svensson EM, Gilbert MT, Willerslev E, Gotherström A, Holmlund G (2007) More on contamination: the use of asymmetric molecular behavior to identify authentic ancient human DNA. Mol Biol Evol 24:998-1004

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc 2010:pdb prot5448

Yang DY, Eng B, Waye JS, Dudar JC, Saunders SR (1998) Technical note: improved DNA extraction from ancient bones using silica-based spin columns. Am J Phys Anthropol 105:539-543

Supplementary Information section 4 - aDNA data processing

4.1 Initial data processing

Adapters were trimmed and the pair-end reads of each library were merged if the two reads overlapped at at least 11 base pairs using the script MergeReadsFastQ_cc.py (Kircher 2012). Bwa aln 0.7.13 (Li and Durbin 2010) was then used to map them as single end reads to the human reference genome (hg18 and hg19). Non-default parameters for bwa were -l 16500 -n 0.01 -o 2 (Lazaridis et al. 2014; Skoglund et al. 2014). Reads with less than 10% mismatches to the human reference genome and longer than 35 base pairs were retained for further analysis. To determine biological sex we implemented the method described in Skoglund et al. (Skoglund et al. 2012; Skoglund et al. 2013). It uses reads with a mapping quality of at least 30 and calculates the ratio of reads mapping to the Y chromosome and those reads mapping to both X and Y chromosomes.

Sequence data were then merged on library level to ensure maximal retention of reads using samtools merge tool (v.0.1.19) (Li et al. 2009) before removal of PCR duplicates. Reads with identical start and end positions were identified as PCR duplicates and collapsed using a modified version of FilterUniqSAMCons_cc.py (Kircher 2012) which ensures the random assignment of bases in a 50/50 case. Non-UDG and UDG-treated libraries where then merged per individual. Eight individuals were processed and four of these individuals, one male and three females, had an estimated genome coverage over 6x (Table S4.1). The sequence data generated from the Ballito Bay B remains with accession nos. 2009/008.001 and 2009/008.002 were initially treated as two separate individuals. To investigate whether these two bone fragments belonged to the same individual and/or if they were related to the other ~2,000 year old coastal samples, we analyzed baa001, bab001, bab002 and doo001 using READ (Kuhn et al. 2017). READ calculates the proportion non-matching alleles inside non-overlapping windows and then classifies the samples as unrelated, second-degree, first-degree or identical individuals or twins. The coverage of bab002-dr and doo001-dr did not contain enough overlapping data to estimate kinship. All other comparisons were classified as unrelated except bab001 and bab002, which were identified as the same individual or identical twins (Table S4.2).

Table S4.1: Individual and library information.

Individual	Library	Avg. proportion human	Avg. read length	Genome cov	MT cov	Biological Sex
Ballito Bay A	baa001-dr	0.162	58.4562	11.3586	908.332	XY
Ballito Bay B	bab001-dr	0.024	62.1155	0.877068	59.715	XY
Ballito Bay B	bab002-dr	0.003	61.7779	0.0031745	0.323315	consistent with XY but not XX
Champagne Castle	cha001-dr	0.008	65.9827	0.302505	156.993	XX
Doonside	doo001-dr	0.001	52.6248	0.000833784	0.207255	consistent with XY but not XX
Eland Cave	ela001-dr	0.119	60.9033	9.33034	5621.84	XX
Mfongosi	mfo001-dr	0.085	65.358	6.1	482.422	XX
Newcastle	new001-dr	0.072	55.4451	9.73943	514.755	XX
Ballito Bay A	baa001	0.226	66.3037	1.5861	127.061	XY
Ballito Bay B	bab001-b3e1l1	0.051	74.7823	0.360675	22.6492	XY
Ballito Bay B	bab002	0.005	69.3384	0.00631018	1.2588	XY
Champagne Castle	cha001-b1e1l1	0.019	71.9981	0.0589717	29.6469	XX
Doonside	doo001	0.003	57.7276	0.0120243	2.38717	consistent with XY but not XX
Eland Cave	ela001	0.203	68.5227	3.89613	1975.24	XX
Mfongosi	mfo001-b1e1l1	0.108	71.0634	0.84215	79.3689	XX
Newcastle	new001	0.122	59.2867	0.913927	101.383	XX

Table S4.2 READ results for four samples from the ~2,000-year-old remains. The DNA libraries bab001-dr and bab002-dr indicate that these two bone elements with different museum accession numbers originate from the same individual: Ballito Bay B. The overlapping coverage was not sufficient to calculate kinship for the bab002-dr and doo001-dr samples.

Ind/Sample 1	Ind/Sample 2	Relationship	Z upper	Z lower
baa001-dr	Bab001-dr	Unrelated	NA	-18.4991343501
baa001-dr	bab002-dr	Unrelated	NA	-3.86762077351
baa001-dr	doo001-dr	Unrelated	NA	-2.48933227105
bab001-dr	bab002-dr	IdenticalTwins/SameIndividual	4.08507125493	NA
bab001-dr	doo001-dr	Unrelated	NA	-1.78533153114
bab002-dr	doo001-dr	-	-	-

For population genetic analyses, the ancient shotgun data were merged with comparative data sets (see Section 6.1). For low coverage shotgun data, the SNPs in the ancient samples were called as follows: at each SNP site, a random read with minimum mapping and base quality 30 was drawn and the allelic status at that read was coded to be the hemizygous genotype of the individual (file-formats require diploid genotypes and we use the homozygote code for the record, but the data are treated as hemizygote in all downsteam analyses). Sites showing additional alleles or indels were removed from the data. For non-UDG treated sequence data, all transition sites were coded as missing data to avoid the effect of post-mortem damage. For the sequenced individuals we had both UDG-treated and non-UDG treated libraries. At non-transition sites, a read from either of the two library types were randomly sampled, and for transition sites only reads from damage-repaired libraries were sampled.

The three high coverage ancient individuals used in this study (BBayA, ELA, and NEW) as well as high-coverage reference ancient individuals (Mota and LBK) were subjected to diploid genotype calling. We restricted the genotype calling for BBayA to UDG-treated libraries. Base qualities of all Ts in the first five base pairs of each read and all As in the last five base pairs were set to 2. Picard (http://broadinstitute.github.io/picard/) was used to add read groups to the files and indel realignment was conducted using GATK v3.5.0 (McKenna et al. 2010) and the indels from phase 1 of the 1000 genomes project as references (The 1000 Genomes Project Consortium 2010). Diploid genotypes were called with GATK's UnifiedGenotyper and the following parameters -stand_call_conf 50.0, -stand_emit_conf 50.0, -mbq 30, -contamination 0.02 and --output_mode EMIT_ALL_SITES using dbSNP version 142 as known SNPs. Vcftools (Danecek et al. 2011) was used to extract the relevant SNP positions from the VCF if they were not marked as low quality calls. The alleles from the non-UDG treated Mota were set to missing data for all transition sites and the different data sets were merged using Plink v1.9 (Chang et al. 2015).

4.2 Authentication of DNA sequence data and estimation of mitochondrial contamination.

Ancient DNA sequences have a high frequency of cytosine to thymine (C to T) transitions at the 5' ends and of guanidine to adenine (G to A) at 3' ends due to post mortem deamination (Sawyer et al. 2012). Figure S4.1 show these typical damage patterns for ancient DNA for the non-damage repaired libraries.

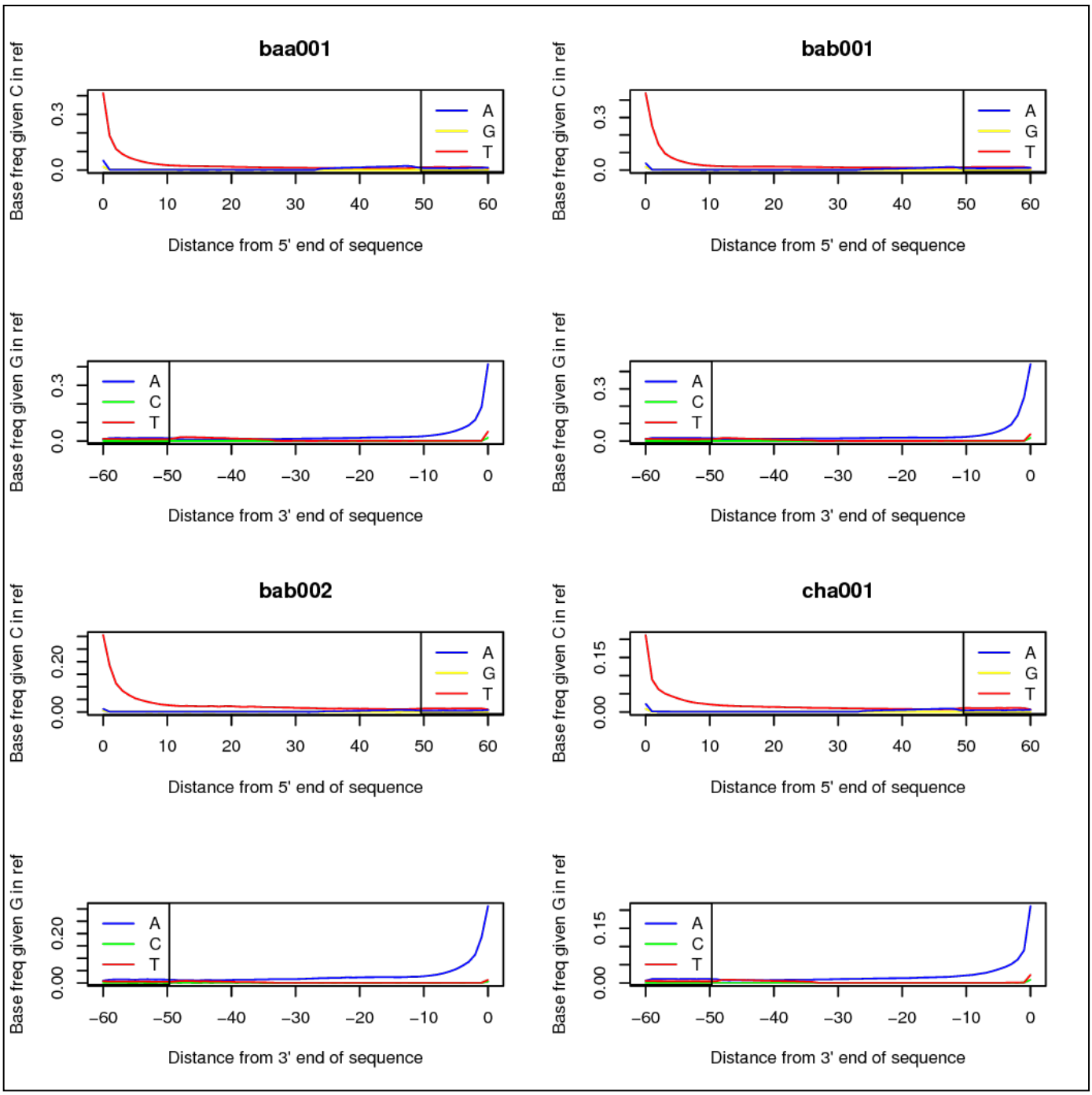

Figure S4.1: Cytosine deamination patterns for non-damage repaired libraries.

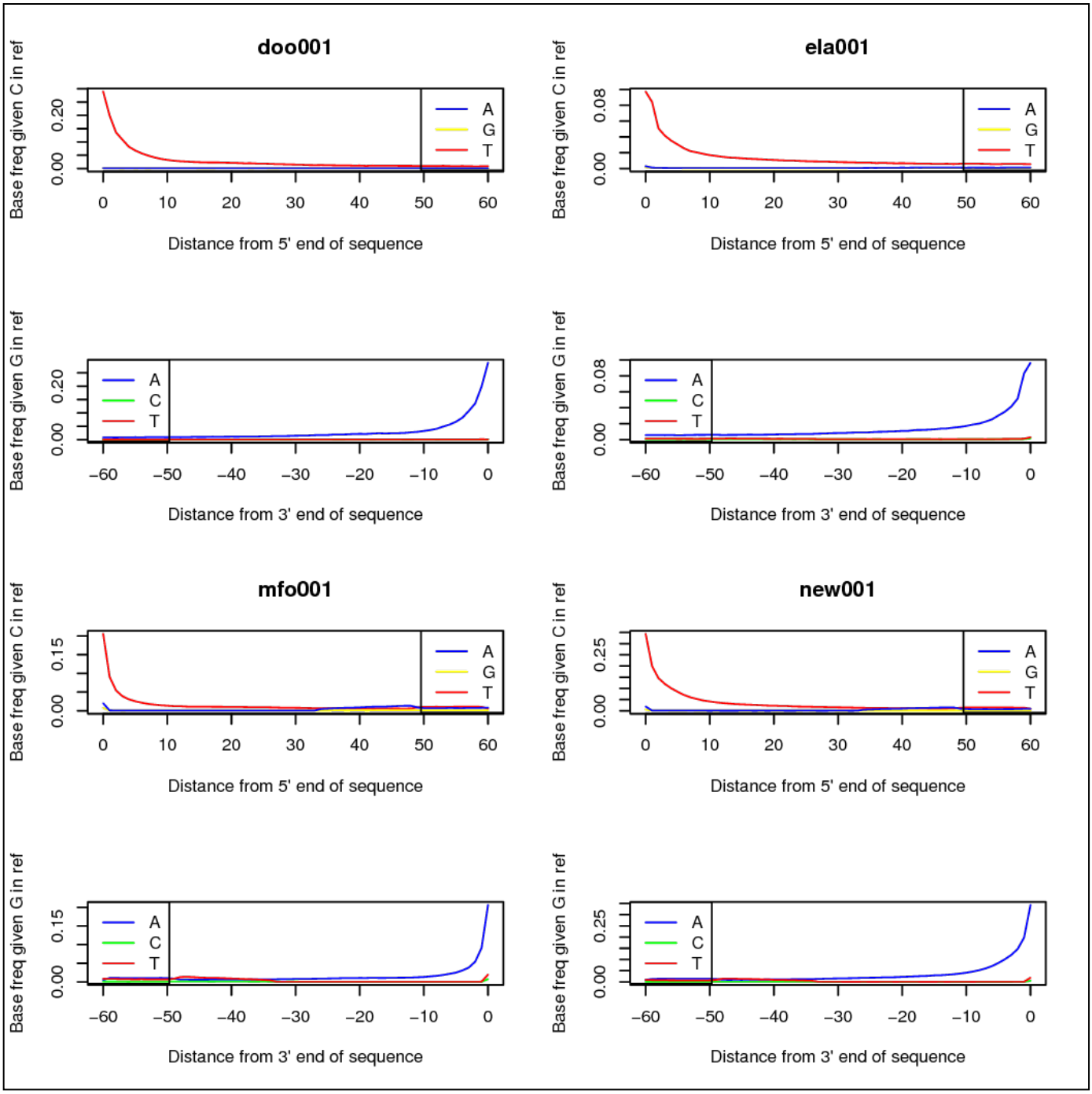

Figure S4.1 (continued): Cytosine deamination patterns for non-damage repaired libraries.

We investigated potential mitochondrial contamination for all samples using the approach of (Green et al. 2008) that utilizes private or near-private consensus alleles in modern-day individuals (<5% in 311 modern mtDNAs), and bases with mapping quality of 30 or higher, as well as a coverage of at least 10x for the ancient DNA data. Positions with a consensus allele of either C or G and where a transition substitution was detected were filtered out to avoid postmortem damage. To obtain a contamination estimate, the counts of consensus

and alternative alleles were added together across all sites (Green et al. 2008). The mitochondrial contamination estimates were less than 4.5% for all ancient individuals (Table S4.3).

Table S4.3: Investigating potential mitochondrial contamination.

Sample	Library	Point estimate (%)	Informative sites	Consensus alleles	Total alleles	Lower C.I.	Higher C.I.
Ballito Bay A	baa001-dr	1.093491124	26	20894	21125	0.9532492265	1.233733022
Ballito Bay B	bab001-dr	3.29847144	21	1202	1243	2.305598806	4.291344074
Ballito Bay B	bab002-dr	-					
Champagne Castle	cha001-dr	1.657940663	25	3381	3438	1.231108364	2.084772963
Doonside	doo001-dr	-					
Eland Cave	ela001-dr	0.146710594	12	38795	38852	0.1086512493	0.1847699388
Mfongosi	mfo001-dr	4.42556996	5	2138	2237	3.573297327	5.277845923
Newcastle	new001-dr	0.7432432432	5	2938	2960	0.4338180001	1.052668486
Ballito Bay A	baa001	0.8980866849	25	2538	2561	0.5327018244	1.263472287
Ballito Bay B	bab001-b3e1l1	1.049868766	17	377	381	0.02641252551	2.073325007
Ballito Bay B	bab002	-					
Champagne Castle	cha001-b1e1l1	1.15384616	17	514	520	0.2359189506	2.071773357
Doonside	doo001	-					
Eland Cave	ela001	0.2278913718	10	15761	15797	0.1535317317	0.3022510119
Mfongosi	mfo001-b1e1l1	0.6153846154	5	323	325	0	1.465635885
Newcastle	new001	0	5	526	526	0	0.5679120981

To estimate errors in the ancient samples coming from sequencing errors, mapping errors and chemical modifications of bases, we used ANGSD's (Korneliussen et al. 2014) error estimation procedure that utilizes an out-group individual (chimpanzee mapped against hg19) and an *ad hoc* "error-free" individual. To generate an "error-free" individual, sequence reads with a mapping quality higher than 35 from a 1000genomes (Auton et al. 2015) CEU male, NA12342, were used. By comparing the quantity of derived alleles in the samples, in relation to the "error-free" individual, to the ancestral state a relative error for each test individual can be calculated. All sequence reads were used, but only sites where ancestral, "error-free" sample, and the target sample have a coverage of ≥1x with a base quality higher than 30 were used for computing the error rate (Fig S4.2). The error rate of Ballito Bay A (0.1%) is on par with previous good-coverage damage repaired data, such as the Loschbour individual from Lazaridis et al. (Lazaridis et al. 2014). The overall error rate for the ancient individuals (with UDG treated sequence data) is around one false positive variant in a thousand called variants, about twice as large as for modern-day DNA samples that show just over one in 2,000 called variants.

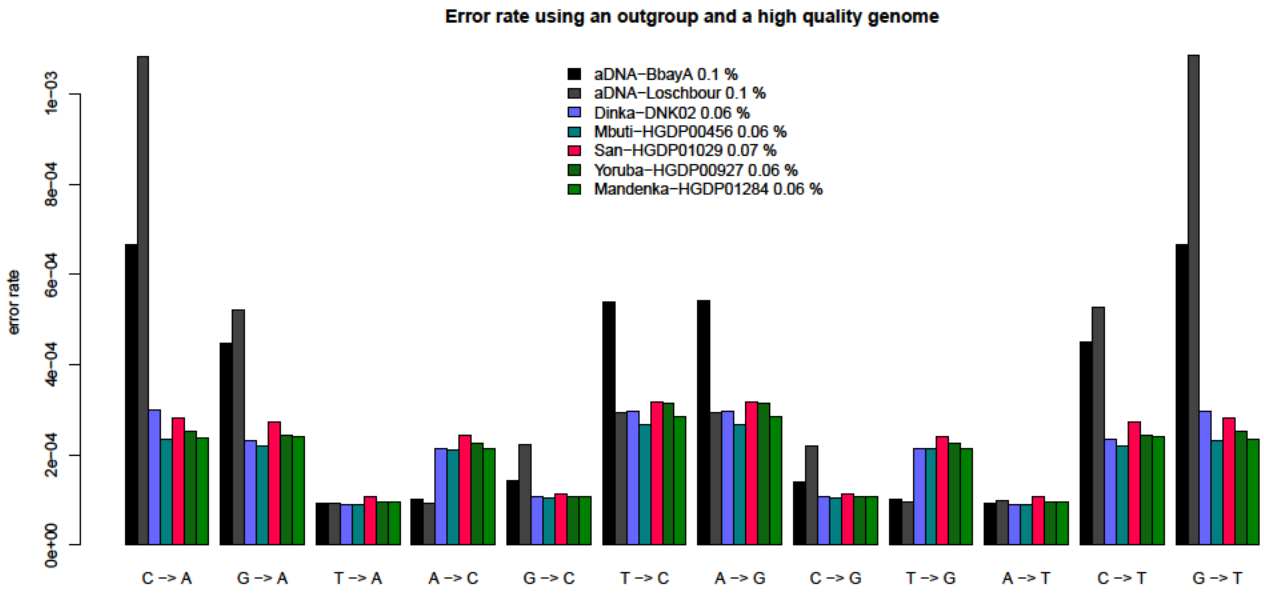

Figure S4.2 Estimated error rates using an outgroup and an "error-free" individual for specific base changes. The average error rate is given in the figure legend.

References

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526:68-74

Bonora M, Wieckowsk MR, Chinopoulos C, Kepp O, Kroemer G, Galluzzi L, Pinton P (2015) Molecular mechanisms of cell death: central implication of ATP synthase in mitochondrial permeability transition. Oncogene 34:1608

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27:2156-2158

Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, Prufer K, Siebauer M, Burbano HA, Ronan M, Rothberg JM, Egholm M, Rudan P, Brajkovic D, Kucan Z, Gusic I, Wikstrom M, Laakkonen L, Kelso J, Slatkin M, Paabo S (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 134:416-426

Kircher M (2012) Analysis of High-Throughput Ancient DNA Sequencing Data. Available from: . In: Shapiro B, Hofreiter M (eds) Ancient DNA (Internet). Vol cited 2017 Jan 20. Humana Press, Totowa, NJ, pp 197–228

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15:356

Kuhn JMM, Jakobsson M, Günther T (2017) Estimating genetic kin relationships in prehistoric populations. bioRxiv 100297

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, et al. (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409-413

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589-595

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303

Sawyer S, Krause J, Guschanski K, Savolainen V, Paabo S (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PLoS One 7:e34131

Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes 9:88

Skoglund P, Malmström H, Omrak A, Raghavan M, Valdiosera C, Günther T, Hall P, Tambets K, Parik J, Sjögren K-G, Apel J, Willerslev E, Storå J, Götherström A, Jakobsson M (2014) Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers. Science Published online 24 April 2014 [DOI:10.1126/science.1253448]

Skoglund P, Malmström H, Raghavan M, Stora J, Hall P, Willerslev E, Gilbert MT, Gotherström A, Jakobsson M (2012) Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 336:466-469

Skoglund P, Storå J, Götherström A, Jakobsson M (2013) Accurate sex identification of ancient human remains using DNA shotgun sequencing. Journal of Archaeological Science 40:4477–4482

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

Supplementary Information section 5 - Uniparental markers

5.1 Y-chromosomes

Samtools v.1.3 (Li et al. 2009) mpileup were used to call single base substitutions from Phylotree (version of 09/03/2016; (van Oven et al. 2014)) from bam files mapped to hg19 (UDG treated data only). Sites with mapping quality and base quality of at least 30 were extracted. Insertions, deletions and sites with chimeric alleles were excluded. Transition sites and A>T and G>C SNPs were kept, to maximize the number of haplogroup defining substitutions. All derived states in the hierarchal phylogeny as well as all ancestral states downstream of the last of the derived alleles within the branch, are reported to show the certainty of the haplogroup call. Additionally, we double-checked that there were no ancestral alleles upstream (in the hierarchal phylogeny) of the defined haplogroup that would contradict the call. The nomenclature of the International Society of Genetic Geneaology (ISOGG) version 11.224 (http://isogg.org) was used. The definitions in the minimal reference phylogeny of Phylotree (http://www.phylotree.org/Y/tree/) were used for sites not present in ISOGG.

The Ballito Bay A boy belongs to Y-chromosomal haplogroup A1b1b2, as supported by 12 derived allele states (Table S5.1). The further downstream subtype is, however, unclear as two additional sites displayed derived substitution for M51 and M118 defining A1b1b2a and A1b1b2b1, respectively. We also note that this individual is ancestral for M13 and M201, both defining haplogroup A1b1b2b. As ancient individuals may belong to branches not found among extant populations (Kivisild 2017), Ballito Bay B could possibly represent an ancient hitherto unknown sub-branch of A1b1b2. Some additional sites displayed derived alleles that do not fit within the A1 phylogeny and they were cross-checked against an updated version of ISOGG (version 11.325, updated 19 November 2016) (the minimal reference phylogeny in Phylotree had not been updated since our last check). These were M236 (G>C) (B1 according to ISOGG), M10072 (G>A) (S according to ISOGG and M2 according to Phylotree), CTS4385 (A>T) and R-Y40 (C>T) (R1a according to Phylotree, marker not present in ISOGG). However, as they were sporadically shattered over the phylogeny, and as multiple upstream sites displayed ancestral states, they may be false positives resulting from strand misidentifications, sequencing errors or postmortem deaminations (Briggs et al. 2007).

Table S5.1: The Y-chromosome haplogroup support for Ballito Bay A including markers, their position in hg19 and information about the mutations.

Hg ISOGG	SNP/ marker	RefSNP ID	Position hg19	Mutation	Obs. allele	No. reads	Allele state
A0-T	L1085	-	2790726	T>C	C	13	derived
A0-T	L1130	-	16661010	T>G	G	14	derived
A (Investigation)	PK1	rs373116908	22583507	C>A	A	7	derived
A1	V168	rs191505182	17947672	G>A	A	2	derived
A1	V171	rs2524861	4898665	C>G	G	5	derived
A1b	P108	-	15426248	C>T	T	3	derived
A1b	V221	rs188292317	7589303	G>T	T	11	derived
A1b1	L419	rs111762602	15204887	G>A	A	5	derived
A1b1b	M32	-	21740436	T>C	C	5	derived
A1b1b2	M144	rs2032619	21925500	T>C	C	4	derived
A1b1b2	M190	rs2032603	14968527	A>G	G	8	derived
A1b1b2	P289	rs372246020	8467082	C>G	G	2	derived
A1b1b2a	M51	rs34078768	21868863	G>A	A	7	derived
A1b1b2b1	M118	-	21763965	A>T	T	9	derived
A1b1b2b	M13	rs3904	21722098	G>C	G	6	ancestral
A1b1b2b	M202	rs2032649	15029492	T>G	T	3	ancestral

Similar to Ballito Bay A, the Ballito Bay B male belongs to haplogroup A1b1b2, and likely even to A1b1b2b1 (Table S5.2). Seven markers displaying derived alleles support the former as does A1b1b2b1-M118 for the latter. No ancestral sites were found downstream of A1b1b2b1. Two additional sites displayed derived alleles, namely M236 and M10072. These were also observed in Ballito Bay A; see above for possible explanations of these discrepant alleles.

Table S5.2: The Y-chromosome haplogroup support for Ballito Bay B including markers, their position in hg19 and information about the mutations.

Hg ISOGG	SNP/ marker	RefSNP ID	Position hg19	Mutation	Obs. allele	No. reads	Allele state
A0-T	L1085	-	2790726	T>C	C	1	derived
A0-T	L1130	-	16661010	T>G	G	1	derived
A1	V171	rs2524861	4898665	C>G	G	1	derived
A1b1b	M32	-	21740436	T>C	C	1	derived
A1b1b2	M144	rs2032619	21925500	T>C	C	1	derived
A1b1b2	M190	rs2032603	14968527	A>G	G	1	derived
A1b1b2	P289	rs372246020	8467082	C>G	G	1	derived
A1b1b2b1	M118	-	21763965	A>T	T	1	derived

Haplogroup A is the oldest Y-chromosomal lineage with an estimated age of circa 150,000 years (Karmin et al. 2015). The sub-haplogroup A1b1b2a (A-M51, previously known as A3b1), is together with A1b1a (A-M14, previously known as A2), common among southern African Khoe-San populations while being rare in Bantu-speaking populations (Naidoo et al. 2010; Schlebusch 2010; Barbieri et al. 2016). The only other ancient Y-chromosomal data available to date from Africa, is the 4,500-year-old Mota hunter-gatherer from Ethiopia, who belonged to haplogroup E1b1 (Gallego Llorente et al. 2015).

5.2 Mitochondrial DNA

Consensus sequences were generated using samtools' mpileup and vcfutils.pl (and vcf2fq) (v0.1.19, (Li et al. 2009)) coupled with ANGSD (Korneliussen et al. 2014). A minimum base quality and mapping quality score of 30 and a coverage of at least three sequence reads were used to call the consensus sequences. Haplogroups were assigned to the sequences using HaploFind (Vianello et al. 2013) and PhyloTree mtDNA Build 17 (18 Feb 2016) (van Oven and Kayser 2009). Doonside had low mtDNA coverage (2.6x) and the haplogroup was called manually from PhyloTree without restrictions on coverage. The variants are reported against the Reconstructed Sapiens Reference Sequence, RSRS (Behar et al. 2012). The mitochondrial coverage, haplogroups, variants supporting the called haplogroup and private variants are reported in Table S5.3. There were a few regions where none of the consensus sequences had any data after filtering. The majority of these positions are situated between the ND1 and CO3 genes and have previously been reported as regions that are difficult to map when working with short sequence reads (Marinov et al. 2014). We noted that HaploFind was not well-adjusted to L0-lineages and therefore we manually curated our variant table to fit the phylogeny in PhyloTree (Build 17, 18 Feb 2016). Haplofind assigned haplotypes correctly, but reported that several variants were missing from the L0d-haplotypes although some of these missing variants defined haplotypes within L1'2'3'4'5'6 and not within L0 (e.g. 146T, 182T, 10664C, 10915T, 11914G, 13276A, 16230A). These errors have been reported to HaploFind.

The mitochondrial genome of the Ballito Bay A boy (BBayA) has 40 variants leading to L0d2c1 (Table S5.3). There are five other variants associated with this haplotype. Two of them were ancestral in this sequence (BBayA lacked a deletion at np 498 and a transition at np 8251) and for the remaining three sites (nps 4204, 4232 and 7154), there were not enough high-quality sequence data. This individual has six additional variants (T4312C, A4732G, T7256C, T8655C, G8701A and A16129G), that are present in between one and 18 other haplotypes in Phylotree. It is not likely that these variants are caused by post-mortem deamination as i) the majority of the data are based on UDG-treated DNA libraries in which the majority of these types of damages are removed, and ii) only one of the sites comprised of a G to A transition.

The Ballito Bay B (BBayB) male belongs to L0d2a1 and displays 33 of 42 expected variants for this haplotype (Table S5.3). Three of the sites have the ancestral allele (i.e. no deletion at np 498 and no transitions at nps 7154 and 8392), while there were not enough data for the remaining six sites (nps 4025, 4044, 4225, 4232, 5153 and 6815). BBayB displays two additional variants; T310C and T16187C. The former has not previously been found in any L-haplotypes but the latter is recurrently found within L-lineages, including in L0d2a1b. BBayB is, however, ancestral for C463T and T7861C (which together with T16178C defines L0d2a1b).

Due to the low coverage, the Doonside (DOO) consensus contained several sites displaying C to T transitions likely caused by post-mortem damage and positions lacking data. We could not use HaploFind to call the haplotoype for this individual. Instead we manually investigate what ancestral and derived states the DOO mt data displayed for haplogroup defining positions, first within the L0 lineage, and then following L1'2'3'4'5'6 lineage leading to all other non-L0 lineages. We conclude that DOO belongs to haplogroup L0d2 as only derived states were present for defining positions leading to this haplogroup. The derived states were G263A, C1048T, C3516A, T5442C, T6185C, C9042T, A9347G, A12720G (leading to L0); G1438A, G8251A, T12121C, G15466A, G15930A, T15941C, T16243C (leading to L0d); A3756G, G9755A, T16278C (leading to L0d1'2); and T11854C, A15766G (leading to L0d2) (Table S5.3). DOO further displayed ancestral alleles for the downstream lineages L0d2a'b'd (16212A), L0d2a (12172A), L0d2b (1386T, 9932G, 10084T, 16069C, 16169C), L0d2d (125T, 127T, 188A, 8434C, 9254A, 9476A, 10745C, 14094T), L0d2c (294T, 4937T, 6644C, 8420A, 9230T, 9305G, 13827A, 14007A, 15346G) and L0d3 (721T, 1243T, 2755A, 5460G, 6377C, 8459A, 9027C, 9488C, 11061C, 13359G, 15236A, 15312T, 16290C, 16300A) with the exception of a few positions with derived alleles (G16390A found in L0d2a, C152T found in L0d2b, C150T found in L0d2d and L0d3). It is highly unlikely that DOO would belong to a non-L0 lineage as it displayed ancestral alleles for L1'2'3'4'5'6 (146C, 182C, 10664T, 13276G), L2'3'4'5'6 (2758A, 2885C, 8468T) and L2'3'4'6 (195T, 247A, 10688A, 13105G, 13506T, 15301G, 16129A) and only derived states at seven positions (10915T, 16230A, 152T, 8655C, 10810T, 16187C, 16189T) which may largely be due to low read coverage at the positions combined with post-mortem damage. As the DOO consensus is uncertain, only the derived states leading to L0d2 are reported in Table S5.3.

The Champagne Castle (CHA) female has 36 of 43 variants leading to L0d2a1a (Table S5.3). This individual displayed the ancestral state for one variant (i.e. did not have a deletion at np 498, similar to BBayA and BBayB) and the remaining six sites lacked high-quality sequence data (np 4025, 4044, 4225, 4232, 7154 and 8392). There were three additional variants present in this mitochondrial genome. The C11881T and G15077A transitions are present in three other haplotypes, while the T16093C transition is highly recurrent, and was present in over 50 haplotypes dispersed over the PhyloTree mitochondrial phylogeny.

The Eland Cave (ELA) female displays all 51 of the expected variants leading to L3e3b1 (Table S5.3). This individual has two additional variants, G10373A and T15071C, which are also found elsewhere in the phylogeny (in 12 other haplotypes and one other haplotype in PhyloTree, respectively).

The Mfongosi (MFO) female belongs to L3e1b2 (Table S5.3). This individual has 46 of the expected variants for this haplotype but displays the ancestral state for the remaining two sites (i.e. does not have a deletion at np 16325 and lacks the C16327T transition). In addition, there are four private variants (T310C, T15115C, C16239T and C16519T) that are also present in different haplotypes in PhyloTree.

The Newcastle (NEW) female displays all 42 expected variants for L3e2b1a2. NEW has two additional transitions, G4769A and T15721C, and an additional transversion mutation, A16183C.

The three approximately 2,000-year-old individuals from the coastal region of eastern South Africa, Ballito Bay A, Ballito Bay B and Doonside belong to L0d lineages (L0d2c, L0d2a1 and L0d2, respectively). The deepest split in the mitochondrial phylogeny is between L0 and L1'2'3'4'5'6 (which comprise all other haplogroups) (Behar et al. 2008; Chan et al. 2015). This lineage is highly divergent and common in the Khoe-San populations of southern Africa (Vigilant et al. 1991; Chen et al. 2000; Tishkoff et al. 2007; Behar et al. 2008; Schlebusch et al. 2013). The L0d2c haplogroup, found in the Ballito Bay A individual, is most common in present-day Nama and ≠Khomani (12%-14%) from Namibia and South Africa, but it is also found at lower frequencies in other Khoe-San populations and in Coloured populations (Schlebusch et al. 2013) and has recently been identified in some Bantu-speaking populations (Chan et al. 2015). Furthermore, a 2,330-year-old forager skeleton from St. Helena Bay on the south west coast of South Africa, displays the sub-haplogroup L0d2c1c (Morris et al. 2014). L0d2a is more frequently observed in present-day populations than L0d2c, and this haplogroup is carried by Ballito Bay B. The highest frequency is found in the Karretjie People (60%), ≠Khomani (33%) and Nama (21%) (Schlebusch et al. 2013). L0d2a is further found in Bantu-speaker populations (12%), in Coloured populations and in 'Baster' (Schlebusch et al. 2013; Chan et al. 2015). This potential Khoe-San maternal contribution into non-Khoe-San populations can be observed in one of the younger, 300-500-year-old, individuals (the Champagne Castle female, who carries an L0d2a1a mitochondria and an otherwise typical Bantu-speaker genomic signature).

Three of the younger individuals (dated to ~300-500 BP); Eland Cave, Mfongosi and Newcastle, belong to L3e-lineages (L3e3b1, L3e1b2 and L3e2b1a2, respectively). L3 lineages are frequent in modern-day individuals in East Africa and the L3e lineage, the most frequent of the L3 lineages, is common in Central/West African groups, and has been suggested to have reached southern Africa with the Bantu expansion ~1,800 years ago (Bandelt et al. 2001; Salas et al. 2002; Torroni et al. 2006; Soares et al. 2012). These lineages are generally absent in Khoe-San populations, with the exception of the Khwe, but are common among many present-day Bantu-speaking populations (Schlebusch et al. 2013). The 4,500-year-old 'Mota' hunter-gatherer from Ethiopia also carries an L3-lineage, L3x2a (Gallego Llorente et al. 2015).

Table S5.3: Mitochondrial coverage, haplogroup assignment, polymorphisms supporting the assigned haplogroup, variants associated with assigned haplogroup that either display the ancestral allele (sites/ancestral) or for which no data are available (sites/no data), and private variants in the ancient African consensus sequences are shown.

Individual	**Mt coverage**	**Mt hg**	**Polymorphisms for called hg (against RSRS)***	**sites/ ancestral**	**sites/ no data**	**private variant**
Ballito Bay A	1035	L0d2c1	263A 294A 1048T 1438A 3516A 3756G 3981G 4025T 4038G 4044G 4937C 5442C 6185C 6249A 6644T 6815C 8113A 8152A 8284T 8420G 9042T 9230C 9305A 9347G 9755A 10589A 11854C 11974G 12007A 12121C 12720G 13827G 14007G 15346A 15466A 15766G 15930A 15941C 16243C 16278C	498del 8251A	4204C 4232C 7154G	4312C 4732G 7256C 8655C 8701A 16129G
Ballito Bay B	84	L0d2a1	198T 263A 597T 1048T 1438A 3516A 3756G 3981G 5442C 6185C 8113A 8152A 8251A 9042T 9347G 9755A 10589A 11854C 12007A 12121C 12172G 12234G 12720G 12810G 14221C 15466A 15766G	498del 7154G G8392A	4025T 4044G 4225G 4232C	310C 16187C

			15930A 15941C 16212G 16243C 16278C 16390A		5153G 6815C	
Doonside	2.6	L0d2	263A 1048T 1438A 3516A 3756G 5442C 6185C 8251A 9042T 9347G 9755A 11854C 12121C 12720G 15466A 15766G 15930A 15941C 16243C 16278C			
Champagne Castle	186	L0d2a1a	198T 263A 597T 1048T 1438A 3516A 3756G 3981G 5153G 5442C 6185C 6815C 8113A 8152A 8251A 8545A 9042T 9347G 9755A 10589A 11854C 12007A 12121C 12172G 12234G 12720G 12810G 14221C 15466A 15766G 15930A 15941C 16212G 16243C 16278C 16390A	498del	4025T 4044G 4225G 4232C 7154G 8392A	11881T 15077A 16093C
Eland Cave	7597	L3e3b1	146T 150T 152T 247G 750A 769G 825T 1018G 2000T 2352C 2758G 2885T 3594C 4104A 4312C 4655A 5262A 6261A 6524C 7146A 7256C 7521G 8468C 8655C 9554A 10664C 10667C 10688G 10810T 10816G 10819G 10915T 11914G 12248G 13101C 13105A 13197T 13276A 13506C 13650C 13651G 14212C 15301A 15812A 16129G 16187C 16189T 16230A 16265T 16278C 16311T			10373A 15071C
Mfongosi	562	L3e1b2	146T 150T 152T 185A 189G 195T 247G 769G 825T 1018G 2352C 2758G 2885T 3594C 4104A 4312C 6587T 7146A 7256C 7521G 8468C 8577G 8655C 10664C 10688G 10810T 10819G 10915T 11914G 12192A 13105A 13276A 13506C 13650C 14152G 14212C 14926G 15301A 15670C 15942C 16129G 16187C 16189T 16230A 16278C 16311T	16325del 16327T		310C 15115C 16239T 16519T
Newcastle	616	L3e2b1a2	146T 150T 152T 247G 769G 825T 1018G 2352C 2483C 2758G 2885T 3277A 3594C 4104A 4312C 7146A 7256C 7521G 8468C 8655C 9377G 10664C 10688G 10810T 10819G 10915T 11914G 12406A 13105A 13276A 13506C 13650C 14212C 14905A 15301A 16129G 16172C 16187C 16230A 16278C 16311T 16320T			4769A 15721C 16183C

* Only derived states within L0d2 are displayed for Doonside, see Section 5.2 for more details.

References

Bandelt HJ, Alves-Silva J, Guimaraes PE, Santos MS, Brehm A, Pereira L, Coppa A, Larruga JM, Rengo C, Scozzari R, Torroni A, Prata MJ, Amorim A, Prado VF, Pena SD (2001) Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. Ann Hum Genet 65:549-563

Barbieri C, Hubner A, Macholdt E, Ni S, Lippold S, Schroder R, Mpoloka SW, Purps J, Roewer L, Stoneking M, Pakendorf B (2016) Refining the Y chromosome phylogeny with southern African sequences. Hum Genet 135:541-553

Behar DM, van Oven M, Rosset S, Metspalu M, Loogvali EL, Silva NM, Kivisild T, Torroni A, Villems R (2012) A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet 90:675-684

Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, Rosset S (2008) The dawn of human matrilineal diversity. Am J Hum Genet 82:1130-1140

Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prufer K, Meyer M, Krause J, Ronan MT, Lachmann M, Paabo S (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci U S A 104:14616-14621

Chan EK, Hardie RA, Petersen DC, Beeson K, Bornman RM, Smith AB, Hayes VM (2015) Revised timeline and distribution of the earliest diverged human maternal lineages in southern Africa. PLoS One 10:e0121223

Chen YS, Olckers A, Schurr TG, Kogelnik AM, Huoponen K, Wallace DC (2000) mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. Am J Hum Genet 66:1362-1383

Gallego Llorente M, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, Curtis MC, Stock JT, Coltorti M, Pieruccini P, Stretton S, Brock F, Higham T, Park Y, Hofreiter M, Bradley DG, Bhak J, Pinhasi R, Manica A (2015) Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science 350:820-822

Karmin M, Saag L, Vicente M, Wilson Sayres MA, Jarve M, Talas UG, Rootsi S, et al. (2015) A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res 25:459-466

Kivisild T (2017) The study of human Y chromosome variation through ancient DNA. Hum Genet

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15:356

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079

Marinov GK, Wang YE, Chan D, Wold BJ (2014) Evidence for site-specific occupancy of the mitochondrial genome by nuclear transcription factors. PLoS One 9:e84713

Morris AG, Heinze A, Chan EK, Smith AB, Hayes VM (2014) First ancient mitochondrial human genome from a prepastoralist southern African. Genome Biol Evol 6:2647-2653

Naidoo T, Schlebusch CM, Makkan H, Patel P, Mahabeer R, Erasmus JC, Soodyall H (2010) Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. Investig Genet 1:6

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A (2002) The making of the African mtDNA landscape. Am J Hum Genet 71:1082-1111

Schlebusch CM (2010) PhD Thesis: Genetic variation in Khoisan-speaking populations from southern Africa. University of the Witwatersrand, Johannesburg

Schlebusch CM, Lombard M, Soodyall H (2013) MtDNA control region variation affirms diversity and deep sub-structure in populations from Southern Africa. BMC Evol Biol 13:56

Soares P, Alshamali F, Pereira JB, Fernandes V, Silva NM, Afonso C, Costa MD, Musilova E, Macaulay V, Richards MB, Cerny V, Pereira L (2012) The Expansion of mtDNA Haplogroup L3 within and out of Africa. Mol Biol Evol 29:915-927

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, Reed FA, Mountain JL (2007) History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol 24:2180-2195

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H-J. 2006. Harvesting the fruit of the human mtDNA tree. Trends Genet. 22:339–345.

van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386-394

van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH (2014) Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. Hum Mutat 35:187-191

Vianello D, Sevini F, Castellani G, Lomartire L, Capri M, Franceschi C (2013) HAPLOFIND: a new method for high-throughput mtDNA haplogroup assignment. Hum Mutat 34:1189-1194

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503-1507

Supplementary Information section 6 - Population Structure and Admixture

6.1. Comparative Data

Comparative SNP study data were downloaded for both Illumina and Affymetrix Human Origins SNP platforms. The SNP sets of the two platforms were kept separate for analyses to maximize the SNP overlap (aDNA data handling for merging with SNP data is described in section S4.1). The Illumina platform southern African datasets containing Khoe-San and Bantu-speaker groups, typed on the 2.5 Omni array (Schlebusch et al. 2012; Schlebusch et al. 2016), were merged with the data from the ancient individuals. In this southern African dataset 1,989,349 SNPs were retained from the merge, and the number of SNPs for each of the ancient individuals is indicated in Table S6.1. This dataset was merged with 6 additional populations (YRI, MKK, LWK, TSI, CEU, JPT) from the 1000 genomes project (KGP) global dataset typed on the Illumina 2.5 Omni array (Auton et al. 2015). In this global extended dataset 1,984,902 SNPs were retained and the number of SNPs of each ancient individual is indicated in Table S6.1. To expand the modern-day East African representation of the dataset, we merged the data with 6 additional populations (AMHARA, OROMO, ARI-BLACKSMITH, GUMUZ, SUDANESE, SOMALI) from diverse East African groups, typed on the Illumina 1M Omni array (Pagani et al. 2012). For this 'East Africa extended' dataset, 527,131 SNPs were retained and the number of SNPs for each of the ancient individuals are indicated in Table S6.1. All the mergers of datasets were performed using Plink v. 1.9 (Chang et al. 2015) and A/T and C/G SNPs were removed before merging the datasets. During merging, mismatching SNPs were strand-flipped once and remaining mismatching SNPs were excluded. Only intersecting SNPs were kept after merging. To include more African populations, but to retain high SNP density, we also merged 16 additional populations from the African Genome Variation Project (Gurdasani et al. 2015) with the Global Comparative dataset to form the AGV comparative dataset and retained 1,421,001 SNPs (Table S6.1).

We downloaded the Affymetrix Human Origins fully public dataset as described in (Lazaridis et al. 2014) from (https://reich.hms.harvard.edu/datasets). This dataset also contained Khoe-San populations from (Pickrell et al. 2012). We merged the ancient individuals with this dataset to form the Human Origin comparative dataset (548,476 retained SNPs).

Comparative full genome data, consisting of bam files of 11 HGDP samples (HGDP: 1 individual from Dinka, Mbuti, French, Papuan, Sardinian, Han, Yoruba, Karitiana, San, Mandenka, and Dai populations) were downloaded from (http://www.cbs.dtu.dk/suppl/malta/data/Published_genomes/bams/). The data were originally generated by Meyer et al. (Meyer et al. 2012) and the re-mapping and generation of the bam files was done and described in Raghavan et al. (Raghavan et al. 2014). The 11 HGDP bamfiles were used and SNPs called individually for each bamfile using the Unified Genotyper of GATK v. 3.2.0 (McKenna et al. 2010). SNPs and indels were called separately. A strand call confidence of 30.0 was used, all sites present in reference genome were emitted (not just variant sites) and vcfs were extensively annotated (SpanningDeletions, Coverage, DepthPerAlleleBySample, QualByDepth, FisherStrand, MappingQualityRankSumTest, ReadPosRankSumTest, GCContent, HaplotypeScore, HomopolymerRun, TandemRepeatAnnotator, VariantType). After SNP calling we applied a hard filter with the following criteria: "QD < 3.0 || FS > 20.0 || MQ < 55.0 || MQRankSum < -3 || ReadPosRankSum < -4.0 || SOR > 3.0 || HaplotypeScore > 5.0".

Additionally, the Neandertal and Denisova genomes were prepared for comparative data analysis: Denisova (Published originally in Meyer et al. (Meyer et al. 2012), remapped in Raghavan et al. (Raghavan et al. 2014) and obtained for this study from http://www.cbs.dtu.dk/suppl/malta/data/Published_genomes/bams/), Neandertal (Published in Prüfer et al. (Prüfer et al. 2014) and obtained from http://cdna.eva.mpg.de/Neanderthal/altai/AltaiNeanderthal/bam/). The SNPs for Neandertal and Denisova were called similarly to the HGDP bams and the following hard filter was applied: "QD < 3.0 || FS > 20.0 || MQ < 30.0 || MQRankSum < -3 || ReadPosRankSum < -4.0 || SOR > 3.0 || HaplotypeScore > 10.0".

To be able to compare directly the ancient individuals to genome sequence data from a larger diverse set of modern-day individuals (which is not affected by ascertainment bias present in SNP array genotype data), we downloaded the called variants from the Simons Genome project (Mallick et al. 2016) (https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/PS3_multisample_public/). These genotype data were merged with the Ballito Bay A diploid called sites to be used for confirming results obtained from SNP array data (that contain many more individuals compared to the genome sequence datasets).

Table S6.1: Comparative dataset with the number of SNPs present in ancient individuals from this study

	Southern African dataset	Global Extended Dataset	East African Extended Dataset	AGV Extended Dataset	Human Origins dataset	Simons Genome Variant Sites
Full merged dataset	1,989,349	1,984,902	527,131	1,421,001	548,476	28,622,172
BBayA	1,962,247	1,957,905	526,465	1,402,541	548,153	24,671,536
BBayB1	1,268,240	1,265,495	341,135	908,376	363,171	na
BBayB2	9,854	9,831	2,564	7,040	2,849	na
Champagne Castle	479,547	478,510	127,778	344,026	136,456	na
Doonside	7,261	7,249	1,859	5,272	1,987	na
Eland Cave	1,961,630	1,957,282	526,467	1,402,052	548,172	na
Mfongozi	1,957,298	1,952,973	525,283	1,399,297	547,083	na
Newcastle	1,961,140	1,956,800	526,280	1,401,827	548,009	na

6.2. Principal Component Analysis

Principal Component Analysis (PCA) was done on haploidized versions of comparative datasets (random draw of one of the alleles at each locus). PCA was performed using EIGENSOFT (Patterson et al. 2006; Price et al. 2006) with the following parameters: r2 threshold of 0.9, sample size limit of 20, 10 iterations of outlier removal.

The first PC of the southern African dataset (Figure S6.1) separates Khoe-San from Bantu-speakers and the second PC separates southeast Bantu-speakers from southwest Bantu-speakers. The two older samples (BBayA and BBayB) cluster with the current-day Khoe-San groups, while the four younger samples cluster with southeast Bantu-speakers.

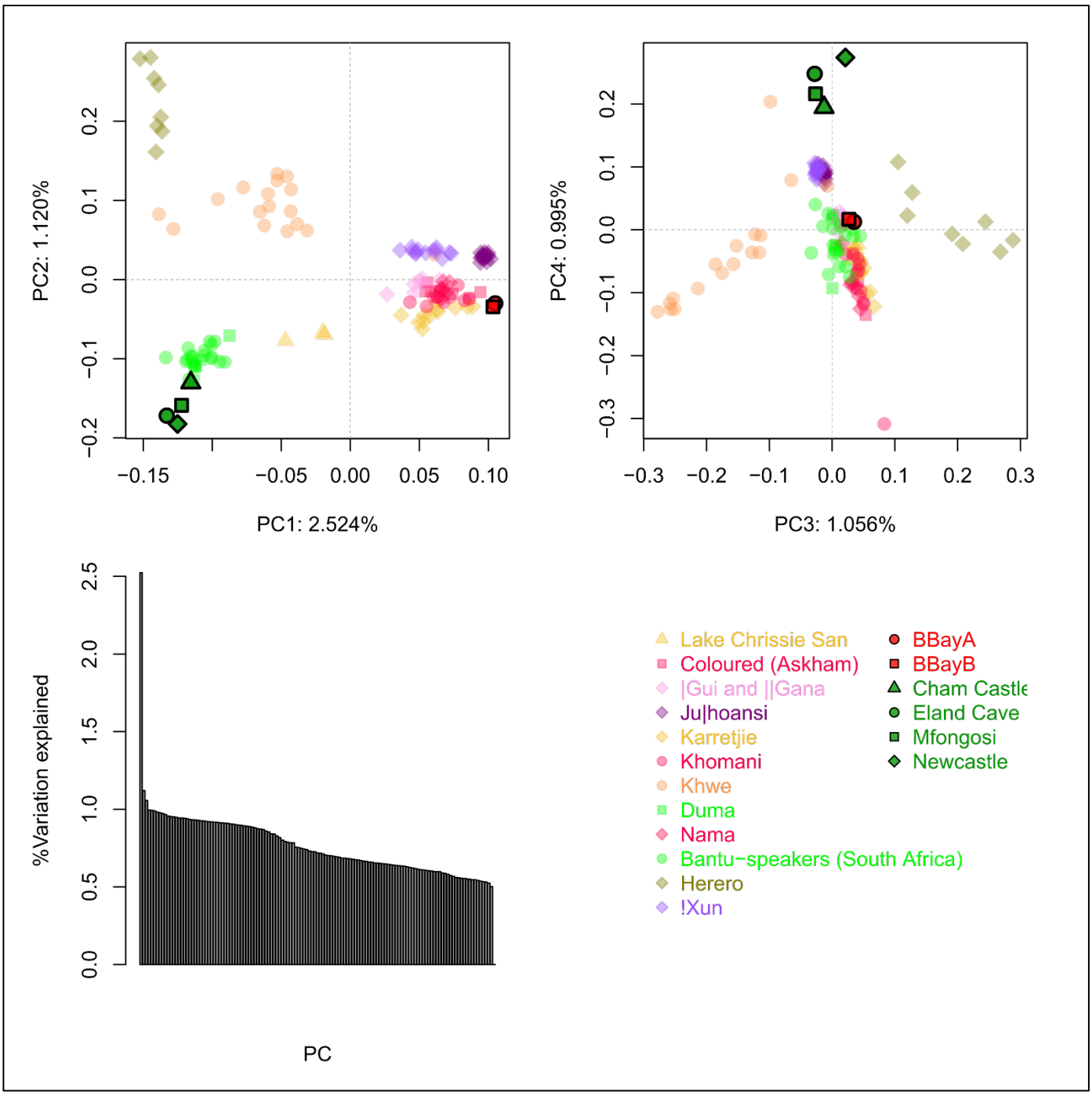

Figure S6.1: Principal Component analysis of the Southern African dataset, showing first four PCs.

When adding comparative African and non-African groups from the KGP panel (Figure S6.2), PC1 separates non-Africans from Africans and PC2 separates West African origin populations from southern African Khoe-San populations. On this PC BBayA and BBayB form the one extreme and West African Yoruba the other extreme. It appears that compared to BBayA and BBayB, all current-day Khoe-San groups are shifted towards the non-African and other African extremes of the PCA. The four younger samples (ELA, NEW, MFO, CHA) are located with southern African Bantu-speakers (between the southwestern and southeastern Bantu-speakers), thus showing evidence of Khoe-San admixture (compared to Yoruba), but not quite as much as most of the current-day southeastern Bantu-speakers. ELA appears to be the least admixed and CHA the most.

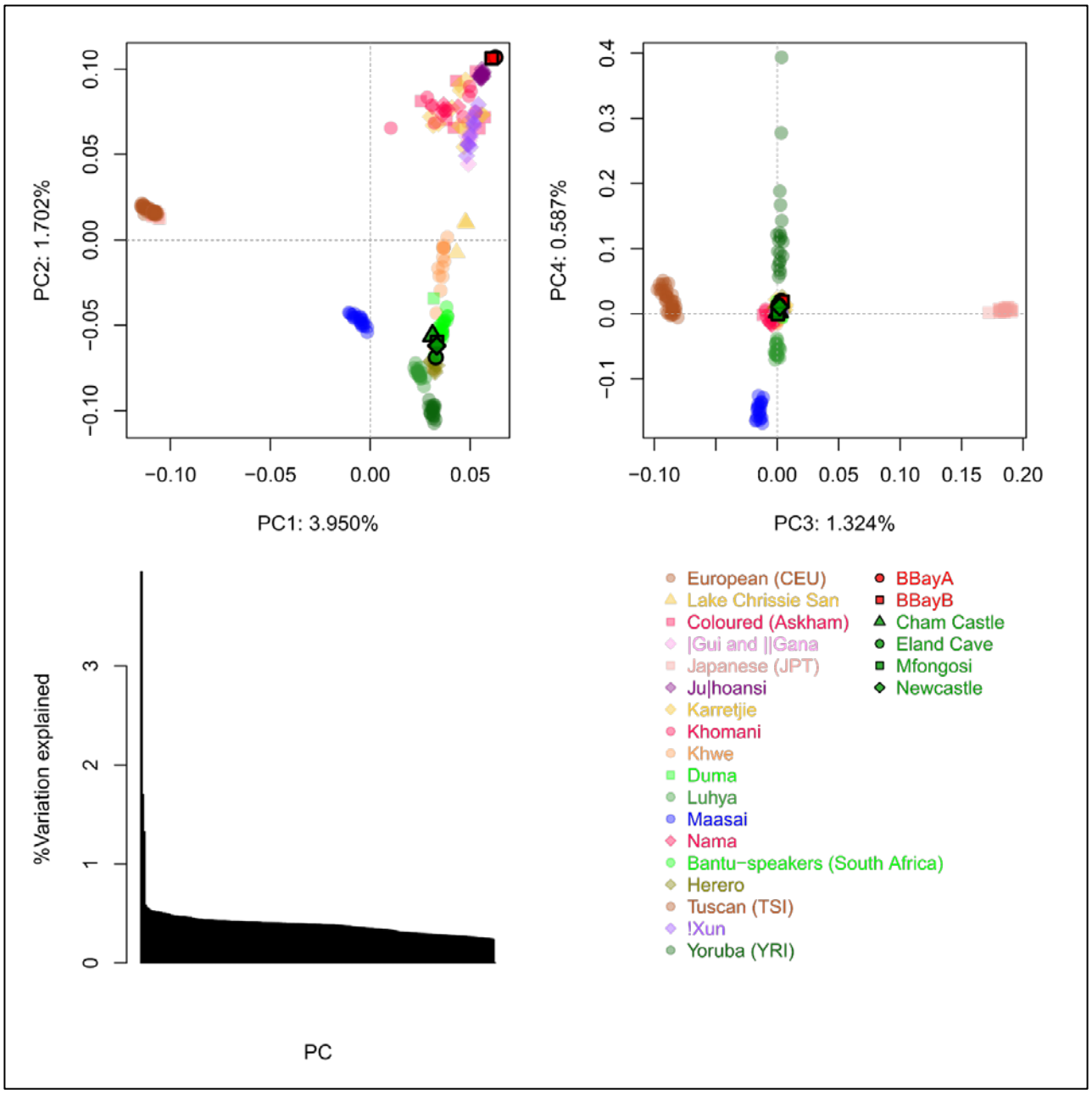

Figure S6.2: Principal Component analysis of the KGP comparative dataset.

To further clarify the association of samples with East and West Africans and northern and southern Khoe-San, only one representative group from East Africa (Maasai), West Africa (Yoruba), northern San (Ju|'hoansi) and southern San (Karretjie People) were included in the PCA. In this analysis it appears that, compared to BBayA, Ju|'hoansi is shifted towards East Africa, while some of the Karretjie individuals are shifted towards East Africans and some towards West Africans (Figure S6.3). BBayA and BBayB appear to cluster with southern San and not with northern San. The Khoe-San admixture in the younger samples also appears to have come from southern San (Figure S6.4).

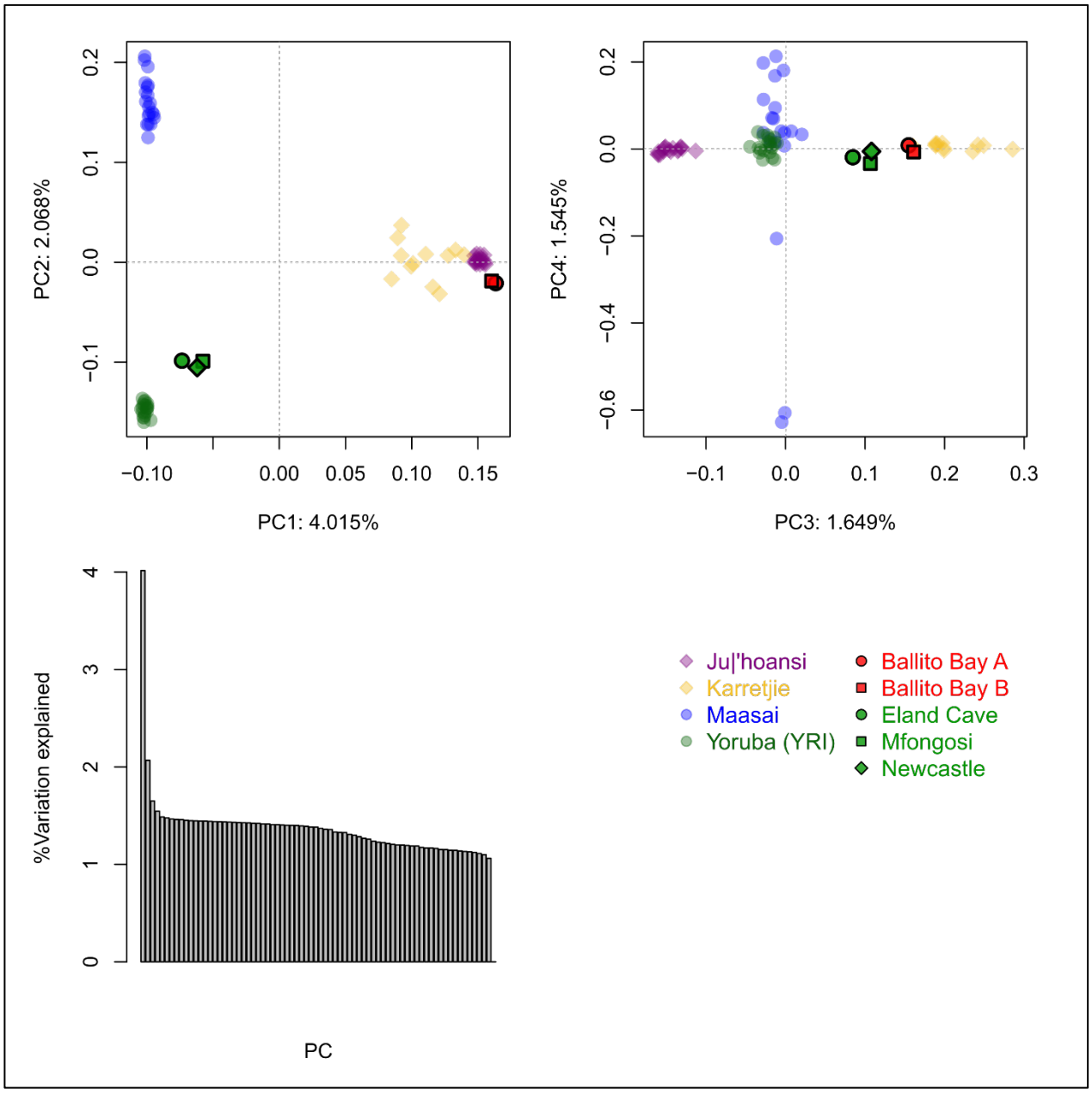

Figure S6.3: Principal Component analysis with comparative East and West Africans (Maasai and Yoruba) and southern and northern Khoe-San (Karretjie and Ju|'hoansi) (maximum Khoe-San SNPs - excluding CHA and DOO).

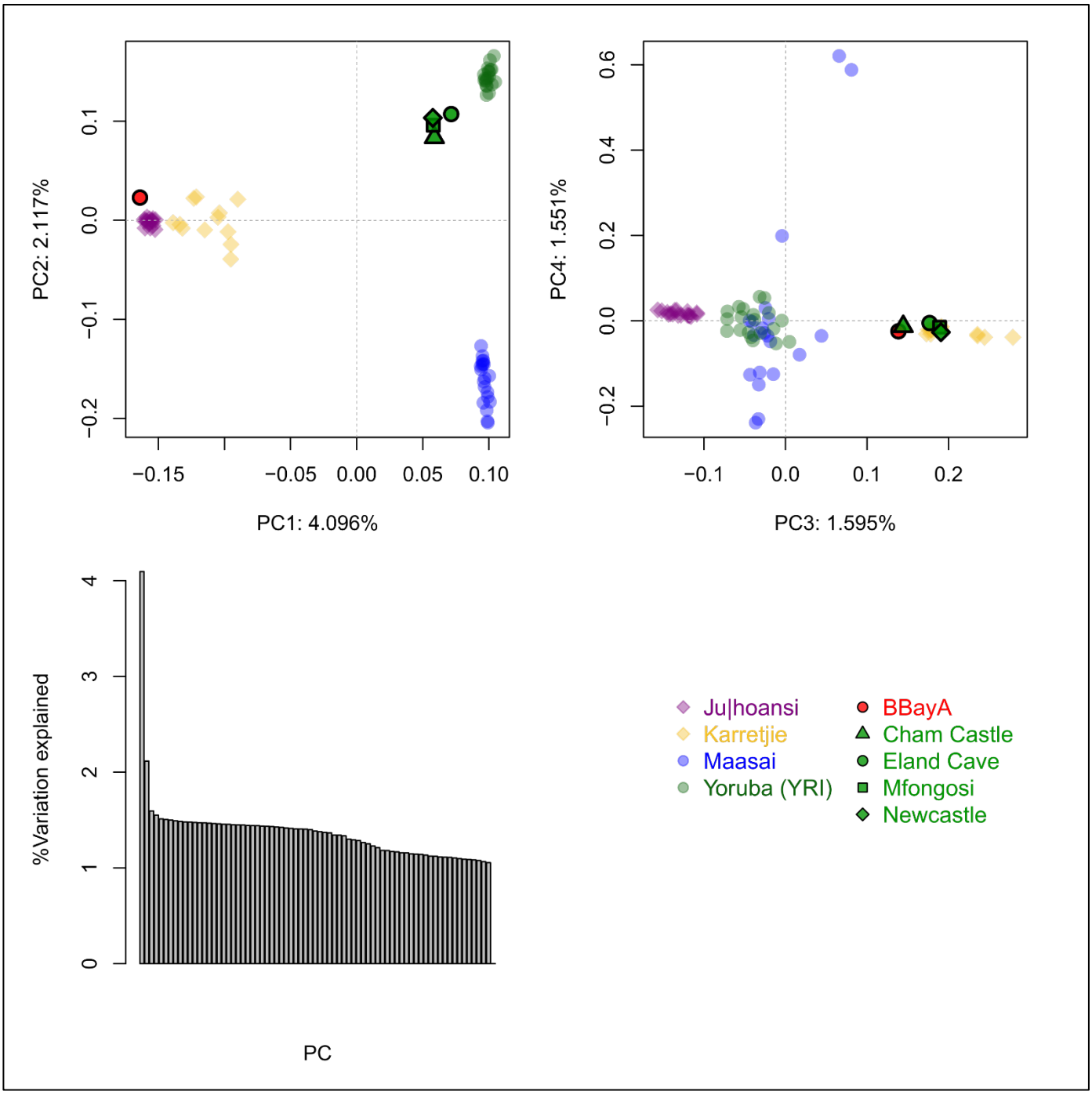

Figure S6.4: Principal Component analysis with comparative East and West Africans (Maasai and Yoruba) and southern and northern Khoe-San (Karretjie and Ju|'hoansi) (maximum Bantu-speaker SNPs, excluding BBayB and DOO).

6.3 Cluster analysis

Admixture fractions were estimated using ADMIXTURE (Alexander et al. 2009) in order to cluster individuals based on SNP genotypes. Default settings and random seeds were used. Between 2 and 13 clusters (K) were tested. A total of 50 iterations of ADMIXTURE were run for each value of K. Iterations for each K were analyzed using CLUMPP (Jakobsson and Rosenberg 2007) and the LargeKGreedy algorithm with 1,000 repeats to identify common modes among replicates. Pairs of replicates yielding a symmetric coefficient G'>=0.9 were considered to belong to common modes. The most frequent common modes were selected and CLUMPP was run a second time for all values of K containing the most frequent common mode (LargeKGreedy algorithm, 10,000 repeats). The results were visualized using DISTRUCT (Rosenberg 2004) (Figure S6.5).

Admixture analyses show that the three ~2000 year old individuals (BBayA, BBayB and DOO) cluster with present-day Khoe-San groups (Figure S6.5, for K≥3), and specifically southern San groups (for K≥7). At K=13, these individuals cluster with the Karretjie People (Schlebusch et al. 2011; Schlebusch et al. 2012) and the Lake Chrissie San (CHR) (Schlebusch et al. 2016). The four ~300-500-year-old individuals (ELA, NEW, MFO and CHA) grouped with populations of West African origin (Figure S6.5, for K≥3), and more specifically southeast Bantu-speakers from South Africa (for K≥12). They have low, but clear signals of admixture with southern Khoe-San groups (Figure S6.6) and the admixture is lowest for the oldest of the four individuals (ELA - 12%) and greatest for the youngest individual (CHA - 18%). Comparatively the levels of admixture in current-day southeastern Bantu-speakers are 19% on average. This observation is consistent with continuous admixture into Bantu-speakers from San groups, however, we note that the time-serial sample is small. CHA also had an L0d mtDNA haplogroup. Although found in its highest frequency in Khoe-San (Schlebusch et al. 2013; Barbieri et al. 2014), L0d occurs at levels of 20% to 40% in present day southeast Bantu-speakers from South Africa (Schlebusch et al. 2013).

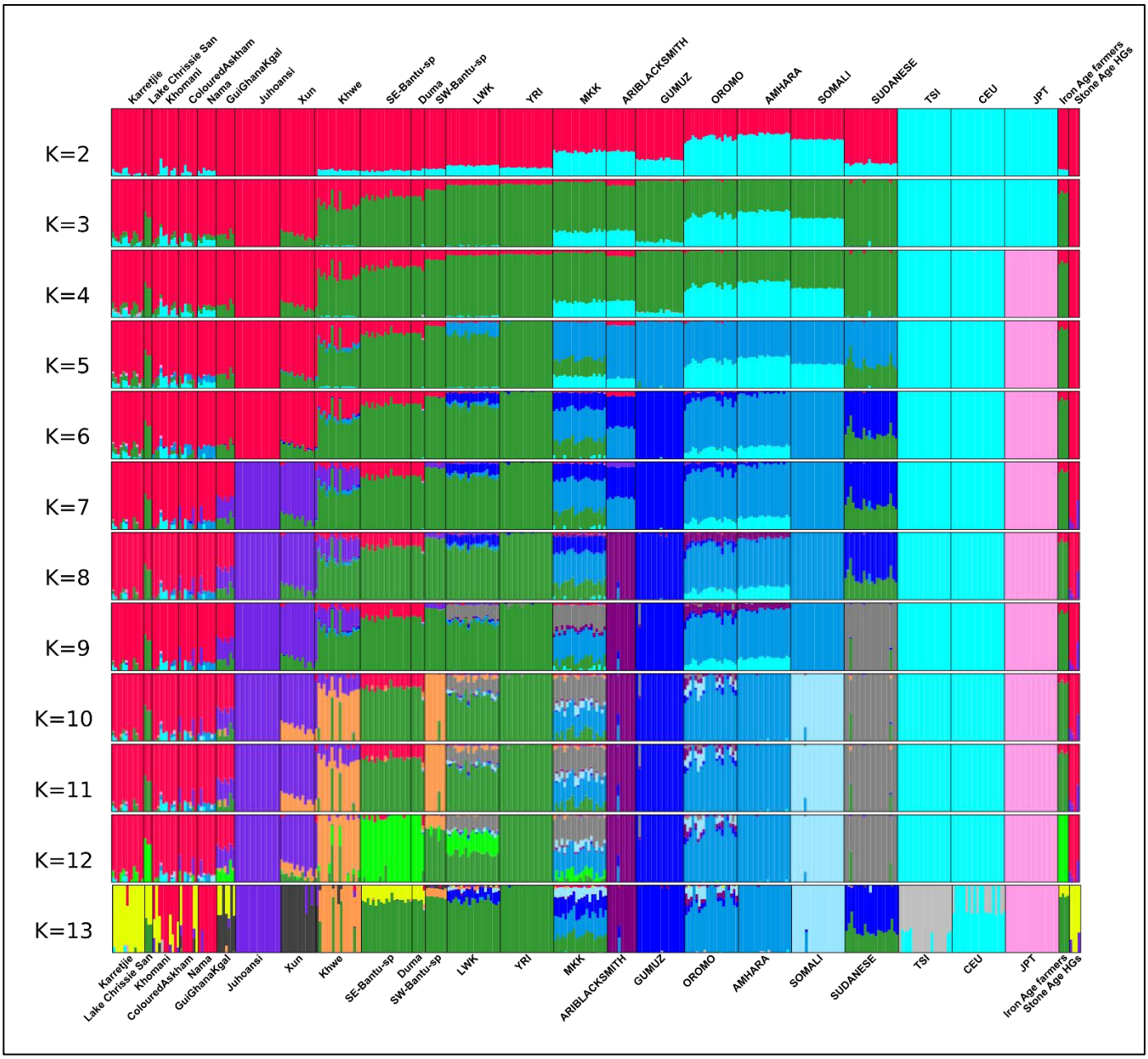

Figure S6.5: Admixture analysis with Illumina datasets for K=2 until K=13.

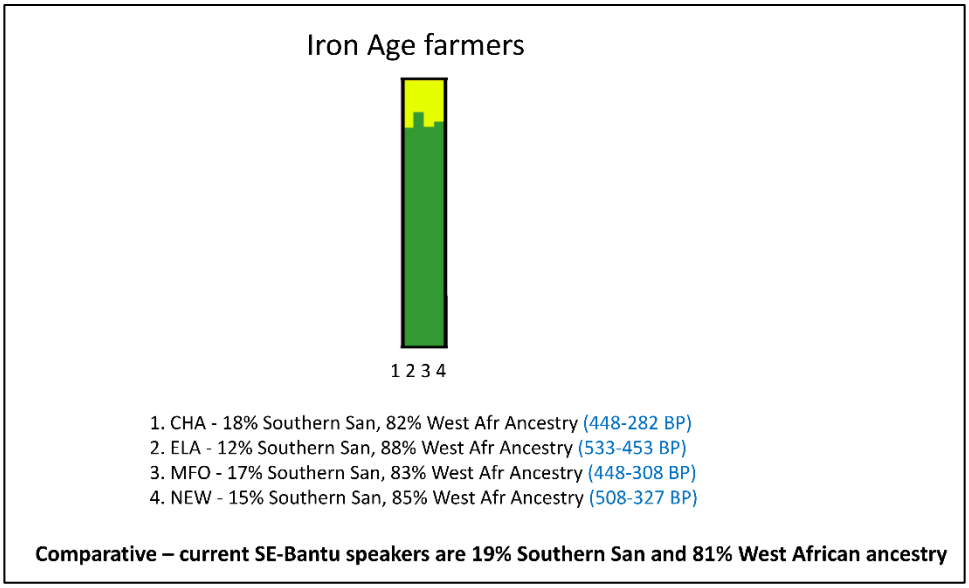

Figure S6.6: Zoom-in on Bantu-speaker aDNA

6.4 Formal tests of admixture and fractions of admixture

f3 tests

To estimate whether the Ju|'hoansi received admixture from an external population, we computed the f3-statistic (Patterson et al. 2012) with Ju|'hoansi as the recipient population, the diploid Ballito Bay A as one source and other populations of the 'East African extended' dataset as the other source population. Full results are shown in Extended Data Table 1. Negative Z scores were observed for all non-Africans and East Africans (except Mota), as a source population in addition to Ballito Bay A. This points to admixture from East Africans and/or Eurasians into the Ju|'hoansi.

f4 ratio statistics

In order to estimate the degree of back-admixture (from non-Africans) into African populations, we followed the approach by Gallego-Llorente et al. (Gallego Llorente et al. 2015) and calculated a ratio of f4 statistics (Patterson et al. 2012). The statistic calculates the proportion of ancestry from a Eurasian source α for each population X by using an ancient Eurasian (LBK in our case; (Lazaridis et al. 2014)) and an ancient African (BBayA in our case) as sources and East Asians (Japanese) and Europeans (French or CEU, depending on the dataset) as outgroups,

.

We calculated this statistic for the different datasets using qpF4ratio from ADMIXTOOLS (Patterson et al. 2012). The estimates were similar between datasets (Tables S6.2-S6.5). We noticed that α_{Mota} is positive in all datasets, which suggests that Mota itself might have received some Eurasian back-admixture compared to BBayA

Table S6.2: Eurasian back-admixture for the East African extended dataset (sorted high to low alpha values).

Population	alpha	error
AMHARA	0.443272	0.012226
OROMO	0.376639	0.011544
SOMALI	0.352891	0.011891
MKK	0.196622	0.011672
ARIBLACKSMITH	0.156234	0.012821
Nama	0.099949	0.009624
≠Khomani	0.094718	0.00932
Coloured (Askham)	0.079101	0.009464
LWK	0.057003	0.012029
Khwe	0.05415	0.01039
GUMUZ	0.053537	0.012882
SWBantu-speaker	0.049844	0.012336
Karretjie	0.047008	0.009168
ǁGui and ǁGana	0.046126	0.009316
YRI	0.039827	0.012073
!Xun	0.039021	0.009316
Ju|'hoansi	0.038835	0.008971
Duma	0.037908	0.01191
SEBantu-speaker	0.036732	0.011228
SUDANESE	0.036503	0.01277
Lake Chrissie San (CHR)	0.027255	0.01182

Table S6.3: Eurasian back-admixture for the AGV extended dataset (sorted high to low alpha values).

Population	alpha	error
AMHARA	0.427872	0.010667
OROMO	0.390445	0.010586
SOMALI	0.335595	0.010513
MKK	0.183071	0.010643
Kikuyu	0.140276	0.010065
Fula	0.126072	0.010167
Kalenjin	0.120642	0.010222
Banyarwanda	0.103852	0.010012
Nama	0.093531	0.008663
≠Khomani	0.090794	0.008549
Barundi	0.078089	0.010269
Coloured (Askham)	0.074879	0.008423
Wolof	0.054592	0.010813
Baganda	0.053819	0.0105
Mandinka	0.047523	0.010613
Khwe	0.046938	0.00953
LWK	0.046892	0.010646
ǁGui and ǁGana	0.039741	0.008474
Karretjie	0.038893	0.008258
SWBantu-speaker	0.036938	0.011209
Jola	0.03603	0.010863
Ga-Adangbe	0.033984	0.010763
Zulu	0.033363	0.009937
!Xun	0.033295	0.008581
Ju|'hoansi	0.031578	0.008158
Sotho	0.031258	0.009879
Igbo	0.03092	0.010771
YRI	0.029805	0.010953
SEBantu-speaker	0.027412	0.010217
Duma	0.027111	0.010993
Lake Chrissie San (CHR)	0.022698	0.010479
Mota	0.013578	0.022693

Table S6.4: Eurasian back-admixture for the Human Origins dataset (sorted high to low alpha values)

Population	alpha	error
Mozabite	0.681708	0.014591
Nama	0.148618	0.010506
≠Khomani	0.120745	0.010962
Hadza	0.090578	0.014182
Shua	0.082231	0.011765
Haiom	0.080085	0.011854
Khwe	0.076263	0.012046
Bantu-sp Kenya	0.072703	0.014146
Mandenka	0.064459	0.01367
Naro	0.062763	0.01081
Tshwa	0.060713	0.011814
Gui	0.059809	0.010937
Damara	0.057348	0.014018
Gana	0.056705	0.011501
Himba	0.055636	0.014962
Hoan	0.054905	0.011505
Taa_North	0.05372	0.010788
Dinka_Hammer	0.052721	0.015036
Kgalagadi	0.052585	0.012897
Mbukushu	0.051359	0.014439
Bantu-sp SouthAfrica	0.050945	0.013909
Yoruba	0.046038	0.013886
Xuun	0.046014	0.01087
Juhoan_North	0.04312	0.010743
Taa_East	0.042393	0.011278
Tswana	0.042139	0.013489
Juhoan_South	0.040093	0.011168
BiakaPygmy	0.039545	0.012617
Taa_West	0.035986	0.010708
Wambo	0.035896	0.014393
MbutiPygmy	0.029326	0.013509

Table S6.5: Eurasian back-admixture for SGDP dataset (sorted high to low alpha values).

Population	alpha	Error
Mozabite	0.670681	0.026642
Saharawi	0.664817	0.027075
Somali	0.303252	0.02758
Masai	0.195705	0.021809
Mandenka	0.068516	0.020048
Gambian	0.054313	0.021615
Bantu-sp Herero	0.053561	0.021034
Luhya	0.051574	0.020709
Bantu-sp Tswana	0.048762	0.019652
Mende	0.047244	0.021645
Ju_hoan_North	0.045141	0.013975
Luo	0.044697	0.02079
Bantu-sp Kenya	0.040586	0.020779
Dinka	0.037619	0.020211
≠Khomani_San	0.036492	0.015546
Esan	0.035596	0.019309
Mbuti	0.033132	0.017235
Yoruba	0.031474	0.020214
Biaka	0.023311	0.019767

As most southern African populations display admixture from a source of mixed east African and Eurasian ancestry (see below), we also used f4 ratios to estimate this ancestry using two different admixed east African (Oromo and Amhara) populations as source. The results of this analysis are shown in Table S6.6 and S6.7.

We do not assume that the admixing source was either Amhara or Oromo; they are just the best representatives in the dataset. Both sources lead to similar and highly correlated estimates. We note that the estimates for non-southern Africans should be interpreted with caution as the two sources might not be a suitable model for those populations.

Table S6.6: Admixed East African ancestry proportions in AGV extended dataset (sorted high to low alpha values in Amahra).

Target population	alpha(OROMO)	stderr(OROMO)	alpha(AMHARA)	stderr(AMHARA)
SOMALI	0.860526	0.012784	0.785708	0.011307
MKK	0.469526	0.018803	0.428661	0.018062
Kikuyu	0.360969	0.019221	0.329539	0.018323
Fula	0.325347	0.020599	0.297013	0.019464
Kalenjin	0.311013	0.020659	0.283924	0.019514
Banyarwanda	0.26791	0.021096	0.244575	0.019735
Nama	0.24433	0.018992	0.223057	0.017678
≠Khomani	0.23594	0.018551	0.2154	0.017239
Barundi	0.202461	0.022839	0.184808	0.021264
Coloured (Askham)	0.193945	0.018938	0.17706	0.017484
Wolof	0.142684	0.025286	0.130216	0.023429
Baganda	0.140508	0.024477	0.128239	0.022635
Mandinka	0.124378	0.02509	0.113504	0.023194
LWK	0.123884	0.025108	0.113053	0.023207
Khwe	0.121989	0.022523	0.11134	0.020796
ǂGui and ǁGana	0.104318	0.02021	0.095219	0.018592
Karretjie	0.102085	0.019775	0.093179	0.018199
SWBantu-sp	0.097868	0.027092	0.089301	0.024937
Jola	0.095553	0.02621	0.087182	0.02415
Ga-Adangbe	0.09052	0.025979	0.082588	0.023927
!Xun	0.089345	0.020666	0.081543	0.019001
Zulu	0.08761	0.023936	0.079941	0.022033
Juhoansi	0.085977	0.019692	0.078475	0.018094
Sotho	0.08229	0.023871	0.075084	0.021964
Igbo	0.081864	0.026172	0.074684	0.02408
YRI	0.079536	0.026657	0.072555	0.024525
Duma	0.072354	0.026884	0.066012	0.024683
SEBantu-sp	0.072167	0.024892	0.065841	0.022872
Lake Chrissie San (CHR)	0.061098	0.025843	0.055745	0.023691
Mota	0.030152	0.056876	0.027437	0.052003

Table S6.7: Admixed East African ancestry proportions in East African extended dataset (sorted high to low alpha values in AMHARA).

Target Population	alpha(OROMO)	stderr(OROMO)	alpha(AMHARA)	stderr(AMHARA)
SOMALI	0.936591	0.015848	0.794275	0.014208
MKK	0.518541	0.020572	0.43969	0.019122
ARIBLACKSMITH	0.4123	0.026744	0.349587	0.023819
Nama	0.268929	0.021813	0.228031	0.019015
≠Khomani	0.255832	0.021049	0.216926	0.018334
Coloured (Askham)	0.212254	0.02197	0.179966	0.019053
LWK	0.149596	0.02916	0.126774	0.025231
Khwe	0.142553	0.025106	0.120816	0.021767
GUMUZ	0.139795	0.031365	0.118447	0.027112
SWBantu-sp	0.130486	0.030468	0.110572	0.026251
Karretjie	0.124186	0.022619	0.105278	0.019415
ǂGui and ǁGana	0.122238	0.022986	0.103623	0.019771
Juhoansi	0.107662	0.022405	0.091258	0.019246
!Xun	0.105042	0.023075	0.089015	0.0199
YRI	0.104296	0.030014	0.088348	0.025809
Duma	0.096901	0.030021	0.082081	0.025796
SEBantu-sp	0.095029	0.028048	0.080496	0.024114
SUDANESE	0.094218	0.031807	0.079786	0.027334
Lake Chrissie San (CHR)	0.071228	0.030134	0.06033	0.025744
Mota	0.012729	0.070952	0.010431	0.059915

6.5 Admixture graphs

Introduction and methods:

Our analyses suggest several admixture events into Khoe-San populations during the last 2000 years. In order to disentangle the contributions from different admixture sources, we used qpGraph v5052 of ADMIXTOOLS (Patterson et al. 2012) to construct admixture graphs. qpGraph takes a user-defined graph as input and estimates drift parameters and admixture proportions by calculating all combinations of f statistics along the graph. Each internal node in the graph can represent a bifurcation into two child populations and/or the recipient or source for two-way admixtures. An admixture graph is usually considered consistent with the data if the differences between all observed (from the data) and expected (from the graph) values of the f statistics are less than 3 standard errors apart from each other ($|Z|<3$). The standard errors are calculated using a block-jackknife across the whole genome. QpGraph was used with the following settings; initmix: 1000, lsqmode: YES, blgsize: 0.005 and diag: 0.0001. We used the chimpanzee reference genome sequence as an outgroup. If the chimpanzee allele was different from the two alleles observed in the human populations, the site was excluded from the analysis.

Basic graph for the East African extended dataset

We started by constructing a graph for the three ancient genomes used in this analysis. We used Mota (Gallego Llorente et al. 2015) as a representative of ancient East Africa, LBK (Lazaridis et al. 2014) to represent Eurasians and BBayA to represent Stone Age southern Africans. A simple model where African populations split into East Africans related to Mota and southern Africans related to BBayA, followed by a split of an out-of-Africa population which is closer to Mota, is consistent with the data (worst $|Z|<0.19$). We then added

Amhara as an East African population with a substantial degree of Eurasian back-admixture (see above and (Pagani et al. 2012)). Consistent with these expectations, Amhara would fit as an admixed population between Eurasians and East Africans (worst |Z|<1.49) with 38 percent contribution from East African and 62% from Eurasian populations respectively. As a first Khoe-San population, we added Ju|'hoansi to the model. Ju|'hoansi are considered to be the least admixed Khoe-San population, but our data suggest that they received East African and/or Eurasian admixture since BBayA lived. Therefore, we tested four different models for Ju|'hoansi:

(I) Ju|'hoansi as an admixed population between Stone Age southern Africans and Eurasians.

(II) Ju|'hoansi as an admixed population between Stone Age southern Africans and East Africans.

(III) Ju|'hoansi as an admixed population between Stone Age southern Africans and a population related to Amhara.

(IV) Ju|'hoansi as an admixed population between southern Africans and a population admixed between Eurasians and East Africans. The ancestry sources for these populations would be identical to those used for Amhara but the admixture contributions would be different.

Models I and II were rejected (worst f4(Mota, LBK; Ju|'hoansi, Amhara), Z=3.58 and f4(baa001, Ju|'hoansi; Mota, LBK), Z=4.58, respectively), but both models III (worst |Z|<2.16) and IV (worst |Z|<1.63) were consistent with the data (Figure S6.7). As the |Z| score for model IV is slightly lower than for model III and since we do not consider Amhara to be the actual source of admixture, we assume model IV to be more representative of the putative population history of Ju|'hoansi, which means that the group mixing with Stone Age southern Africans was probably admixed between Eurasians and East Africans but at different proportions to that of the Amhara. We excluded the Amhara from all further admixture graph modeling of Khoe-San populations.

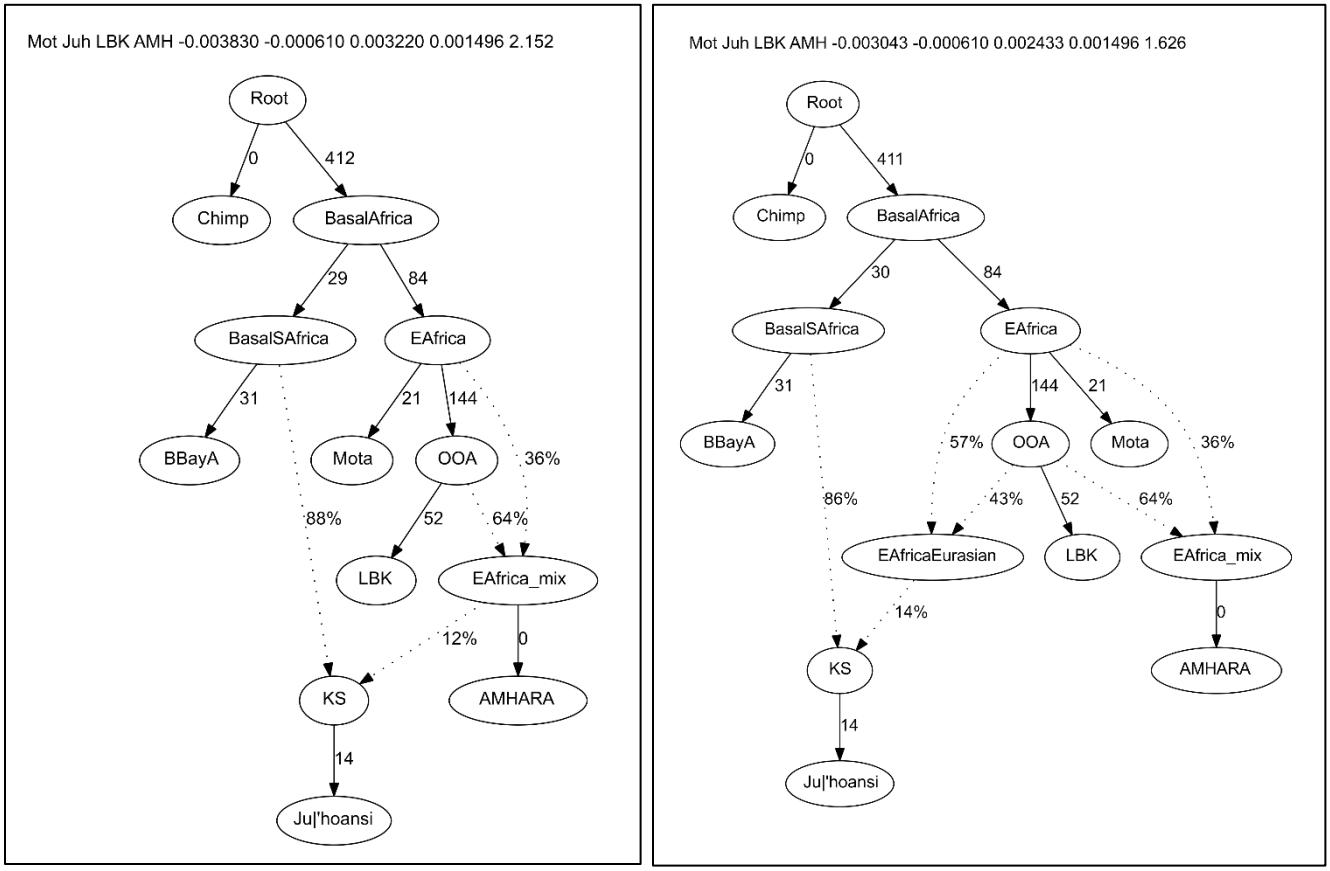

Figure S6.7: Model III and IV.

Modeling West Africans by adding Yoruba

The expansion of Bantu-speaking populations had major impacts on the genomic composition of southern African populations and they contributed ancestry to some Khoe-San populations as well (Pickrell et al. 2012; Schlebusch et al. 2012). Modern-day southern African Bantu speakers have received Khoe-San admixture themselves, which makes it difficult to use them as a source for the Bantu-speaker component in Khoe-San. Due to the lack of West African Bantu-speakers in our data set, we used the closely related West African Niger Kordofanian speakers (Yoruba) as a population related to the source of Bantu-speaker admixture.

We tried to add Yoruba as a simple split off from any internal node of the model or by adding internal nodes from which Yoruba would split off, but none of these models were consistent with the data with $|Z|<3$. Furthermore, models assuming Yoruba as a two-way admixture between any of the internal nodes failed. A model which did fit the data is where Yoruba are modeled as a two-way admixture of two additional nodes: one Basal African node above the split between ancient East Africans and ancient southern Africans and a second group, which is a sister group to the East African population that gave rise to the out-of-Africa groups (worst $|Z|<0.42$). The drift between the Basal African node and the population that splits into Eastern and southern Africans is small compared to the rest of the graph, but it appears to provide a better fit to the data. We did not further investigate the population history of Yoruba as the focus of our analyses is southern Africa.

The model including Yoruba and Ju|'hoansi as a Khoe-San population without admixture from Bantu speakers is consistent with the data. We call this model ‘model A’ (Figure S6.8). In a second approach, we model additional Bantu-speaking admixture into the Khoe-San populations, which we call ‘model B’ (Figure S6.9). We note that model B is also consistent for Ju|'hoansi, (worst |Z|<0.32), but since the simpler model A is also consistent, we conclude that Ju|'hoansi have not received significant admixture from Bantu-speaking populations.

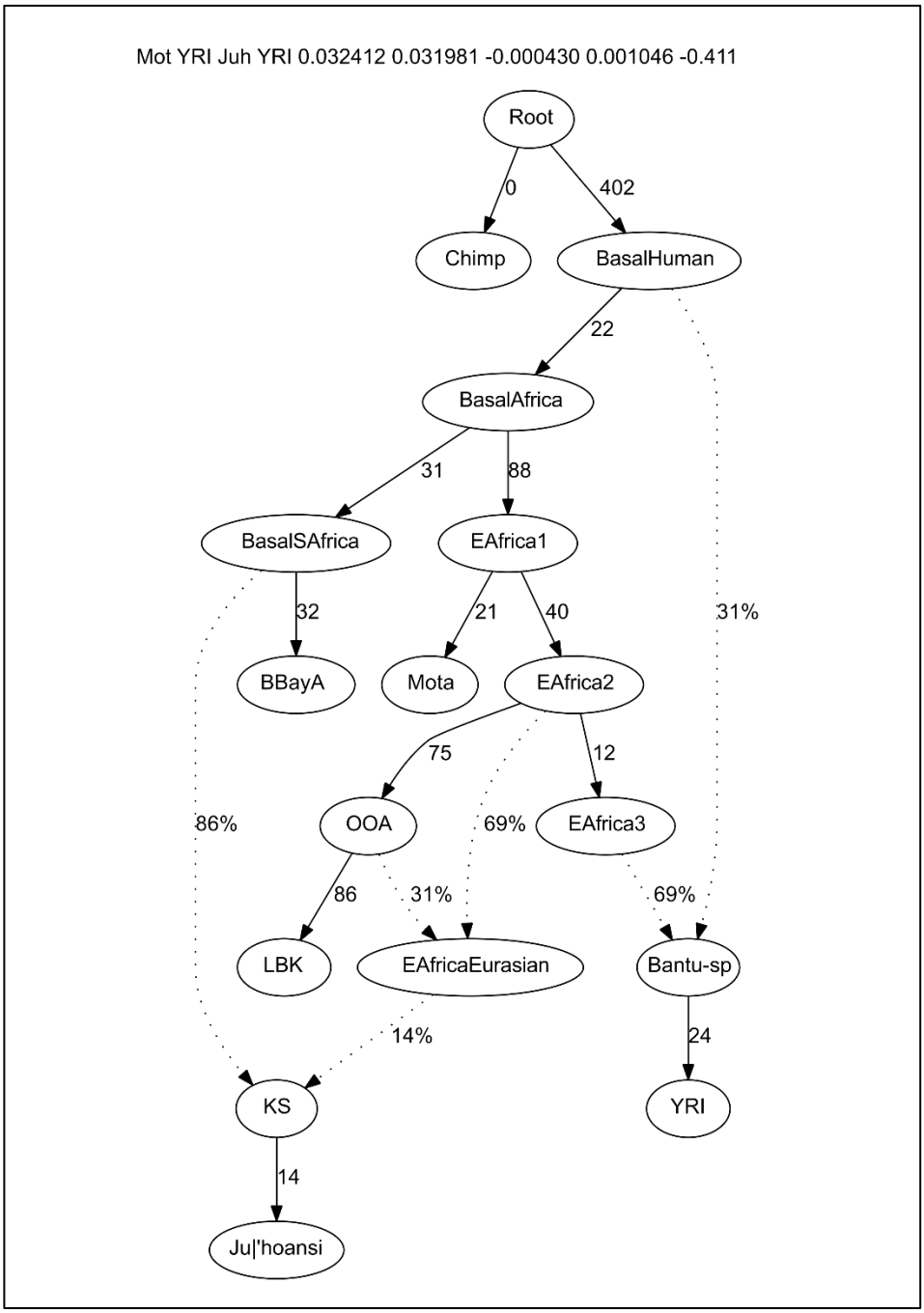

Figure S6.8: Model A for Ju|'hoansi.

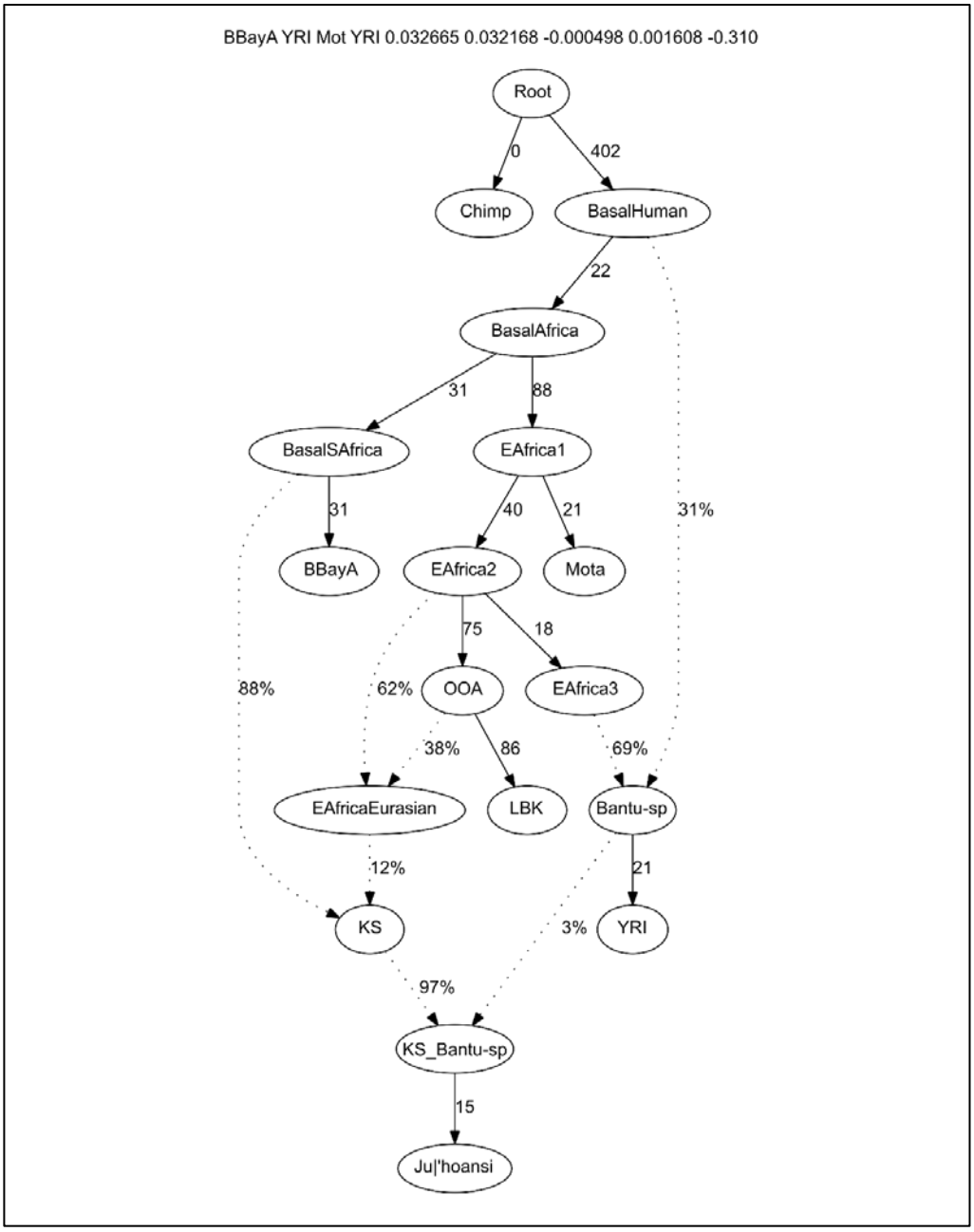

Figure S6.9: Model B for Ju|'hoansi.

Applying model A and B to all KS from the main dataset

We next tried to model all KS populations and the 68,284 overlapping transversion SNPs in the East African extended dataset using both models A and B. Model A is consistent for Ju|'hoansi, Coloured (Askham), ≠Khomani and Nama, indicating no (or minimal) admixture from Bantu speakers. The results of the different admixture proportions are shown in Table S6.8. Notably, some values for Eurasian admixture are different to the values obtained with f4 ratios above this can be attributed to not accounting for East African admixture in the two-source f4 ratio test and also to the fact that the populations used unlikely represent the populations that actually mixed.

Table S6.8: Model A results for East African extended dataset

Population	Ancient southern African Hunter-gatherer	Eurasian	East African	Worst f	Z
Coloured (Askham)	0.76	0.15	0.09	f4(baa001, Coloured (Askham); Mota, YRI)	1.82
Ju|'hoansi	0.86	0.04	0.10	f3(Mota, Ju|'hoansi; YRI)	-0.41
≠Khomani	0.72	0.24	0.04	f3(Mota, ≠Khomani; YRI)	-2.04
Nama	0.70	0.23	0.07	f3(Mota, Nama; YRI)	-0.57

Model B seems to be a suitable representation for all Khoe-San populations. Model B, however, does not always allow us to fully disentangle the East African, West African and Eurasian sources. Part of the ancestry of Yoruba is coming from a population related to East Africans while East Africans also contribute to the mixed East African/Eurasian population, which leads to two possible sources of East African ancestry in Khoe-San when they receive significant West African admixture. Some of the resulting graphs set the East African contribution to the mixed East African/Eurasian population zero, which is balanced by a higher East African component in the Bantu-speaker node. Notably, the Lake Chrissie San (CHR) is the only Khoe-San population where the Eurasian contribution to the mixed East African/Eurasian population is set to zero. This pattern is consistent with results in (Schlebusch et al. 2016) and not unexpected given the probable historical locations of the Khoekhoe populations (Barnard 1992). Generally, these issues make it difficult to interpret and compare the admixture proportions estimated by model B for different populations: the West African proportion might be inflated due to the East African contribution, and different drift parameters between the source nodes make the values hard to compare even if all three proportions are estimated to be non-zero. The results for model B are shown in Table S6.9 but we caution against over-interpreting them.

Table S6.9: Model B results for East African extended dataset

Population	Ancient southern African Hunter-gatherer	Eurasian	East African	Bantu-sp	Worst f	Z
Lake Chrissie San*	0.44	0.00*	0.05	0.51	f3(baa001, Mota; YRI)	-0.34
Coloured (Askham)*	0.73	0.14	0.00*	0.13	f3(baa001, Mota; YRI)	-0.33
|Gui and ||Gana	0.68	0.02	0.09	0.21	f3(baa001, Mota; YRI)	-0.32
Ju|'hoansi	0.85	0.07	0.04	0.03	f3(baa001, Mota; YRI)	-0.31
Karretjie*	0.72	0.13	0.00*	0.15	f3(baa001, Mota; YRI)	-0.32
≠Khomani*	0.69	0.18	0.00*	0.13	f3(baa001, Mota; YRI)	-0.34
Khwe	0.34	0.14	0.03	0.49	f3(baa001, Mota; YRI)	-0.36
Nama	0.68	0.22	0.05	0.05	f3(baa001, Mota; YRI)	-0.33
!Xun	0.70	0.04	0.06	0.21	f3(baa001, Mota; YRI)	-0.32

* difference between East African and Eurasian sources could not be resolved

KS populations from the Human Origins dataset

We also ran qpGraph on the 87,599 overlapping transversion SNPs with the Human Origins dataset (Patterson et al. 2012; Pickrell et al. 2012; Lazaridis et al. 2014; Pickrell et al. 2014), which contains some additional Khoe-San populations. Similar to the results obtained for the East African extended dataset, we see that model A is a consistent model for Ju|'hoansi North, Ju|'hoansi South, ǂKhomani, Naro and Taa North. For Ju|'hoansi North and Taa North, however, the difference between the Eurasian and East African sources could not be resolved – the drift parameter between the two populations was estimated to be very low and all admixture comes from the node slightly closer to the OOA populations. The results for model A are shown in Table S6.10. Model B is consistent with the data for all Khoe-San populations and we observe a similar pattern of balancing of the East African contributions in some Khoe-San (Table S6.11). Overall, the results are quite similar for the two data sets.

Table S6.10: Model A results for the Human Origins dataset.

Population	Ancient southern African Hunter-gatherer	Eurasian	East African	Worst f	Z
Ju|'hoansi North*	0.86	0.14	0.00*	f2(Ju|'hoansi North, Yorubans)	-1.31
Ju|'hoansi South	0.83	0.01	0.16	f2(Ju|'hoansi South, Yorubans)	-1.64
ǂKhomani	0.67	0.28	0.05	f2(ǂKhomani, Yorubans)	-3.00
Naro	0.83	0.04	0.13	f2(Naro, Yorubans)	-1.31
Taa North*	0.83	0.18	0.00*	f2(TaaNorth, Yorubans)	-2.24

* difference between East African and Eurasian sources could not be resolved

Table S6.11: Model B results for the Human Origins dataset.

Population	Ancient southern African Hunter-gatherer	Eurasian	East African	Bantu-sp	Worst f	Z
Gana*	0.57	0.02	0.00*	0.41	f4(Mota, LBK; Gana, Yorubans)	-1.04
Gui	0.69	0.10	0.02	0.19	f3(baa001, Gui; Yorubans)	-0.39
Haiom*	0.54	0.18	0.00*	0.28	f3(baa001, Haiom; Yorubans)	-0.50
Hoan	0.71	0.04	0.02	0.24	f3(baa001, Hoan; Yorubans)	-0.35
Ju|'hoansi North*	0.84	0.08	0.00*	0.08	f3(baa001, Ju|'hoansi North; Yorubans)	-0.48
Ju|'hoansi South	0.80	0.03	0.06	0.11	f3(baa001, Ju|'hoansi South; Yorubans)	-0.32
ǂKhomani*	0.64	0.22	0.00*	0.14	f3(baa001, ǂKhomani; Yorubans)	-0.39
Khwe	0.37	0.07	0.08	0.49	f3(baa001, Khwe; Yorubans)	-0.76
Nama*	0.58	0.24	0.00*	0.18	f3(baa001, Nama; Yorubans)	-0.47
Naro	0.81	0.06	0.05	0.09	f3(baa001, Naro; Yorubans)	-0.32
Shua	0.41	0.10	0.06	0.43	f3(baa001, Shua; Yorubans)	-0.71
Taa East	0.74	0.02	0.01	0.23	f3(baa001, Taa East; Yorubans)	-0.35
Taa North	0.79	0.04	0.03	0.13	f3(baa001, Taa North; Yorubans)	-0.33
Taa West	0.80	0.01	0.05	0.14	f2(Taa West, Yorubans)	-0.35
Tshwa	0.48	0.03	0.11	0.39	f3(baa001, Tshwa; Yorubans)	-0.56
Xuun	0.71	0.05	0.03	0.21	f3(baa001, Xuun; Yorubans)	-0.39

* difference between East African and Eurasian sources could not be resolved

Caveats

We note that while we consider the presented models to be good models to represent the population history of Khoe-San populations, there are very likely other models that would fit the data as well. The extremely large number of possible admixture graph models combined with the need of manually defined graphs restricted us to this kind of analysis. Most applications of studying complex admixture histories with qpGraph have been restricted to less than 200,000 SNP markers so far e.g. (Lazaridis et al. 2014; Fu et al. 2016; Skoglund et al. 2016). As the focus of this analysis was to obtain a simple and general model of the population history of Khoe-San populations, we did not analyze the bigger data sets SGDP and the AGV extended dataset to avoid over-fitting the models due to the large numbers of markers and various minor admixture events between the populations used as proxies for the admixing populations. In summary, the results from this model-fitting are overall consistent with results from other analyses in this study.

6.6 Admixture dating

We used ADMIXTOOLS (Patterson et al. 2012) and the KGP extended dataset to estimate the linkage disequilibrium (LD) decay due to admixture, and thereby infer admixture dates. The date of admixture into current-day Khoe-San groups with no visible Bantu-speaker admixture was estimated using the two ancient southern Africans (BBayA and BBayB) as one parental population and the East African Maasai as the other parental population (Table S6.12). Default parameters were used and the standard error was estimated with a jackknife procedure implemented in the ROLLOFF package. Admixture dates of East Africans into other Khoe-San groups are difficult to distinguish since the admixture with Bantu-speakers will influence signals. Indeed, when admixture dates with East African and West African source populations were inferred for populations with admixture from both groups, the dates were similar (Table S6.13).

Table S6.12: East African admixture dates into populations with no/little Bantu-speaker admixture

Parent1	Parent2	Admixed Pop	Time in gen (SE)	Time in years (SE) (30 y/gen)
BBayA+BBayB	MKK	Ju|'hoansi	50.229 (3.203)	1507 (96)
BBayA+BBayB	MKK	Nama	43.593 (4.335)	1308 (130)

Table S6.13: Comparison of admixture dates in with East and West African source populations

Parent1	Parent2	Admixed Pop	Time in gen
BBayA+BBayB	MKK	Ju|'hoansi	50.2
	YRI		57.4
BBayA+BBayB	MKK	Nama	43.6
	YRI		38.3
BBayA+BBayB	MKK	Karretjie	10.2
	YRI		9.6
BBayA+BBayB	MKK	!Xun	32.5
	YRI		34.3
BBayA+BBayB	MKK	≠Khomani	13.3
	YRI		13.9

6.7 Presence of archaic admixture in current-day Khoe-San

We performed D-tests for testing admixture with Neandertal and Denisovan with P1=HGDP San or BBayA; P2 one of the other 10 HGDP individuals and P3=Neandertal or Denisovan (Figure S6.10 and S6.11). We estimated standard deviations using the weighted block jackknife approach with 5 Mb blocks. We only used sites for which the ancestral state were confidently called (the 3 great apes showed the same variant and exactly one additional variant in the three individuals tested) and set P4 to the ancestral state.

We retrieved the signal for introgression of Neandertals into non-African individuals. Among African individuals, the signal for Neandertal introgression was always around 0 but consistently lower for P1=San than for P1=BBayA. This may reflect a larger Neandertal component (due to admixture) in the HGDP San than in BBayA. When testing for introgression from Denisovans (P3=Denisovan), only the non-African individuals had a mean signal more than 2 SD away from 0 (this is likely to reflect Neandertal admixture) and the well-known strong signal in the Papuan individual was retrieved. Interestingly, the signal for the Mandenka individual was almost as strong as for the French individual, especially for P1=BBayA.

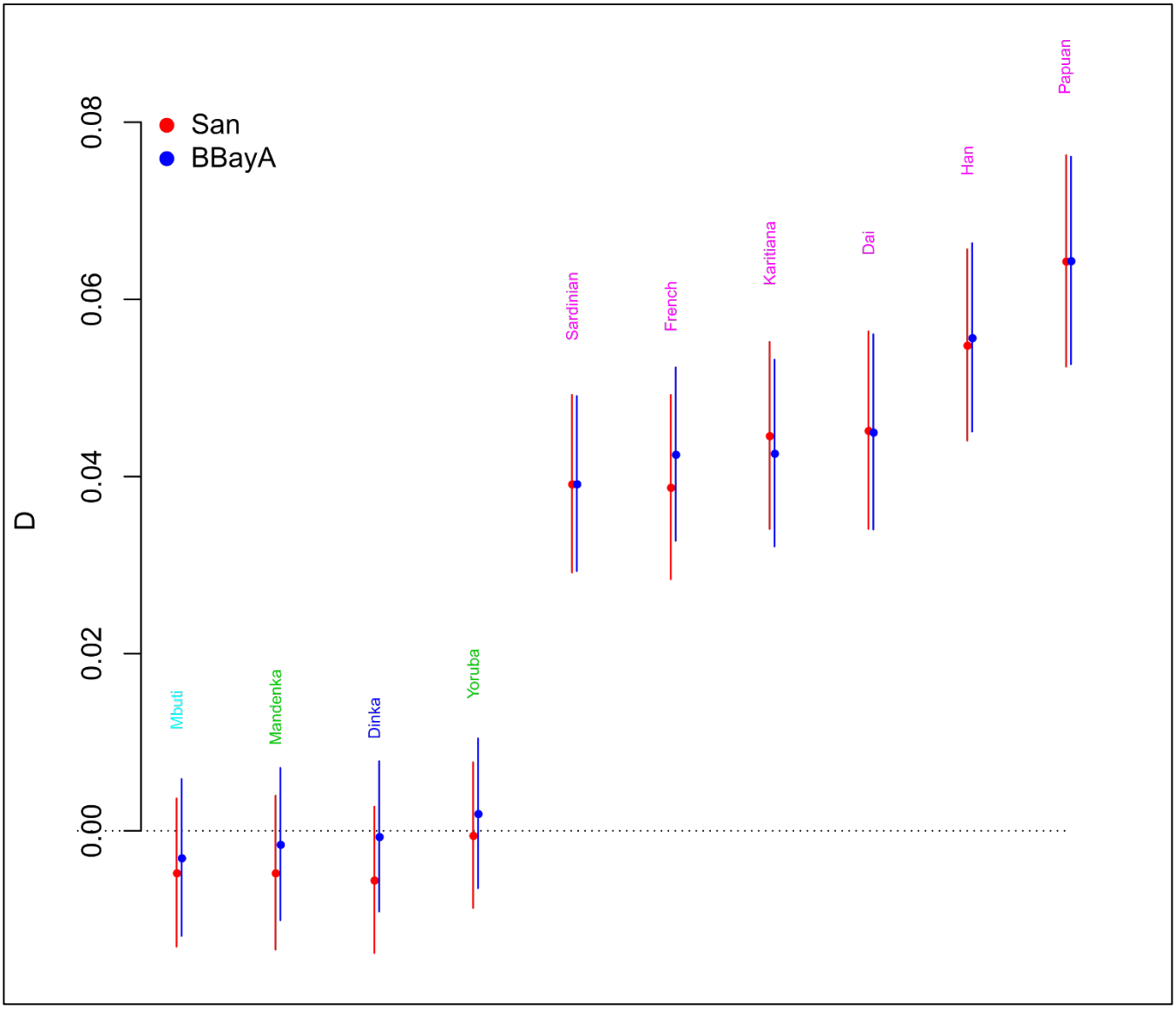

Figure S6.10: D-tests with P1=San or BBayA; P2 one of the 10 HGDP individuals and P3=Neandertal.

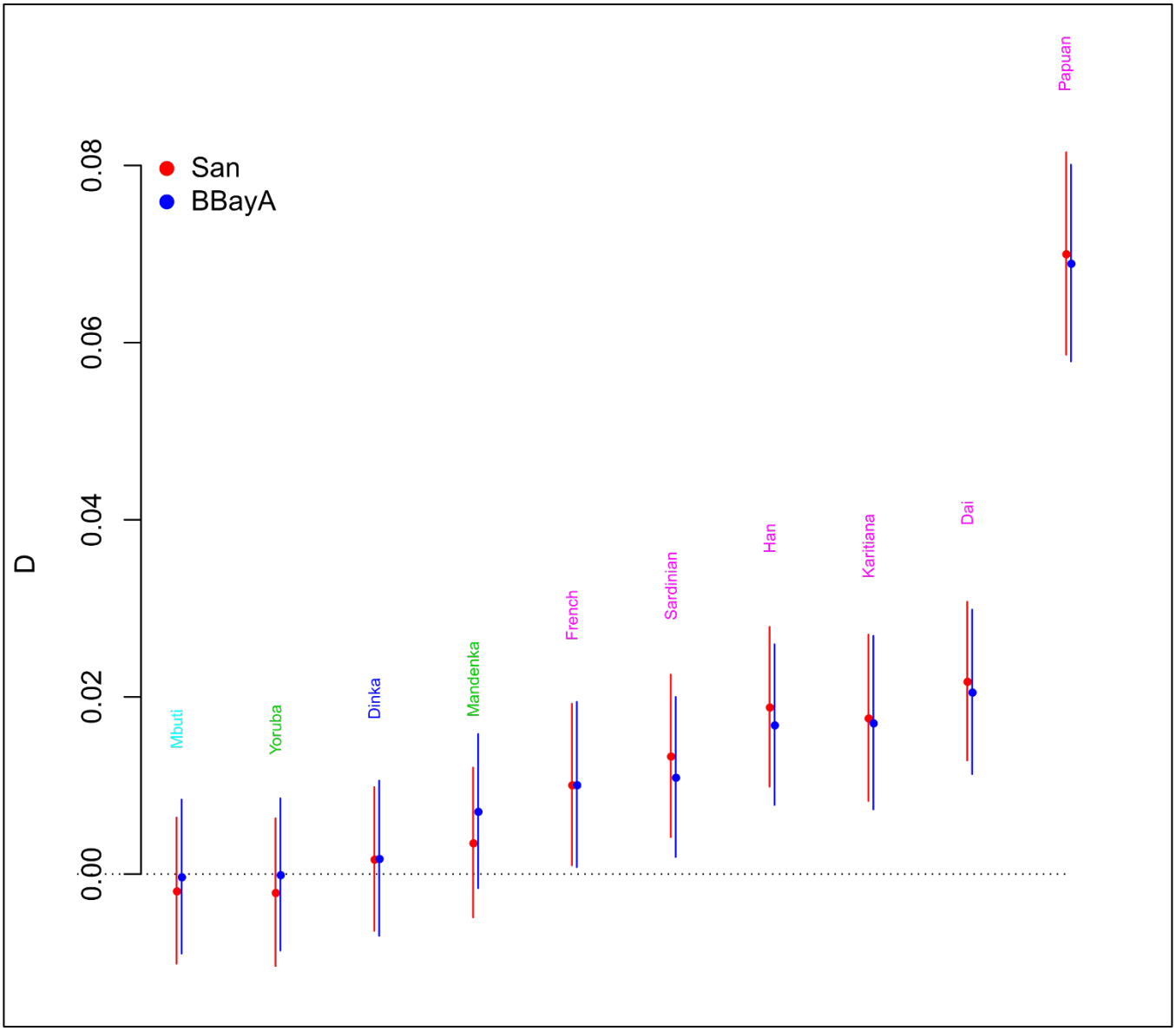

Figure S6.11: D-tests for testing admixture with P1=San or BBayA; P2 one of the 10 HGDP individuals and P3=Denisova.

References

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

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526:68-74

Barbieri C, Guldemann T, Naumann C, Gerlach L, Berthold F, Nakagawa H, Mpoloka SW, Stoneking M, Pakendorf B (2014) Unraveling the complex maternal history of Southern African Khoisan populations. Am J Phys Anthropol 153:435-448

Barnard A (1992) Hunters and herders of southern Africa - A comparitive ethnography of the Khoisan peoples. Vol 85. Cambridge University Press, Cambridge

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7

Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwangler A, et al. (2016) The genetic history of Ice Age Europe. Nature 534:200-205

Gallego Llorente M, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, Curtis MC, Stock JT, Coltorti M, Pieruccini P, Stretton S, Brock F, Higham T, Park Y, Hofreiter M, Bradley DG, Bhak J, Pinhasi R, Manica A (2015) Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science 350:820-822

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, et al. (2015) The African Genome Variation Project shapes medical genetics in Africa. Nature 517:327-332

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801-1806

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, et al. (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409-413

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, et al. (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538:201-206

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222-226

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

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D (2012) Ancient admixture in human history. Genetics 192:1065-1093

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Guldemann T, Kure B, Mpoloka SW, Nakagawa H, Naumann C, Lipson M, Loh PR, Lachance J, Mountain J, Bustamante CD, Berger B, Tishkoff SA, Henn BM, Stoneking M, Reich D, Pakendorf B (2012) The genetic prehistory of southern Africa. Nat Commun 3:1143

Pickrell JK, Patterson N, Loh PR, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D (2014) Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A 111:2632-2637

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

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43-49

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, et al. (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505:87-91

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 4:137 – 138

Schlebusch CM, de Jongh M, Soodyall H (2011) Different contributions of ancient mitochondrial and Y-chromosomal lineages in 'Karretjie people' of the Great Karoo in South Africa. J Hum Genet 56:623-630

Schlebusch CM, Lombard M, Soodyall H (2013) MtDNA control region variation affirms diversity and deep sub-structure in populations from Southern Africa. BMC Evol Biol 13:56

Schlebusch CM, Prins F, Lombard M, Jakobsson M, Soodyall H (2016) The disappearing San of southeastern Africa and their genetic affinities. Hum Genet 135:1365-1373

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, Li S, De Jongh M, Singleton A, Blum MG, Soodyall H, Jakobsson M (2012) Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. Science 338:374-379

Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, Clark GR, et al. (2016) Genomic insights into the peopling of the Southwest Pacific. Nature 538:510-513

Supplementary Information section 7 - Diversity estimates and demographic inferences

7.1 Diversity estimates - Heterozygosity

We compared the proportion of heterozygote sites for the HGDP individuals and BBayA for sites with i) >1x coverage ii) >7x coverage and iii) sites with a coverage >13x and within the 99.95% of the coverage distribution (this cuts off the high and low tail of the coverage distribution and adapted to the specific coverage distribution of an individual in order to avoid regions with unexpected coverage). We also limited the data to sites where the ancestral state could be confidently called (no more than 2 variants including the three apes, no missing data and the 3 apes showing the same variant). The effect of coverage on heterozygosity can be seen in Figure S7.1. We note that BBayA has levels of heterozygosity similar to most other African groups, but that modern-day San (Ju'|hoansi) have greater heterozygosity compared to other African groups.

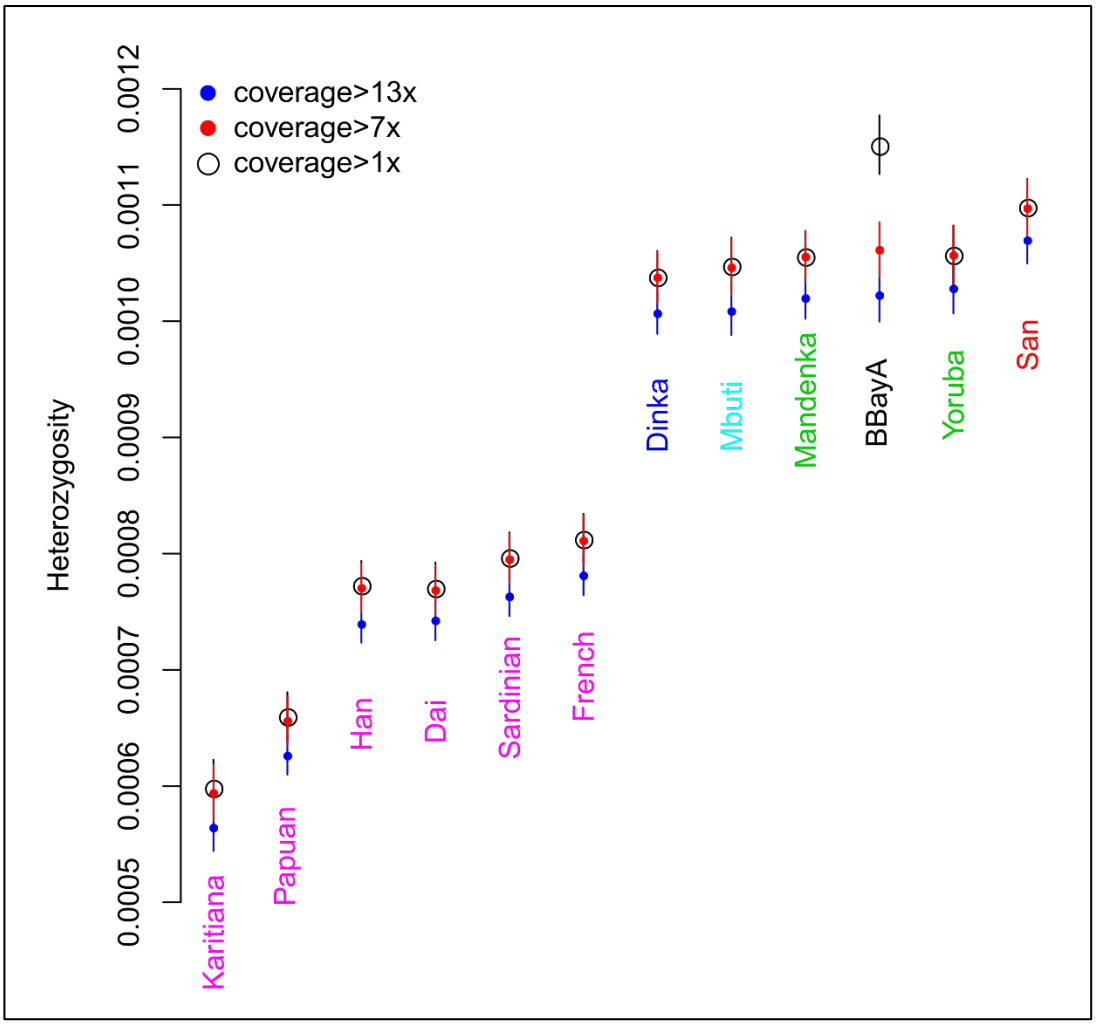

Figure S7.1: Heterozygosity estimates based on sites of different qualities.

7.2 Diversity estimates - Runs of Homozygosity

The distributions of Runs of Homozygosity (RoH) of the data were computed using Plink (v. 1.9) (Chang et al. 2015) with the following parameters: sliding windows of 50 SNPs, allowing 1 heterozygote per window, with an overlapping proportion of 0.05, final window sizes of at least 200 kb and 200 SNPs with a minimum SNP density of 1 in 20kb and a gap of 50 kb between SNPs before the run of homozygosity is split in two. Ballito Bay A had among the longest RoH results among Khoe-San individuals, suggesting lower diversity and lower N_e in Ballito Bay A compared to modern-day Khoe-San groups (Figure S7.2).

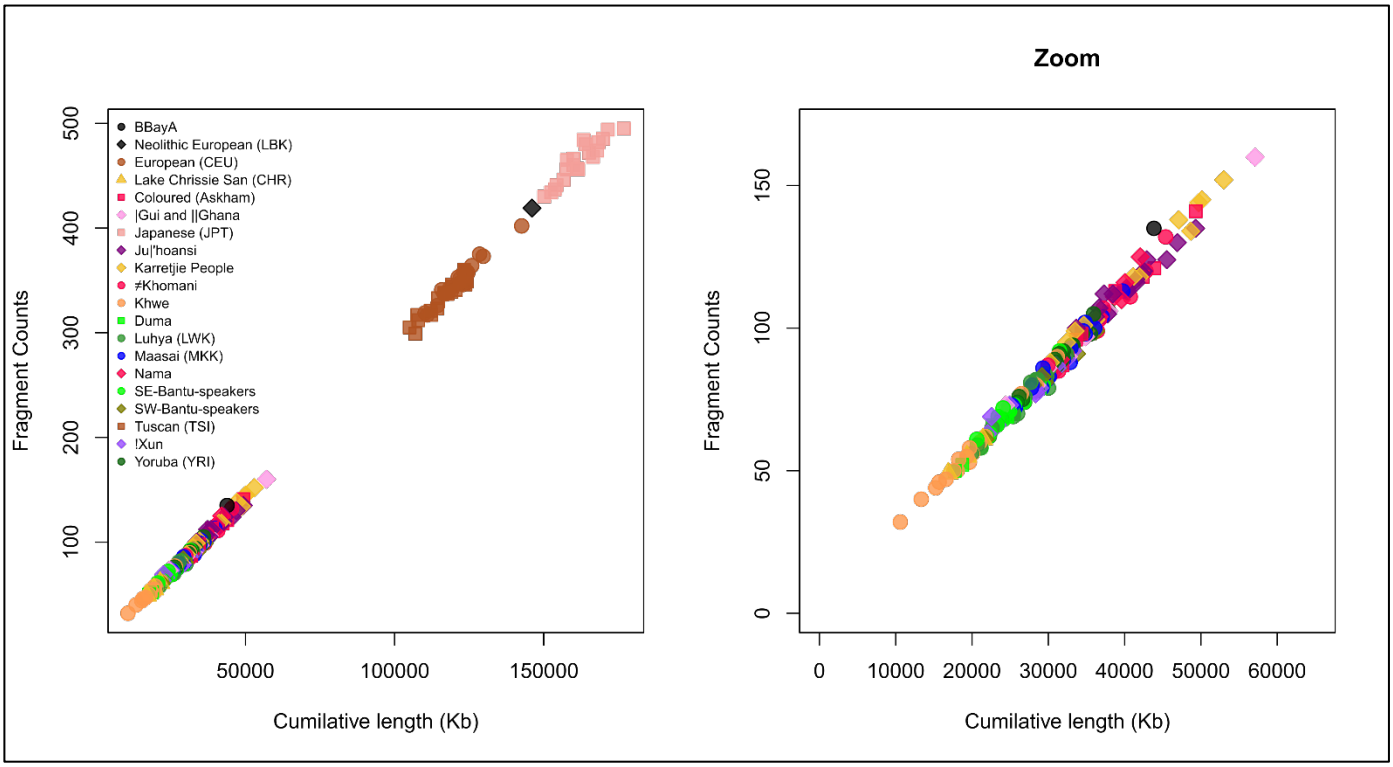

Figure S7.2: RoH of the KGP extended dataset. The cumulative length of RoH (x-axis) plotted against the number of RoH fragments (Y-axis) for the shortest RoH class (200-500Kb). Left: Including non-Africans and the LBK Neolithic European. Right: Zoom-in on African samples with BBayA (black dot) showing among the greatest cumulative lengths of RoH among African samples.

7.3 Demographic inferences - MSMC

To infer past effective population sizes for BBayA and compare it to a number of high-coverage modern-day genomes, we use MSMC's implementation of PSMC' (Schiffels and Durbin 2014). Input files were created using a set of scripts provided with MSMC (). MSMC was run with default parameters except for -r 0.88 in order to represent the ratio of recombination and mutation rate for humans and --fixedRecombination. We plot the effective population size for BBayA together with the HGDP individuals assuming a mutation rate

of 1.25x10e^{-8} per site per generation and a generation time of 30 years (Figure S7.3). The curve for BBayA is shifted according to the radiocarbon date of the individual.

Starting from the past, all populations start reducing their effective population size around 150 kya. Non-African populations go through a drastic reduction in N_e, probably representing the out-of-Africa migration bottleneck. African populations have higher population sizes during this time, but still show signs of a weaker bottleneck, except the San. BBayA's population size more recent than 100 kya is very similar to West Africans (Yorubans, Mandenka) and Dinka. Notably, Mbuti and modern San have greater population sizes during this period. Around 30 kya, BBayA's estimated effective population size starts to increase, which could be an effect of residual deanimation and/or mapping errors in the ancient sample.

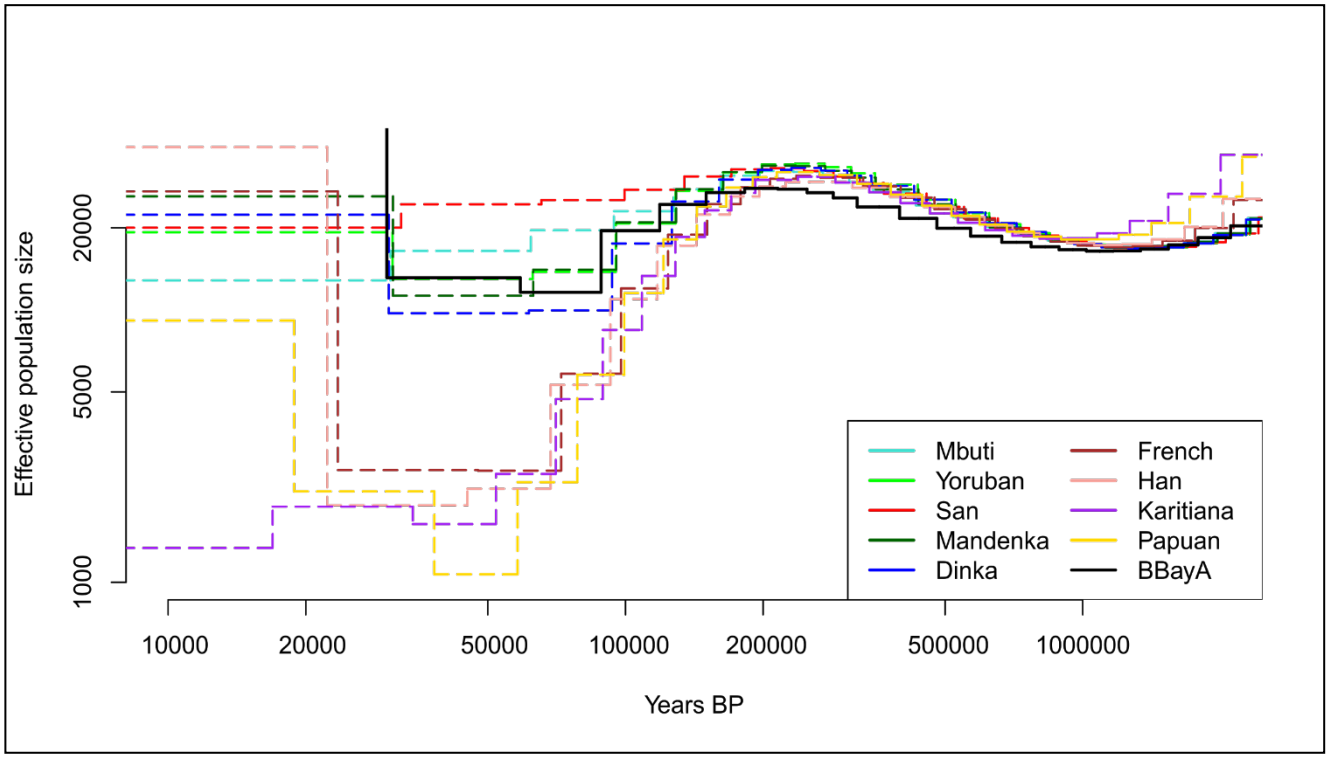

Figure S7.3: MSMC plot of 11 HGDP genomes together with the diploid full genome of Ballito Bay A.

References

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7

Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nat Genet 46:919-925

Supplementary Information section 8 - Dating of population split times (G-PhoCS)

8.1 Data

We used the diploid genotypes of Ballito Bay A together with the 11 HGDP genomes to estimate pairwise population split times and effective population sizes, through coalescence based analyses using G-PhoCS (Gronau et al. 2011). The sequence data for coalescence analysis were prepared according to the guidelines outlined in Gronau et al. (Gronau et al. 2011). Over 30,000 short sequence fragments were sampled from random positions across the autosomes. The length of the fragments was set to 1kb, which is a good length for human genomes, as it represents the optimal trade-off between minimizing the impact of recombination and maximizing information for coalescence analysis (Gronau et al. 2011). For filtering the fragments, we followed the guidelines and recommendations of (Gronau et al. 2011). Five filters were downloaded from the UCSC genome annotation database for hg19 (http://hgdownload.soe.ucsc.edu/goldenPath/ hg19/database/), which targets known genic regions (refGene, knownGene), simple and complex repeat regions (simpleRepeat, genomicSuperDups) and CpG islands (cpgIslandExt). In addition, we also compiled a filter from our own called INDEL regions in the dataset. Positions were set to missing using the 6 different filters and thereafter 1kb fragments containing more than 10% missing data were filtered out. The pipeline thus contained the following steps; random sampling of 1kb fragments from the autosomes, marking positions present in filters as missing, filtering out of fragments containing over 10% missing data and converting the data to the right input format for G-PhoCS. This pipeline is then run until over 30,000 fragments were obtained. The exact number of fragments used in the G-PhoCS run was 32,569 fragments.

8.2 Split times (Tau) in modern humans

G-PhoCS was run for all pairwise combinations of the individuals from the HGDP dataset and Ballito Bay A. Default input parameters were used except that the data were logged every 20 steps instead of every 10. No migration bands were added. The MCMC was run for 200,000 iterations and the first 50,000 were discarded as burn in. The visualization of the trace files showed that both the inferred split time (Tau) and population size (Theta) had already stabilized before reaching the burn-in cut-off.

Mean and median split times (Tau) were calculated for the 150,000 remaining logs of Tau after the burn-in was removed and are visualized as mean split times together with standard deviations as bar plots (Figure S8.1). To convert Tau to calendar years, a mutation rate of 1.25×10^{-8} per site per generation was used and a generation time of 30 years was assumed. Pairwise split times were grouped according to hierarchical split times (Table S8.1) and visualized as violin plots (Figure S8.2) using the vioplot package in R.

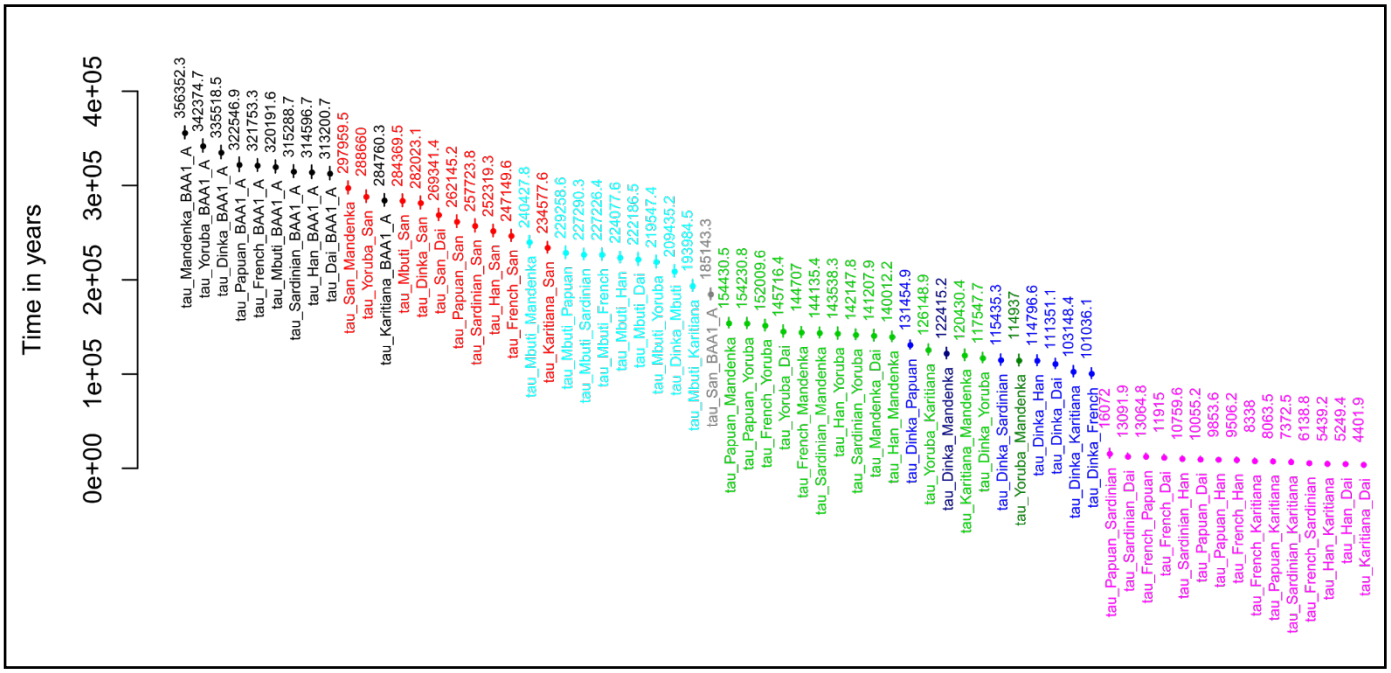

Figure S8.1: Means (dots) and standard deviations (bars) of G-PhoCS pairwise population split times, sorted descending. Colors are according to the hierarchical split times: Ballito Bay A (BAA) vs. all non-San (Black); HGDP-San vs. all non-San (Red); Mbuti vs. all non-San (Turquoise); Ballito Bay A (BAA) vs. HGDP-San (Gray), West Africans (Mandenka and Yoruba) vs. non-Africans (Green) and East Africans (Dark Blue), West Africans vs. West Africans (Dark Green); East Africans vs. non-Africans (Blue); non-Africans from each other (Pink).

Table S8.1 – Mean split times estimated by G-PhoCS.

	BBayA vs. Non-San	**HGDP San vs. Non-San**	**Mbuti vs. Non-San**	**BBayA vs. HGDP San**	**West Afr vs. East Afr& Non-Afr**	**East Afr vs. Non-Afr**	**Non-Afr vs. Non-Afr**
Tau (mean)	1.34441	1.115112	0.922886	0.77143	0.573226	0.470293	0.0387
Tau (SD)	0.081381	0.084897	0.058137	0.025313	0.059693	0.04728	0.017062
Split time mean (Gen)	10,755.30	8,920.90	7,383.10	6,171.40	4,585.80	3,762.30	309.6
Split time mean (Years)	322,658.40	267,626.90	221,492.70	185,143.30	137,574.30	112,870.40	9,288.10
Split time median (Gen)	10,709.00	8,888.50	7,453.00	6,179.30	4,685.80	3,737.70	291
Split time median (Years)	321,271.20	266,654.40	223,591.20	185,378.40	140,575.20	112,130.40	8,728.80
Split time mean Lower (Gen)	10,104.20	8,241.70	6,918.00	5,968.90	4,108.30	3,384.10	173.1
Split time mean Upper (Gen)	11,406.30	9,600.10	7,848.20	6,373.90	5,063.40	4,140.60	446.1
Split time mean Lower (Years)	303,126.90	247,251.50	207,539.90	179,068.30	123,248.00	101,523.20	5,193.20
Split time mean Upper (Years)	342,189.80	288,002.30	235,445.60	191,218.40	151,900.70	124,217.60	13,383.00

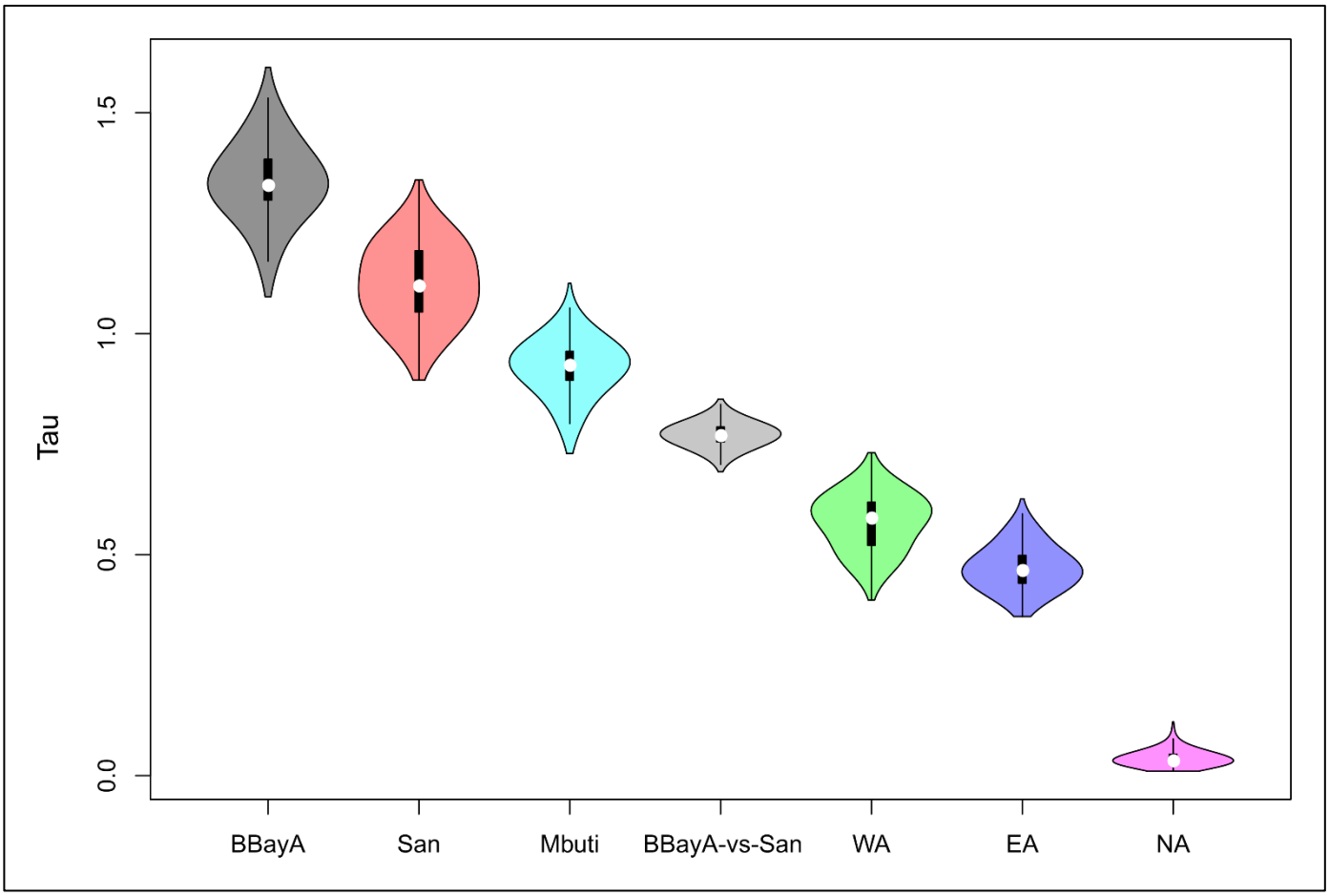

Figure S8.2: Violin plot of G-PhoCS pairwise population split times. Y-axis: Tau (Time in generations = Tau / (10,000×mutation rate)). X-Axis: Groupings are according to the hierarchical split times: Ballito Bay A vs. all non-San (Black); HGDP-San vs. all non-San (Red); Mbuti vs. all non-San (Turquoise); Ballito Bay A vs. HGDP-San (Gray), West Africans (Mandenka+Yoruba) vs. non-Africans and East Africans (Green), East Africans vs. non-Africans (Blue); non-Africans from each other (Pink).

8.3 Ne (Theta) in humans

For each pairwise comparison, G-PhoCS also estimates N_e (Theta) for each population and the ancestral population of the pair. N_e was calculated from Theta for the 150,000 remaining logs of Theta after the burn-in was removed. To convert Theta to N_e, a mutation rate of 1.25×10^{-8} and a generation time of 30 years was assumed. Mean N_e and standard deviation of focus populations (Fig S8.3) and their ancestral populations are visualized as bar plots in Figure S8.4.

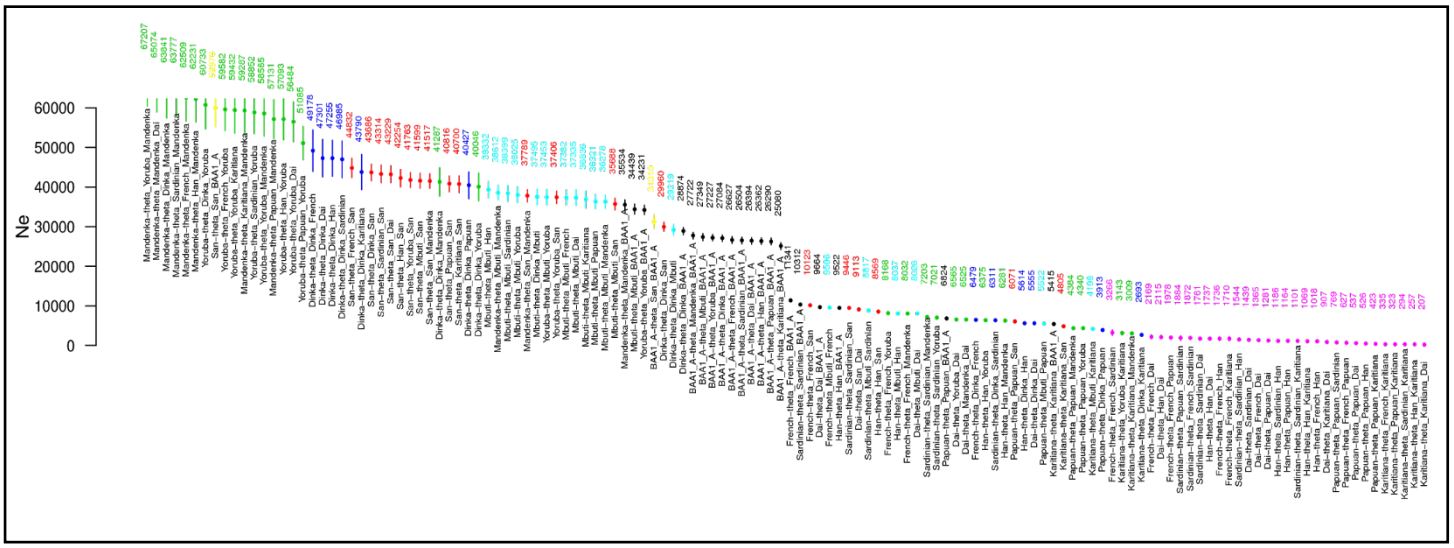

Figure S8.3: Means (dots) and standard deviations (bars) of G-PhoCS estimated effective population sizes for the populations of different pairwise population comparisons, sorted descending. Colors correspond to the specific split, see Figure S8.1.

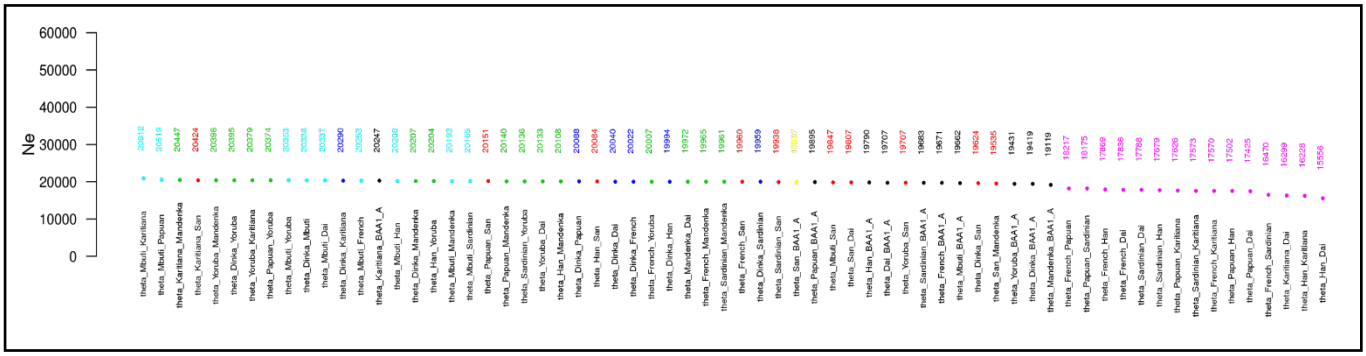

Figure S8.4: Means (dots) and standard deviations (bars) of G-PhoCS estimated effective population sizes for ancestral populations of different pairs of populations, sorted descending. Colors correspond to specific splits, see Figure S8.1.

8.4 Split times to Neandertals

We also analyzed the Altai Neandertal (Prüfer et al. 2014) and compared it to the diploid call-set of Ballito Bay A and the 11 HGDP genomes, with G-PhoCS. Neandertal SNP calling and filtering is described in section S6.1. Additional G-PhoCS specific filtering and run settings were the same as described in 8.1-8.2. The estimated split times and standard deviations are shown in Figure S8.5 and listed in in Table S8.2. Neandertal split times were in general older for comparisons with Africans compared to non-Africans, likely due to archaic admixture in non-Africans (Prüfer et al. 2014). The split with Ballito Bay A is the oldest and is around 10,000 years older than the comparison with HGDP San. The split times estimates against Neandertal are in

general younger than dates estimated with the TT method, see section S9.1 and slightly younger than dates reported previously (i.e. 553,000-589,000 years ago (Prüfer et al. 2014)).

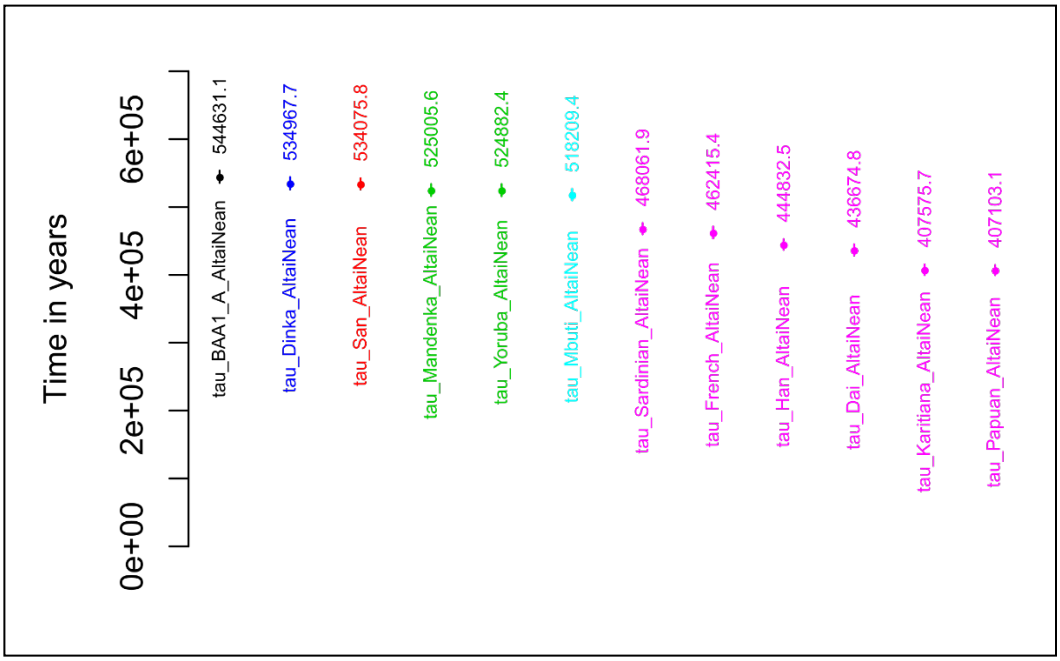

Figure S8.5: Mean and standard deviation of estimated split times of Ballito Bay A (BAA) and 11 individuals from the HGDP panel against Altai Neandertal.

Table S8.2: Mean and standard deviation (SD) of estimated split times (Tau) of Ballito Bay A and 11 individuals from the HGDP panel against Altai Neandertal.

Comparison	Tau	Tau (years)	Tau SD	Tau SD (years)
Ballito Bay A, Altai Neandertal	2.269296	544631.1	0.037944	9106.5
Dinka, Altai Neandertal	2.229032	534967.7	0.038322	9197.2
San, Altai Neandertal	2.225316	534075.8	0.035127	8430.5
Mandenka, Altai Neandertal	2.187523	525005.6	0.036682	8803.7
Yoruba, Altai Neandertal	2.187010	524882.4	0.037056	8893.5
Mbuti, Altai Neandertal	2.159206	518209.4	0.035608	8545.9
Sardinian, Altai Neandertal	1.950258	468061.9	0.034810	8354.4
French, Altai Neandertal	1.926731	462415.4	0.036688	8805.1
Han, Altai Neandertal	1.853469	444832.5	0.032976	7914.2
Dai Altai Neandertal	1.819478	436674.8	0.034689	8325.4
Karitiana, Altai Neandertal	1.698232	407575.7	0.032645	7834.9
Papuan, Altai Neandertal	1.696263	407103.1	0.031114	7467.3

References

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

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43-49

Supplementary Information section 9 – estimations based on sample configuration frequencies

9.1 Inference under a split model with pairwise sampling – the TT method

We developed an approach for estimating population split times that involve samples of two gene copies from each of two populations (denoted the TT method - from Two plus Two). The approach builds on, and extends the 'concordance' approach in Schlebusch et al. (Schlebusch et al. 2012), Skoglund et al. (Skoglund et al. 2011) and Wakeley (Wakeley 2008) that estimates model parameters under a pure split model using single individual samples. In contrast to the 'concordance' approach, the TT method utilizes 2 gene copies from each of a pair of populations and relies on the frequencies of all possible sample configurations (there are 9 such sample configurations but only 7 that are variable) in order to estimate model parameters. The assumptions of the model is an infinite number of sites/small mutation rate per site, independence between sites, a pure split model (no migration between populations) and a constant population size for the ancestral population (of the two daughter populations). The TT approach does not rely on assumptions about i) the population size dynamics in the two daughter populations (more recent than the split event), ii) the mutation rate, or iii) the number of generations since the split time in either of the two daughter populations. The modeled and estimated parameters are: the number of generations from population 1 to the population split, T_1 (scaled by a per site and per generation mutation rate), the number of generations from population 2 to the population split, T_2 (scaled by a per site and per generation mutation rate), the probability of two gene copies not coalescing before the split in population 1, α_1, the probability of 2 gene copies not coalescing before the split in population 2, α_2, and the size of the ancestral population, θ_A (scaled by a per site and generation mutation rate). Each population branch in calendar years, t_1 and t_2, can be estimated by dividing T_1 and T_2 by the per site and per generation mutation rate and multiplying by an assumed generation time in years. The ancestral population size can be estimated by dividing θ_A by the per site and per generation mutation rate. The expected number of generations to coalesce, given that the two lineages (from a specific population) coalesce before the population split, is the only additional parameter that would affect the probability of the different sample configurations under this model. These probabilities are denoted V_1 (the value for population 1 multiplied by the mutation rate per site and per generation) and V_2 (the value for population 2 multiplied by the mutation rate per site and per generation). It is possible to derive closed formulas for the probabilities of all the possible sampling configuration in terms of α_1, α_2, T_1, T_2, θ_A, V_1 and V_2. Assuming two sampled gene copies from each of the two populations, we denote the possible sample configurations of derived variants as:

Configuration	number derived in sample 1	number derived in sample 2
O0	0	0
O1	1	0
O2	0	1
O3	2	0
O4	0	2
O5	1	1
O6	2	1
O7	1	2
O8	2	2

The probability for each of these sample configurations can be derived from considering the probability of the configuration conditioning on either i) all four lineages coalescing before reaching the split in each branch (this is an event with probability $(1-\alpha_1)(1-\alpha_2)$), ii) the lineages in sample 1 coalescing before T_1, but the lineages in sample 2 remain as separate lineages at T_2 (an event with probability $(1-\alpha_1)\alpha_2$), iii) the lineages in sample 2 coalescing before T_2, but the lineages in sample 1 remain as separate lineages at T_1 (an event with probability $\alpha_1(1-\alpha_2)$), iv) both samples remain as separate lineages until the split in each branch (an event with probability $\alpha_1\alpha_2$). We can then derive the following probabilities:

$$P(O1) = 2T_1 - 2(1-\alpha_1)(T_1 - V_1) + \frac{\theta}{3}\alpha_1(4-\alpha_2),$$

$$P(O2) = 2T_2 - 2(1-\alpha_2)(T_2 - V_2) + \frac{\theta}{3}\alpha_2(4-\alpha_1),$$

$$P(O3) = (1-\alpha_1)(T_1 - V_1) + \theta - \frac{\theta}{6}(4\alpha_1 + 2\alpha_2 - \alpha_1\alpha_2),$$

$$P(O4) = (1-\alpha_2)(T_2 - V_2) + \theta - \frac{\theta}{6}(2\alpha_1 + 4\alpha_2 - \alpha_1\alpha_2),$$

$$P(O5) = \frac{\theta}{3}2\alpha_1\alpha_2,$$

$$P(O6) = \frac{\theta}{3}(2-\alpha_1)\alpha_2,$$

$$P(O7) = \frac{\theta}{3}\alpha_1(2-\alpha_1),$$

$$P(O0 \square O8) = 1 - \sum_{i=1}^{7} P(Oi).$$

We denote the number of sites that display the specific sample configuration by:

m_0: number of sites that are O0,
m_1: number of sites that are O1,
m_2: number of sites that are O2,
m_3: number of sites that are O3,
m_4: number of sites that are O4,
m_5: number of sites that are O5,
m_6: number of sites that are O6,
m_7: number of sites that are O7,
m_8: number of sites that are O8.

The total number of sites is then M ($M=m_0+m_1+m_2+m_3+m_4+m_5+m_6+m_7+m_8$). We find the following estimators (the ^ of the estimators have been omitted for simplicity):

$$\alpha_1 = \frac{2m_5}{2m_6 + m_5},$$

$$\alpha_2 = \frac{2m_5}{2m_7 + m_5},$$

$$\theta = \frac{3}{M}\frac{(2m_6 + m_5)(2m_7 + m_5)}{8m_5},$$

$$T_1 = \frac{1}{M}\left(\frac{m_1}{2} + m_3 - \frac{(2m_6 + m_5)(6m_7 + m_5)}{8m5}\right),$$

$$T_2 = \frac{1}{M}\left(\frac{m_2}{2} + m_4 - \frac{(6m_6 + m_5)(2m_7 + m_5)}{8m_5}\right),$$

so that

$$N_A = \frac{3}{\mu M}\frac{(2m_6 + m_5)(2m_7 + m_5)}{8m_5},$$

$$t_1 = \frac{g}{\mu M}\left(\frac{m_1}{2} + m_3 - \frac{(2m_6 + m_5)(6m_7 + m_5)}{8m_5}\right),$$

$$t_2 = \frac{g}{\mu M}\left(\frac{m_2}{2} + m_4 - \frac{(6m_6 + m_5)(2m_7 + m_5)}{8m_5}\right),$$

where μ is the by the per site and generation mutation rate and g is the number of years per generations. Since

$\alpha_1 = \exp\left(\frac{T_1}{\theta_1}\right)$ and

$$\alpha_2 = \exp\left(\frac{T_2}{\theta_2}\right),$$

where θ_1 and θ_2 are the branch specific effective population sizes for population 1 and 2 (multiplied by the per site and generation mutation rate). θ_1 and θ_2 can be estimated as the branch specific effective population size for population 1 and 2 as

$$\theta_1 = \frac{T_1}{\ln(\alpha_1)} \text{ and}$$

$$\theta_2 = \frac{T_2}{\ln(\alpha_2)}.$$

Sjödin et al. (in preparation) described the full derivations of the probabilities for observing the different sampling configurations as well as solutions.

9.2 Split time estimates

We utilized a weighted block jackknife procedure with 5 Mb blocks to estimate the confidence intervals of the parameters. We applied this method to pairwise comparisons of Ballito Bay A and the 11 individuals from the HGPD panel. These 12 individuals are assumed to each represent a population. The genome data for all 12 individuals were filtered with the same criteria, including only retaining sites for which the 3 great apes displayed the same variant. Furthermore, we noticed an effect of genome coverage on the split time estimates, and restrict analyses to positions where both individuals (in a pairwise comparison) passed a coverage filter (≥13x and within 99.95% of the coverage distributions), as described above. The SNP-calling was conducted for each individual separately.

For a comparison between individual A and B, there are branch specific estimates of split time, drift and effective population size. We therefore refer to the estimates of these parameters in the branch leading to individual A (in this particular comparison) as the split time with individual A being ‘focal’ and individual B being ‘reference’ (and vice versa when B is focal and A is reference).

In view of the results in section 6, we tried to model the modern-day San individual’s genome (from the HGDP panel) as a combination of Ballito Bay A, Dinka and Sardinian genomes. Assuming that Ballito Bay A contributed 86%, Dinka contributed 9.66% and Sardinian 4.34%, we randomly sampled for each position a variant from the Ballito Bay A genome with probability 0.8614^2, one allele from the Ballito Bay A genome and one allele from the Dinka genome with probability $2\times0.8614\times0.0966$ and so forth to construct all possible combinations of genotypes (the probabilities to sample an allele from a particular genome correspond to the admixture proportions estimated in section 6). Here, sampling was done without replacement so that if both alleles were drawn from the same source then the site in the modeled genome would be heterozygote if the source genome was heterozygote at this position. This random sampling was reiterated independently for each individual this ‘modeled artificially admixed modern-day San’ (‘AS’ in the figures) was compared to.

The split time estimates based on all comparisons associated with a split between a Khoe-San individual and an individual with a non-Khoe-San origin are shown in figure S9.1 and table S9.1. The estimated population split times do decrease for the modeled modern-day San individual (AS), but relatively little, and the estimated drift parameter recaptures the estimated genetic drift for the San population relatively well (see below), suggesting that we capture some features of the modern-day San genomes by mixing genomic material from Ballito Bay A, Dinka and Sardinian. On the other hand, the estimated split times are lower for the AS genome compared to the modern-day San, suggesting that all features of the mixed modern-day San genome was not recreated by simply mixing genetic material from Ballito Bay A, Dinka and Sardinian. This observation is hardly surprising since Ballito Bay A is likely an ancestor to southern Khoe-San populations, not to northern Khoe-San populations such as the Ju|'hoansi (the HGDP San is a Ju|'hoansi individual), and Dinka is likely not a perfect representative of the East African source population. From this investigation of an 'artificially' admixed individual, we note that: i) small amounts of admixture have limited influence the estimates of split times, and ii) we qualitatively recapitulate the change in estimated genetic drift and population split times using an artificially admixed individual of genomic material from Ballito Bay A, Dinka and Sardinian.

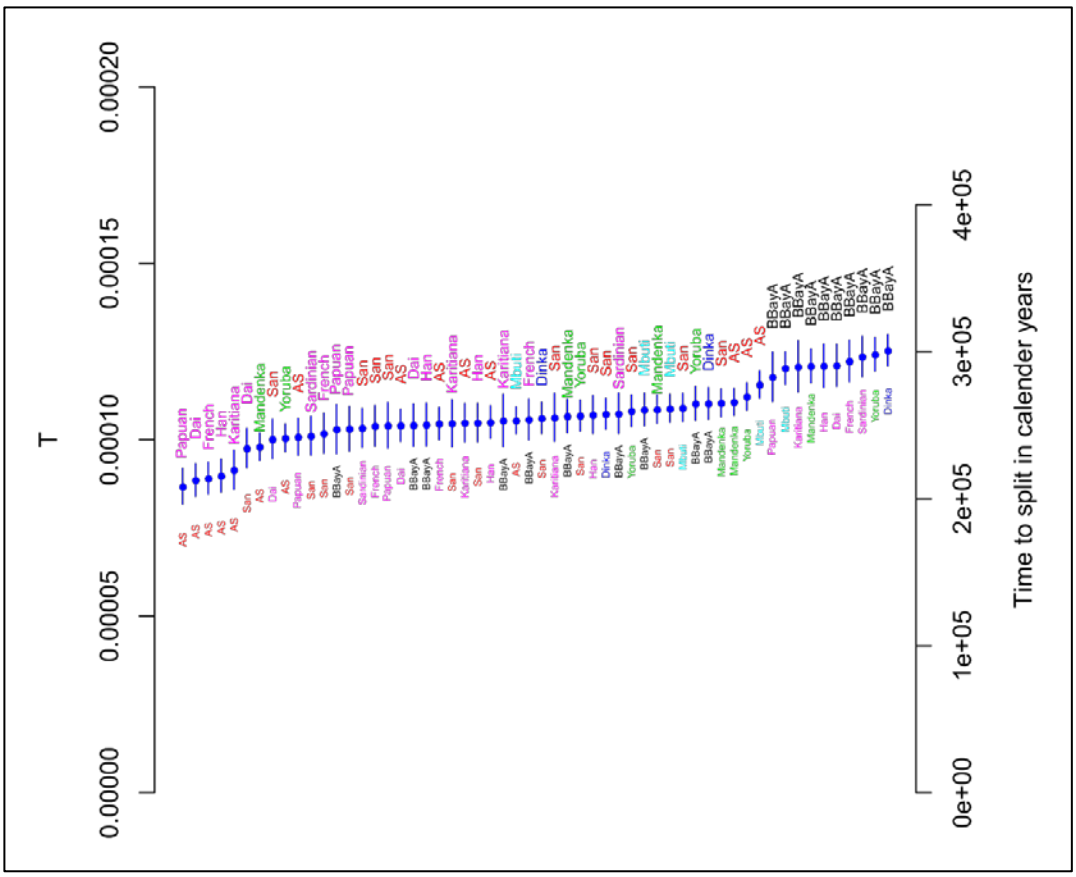

Figure S9.1: Estimates of split time between pairs of individuals. The populations displayed on top and in larger font are focal populations while the populations below in smaller font are the contrasting populations. We assume a mutation rate of 1.25×10^{-8} per site and generation, and a generation time of 30 years to translate the estimated parameter T to time in calendar years. In the figure, 'BBayA' refers to Ballito Bay A and 'AS' to the modeled admixed modern-day San.

We also estimated the split between the Altai Neandertal individual and the 12 non-archaic individuals (the 11 HGDP individuals and Ballito Bay A) as shown in figure S9.2. The estimates are older than those estimated by GPhoCS above, and also less affected by the small proportion of Neandertal admixture in non-African individuals, but overall on par with past estimates (Prüfer et al. 2014; Nielsen et al. 2017).

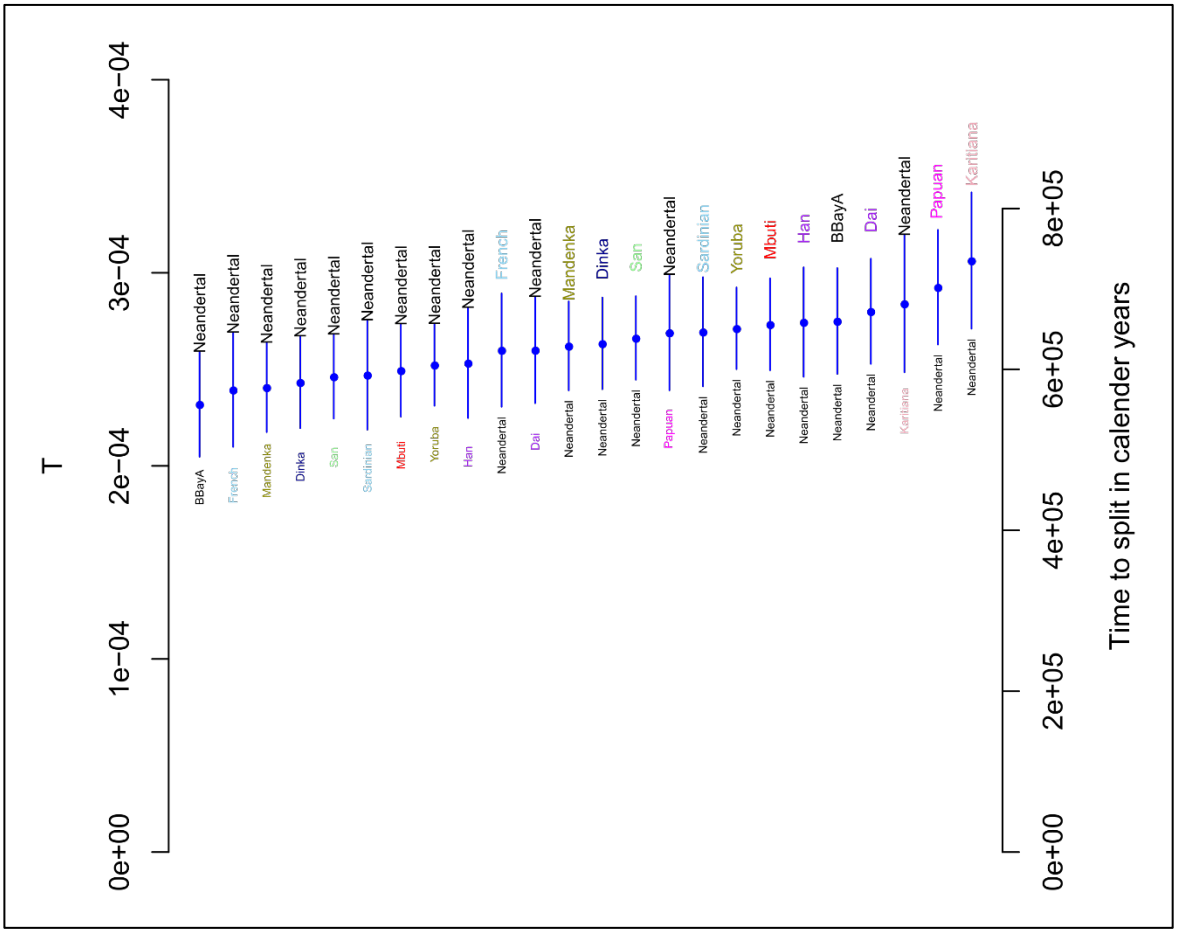

Figure S9.2: Estimation of split time between Altai Neandertal and other populations. The populations above and in larger font are focal while the populations below in smaller font are the contrasting populations. We assume a mutation rate of 1.25×10^{-8} per site and generation, and a generation time of 30 years to translate the estimated parameter T to time in calendar years. In the figure, 'BBayA' refers to Ballito Bay A.

The estimates of the deepest population split among humans (Khoe-San vs non-Khoe-San) using the Ballito Bay A individual consistently produced longer population branches from the Ballito Bay A genome compared to the estimates from the non-Khoe-San branch. Although this effect was mitigated by filtering out low coverage sites, it was not completely removed. This effect is possibly due to additional errors due to nature of ancient DNA, that include chemical lesions, mapping errors due to short reads, and more variable coverage compared to modern-day genome sequences. This (relatively small) effect of aDNA properties will affect the split time estimates in the Neandertal branch as well, however, such effects will also be counteracted by the

age of the remains. The Ballito Bay A individual dates to ~2,000 years ago while the Altai Neandertal individual has an age estimate of around 50,000 years, likely concealing some of the effects of aDNA errors. However, the TT method provides a novel way to overcome the issues with residual aDNA errors by estimating the population branch of a modern-day individual (as a focal group) in a pairwise comparison with an ancient individual.

Table S9.1: Estimated mean and standard deviation of split time assuming a 30 year generation time and a mutation rate of 1.25×10^{-8}. For the 2012 'concordance' method, an effective population size of 17,000 diploid individuals was assumed.

Split	Individuals compared (focal first)	G-PhoCS	TT method	2012 method
KSP North-South	San-BBayA	185,143 ± 6,075	155,917 ± 5,396	15,191 ± 6,660
KSP North-South	BBayA-San	185,143 ± 6,075	183,311 ± 5,289	100,972 ± 7,653
ooAfr	Dinka-Sardinian	115,435 ± 5,866	75,974 ± 6,455	19,012 ± 4,791
Deep human	Dinka-San	282,023 ± 6,802	254,816 ± 5,320	168,795 ± 5,965
Deep human	Dinka-BBayA	335,519 ± 6,989	264,902 ± 5,374	192,780 ± 5,844
Human-Neandertal	Neandertal-BBayA	544,631 ± 9,106	660,118 ± 32,905	659,824 ± 25,117
Human-Neandertal	Neandertal-San	534,076 ± 8,430	639,045 ± 25,961	544,279 ± 18,656
Human-Neandertal	Neandertal-Dinka	534,968 ± 9,197	632,278 ± 28,458	628346 ± 20,502

9.3 Branch specific drift

Using the TT-approach, we estimate the branch specific drift (figure S9.3) as well as branch specific effective population size (given estimates of branch specific drift and split time, the branch specific effective population size is the split time divided by the drift, figure S9.4).

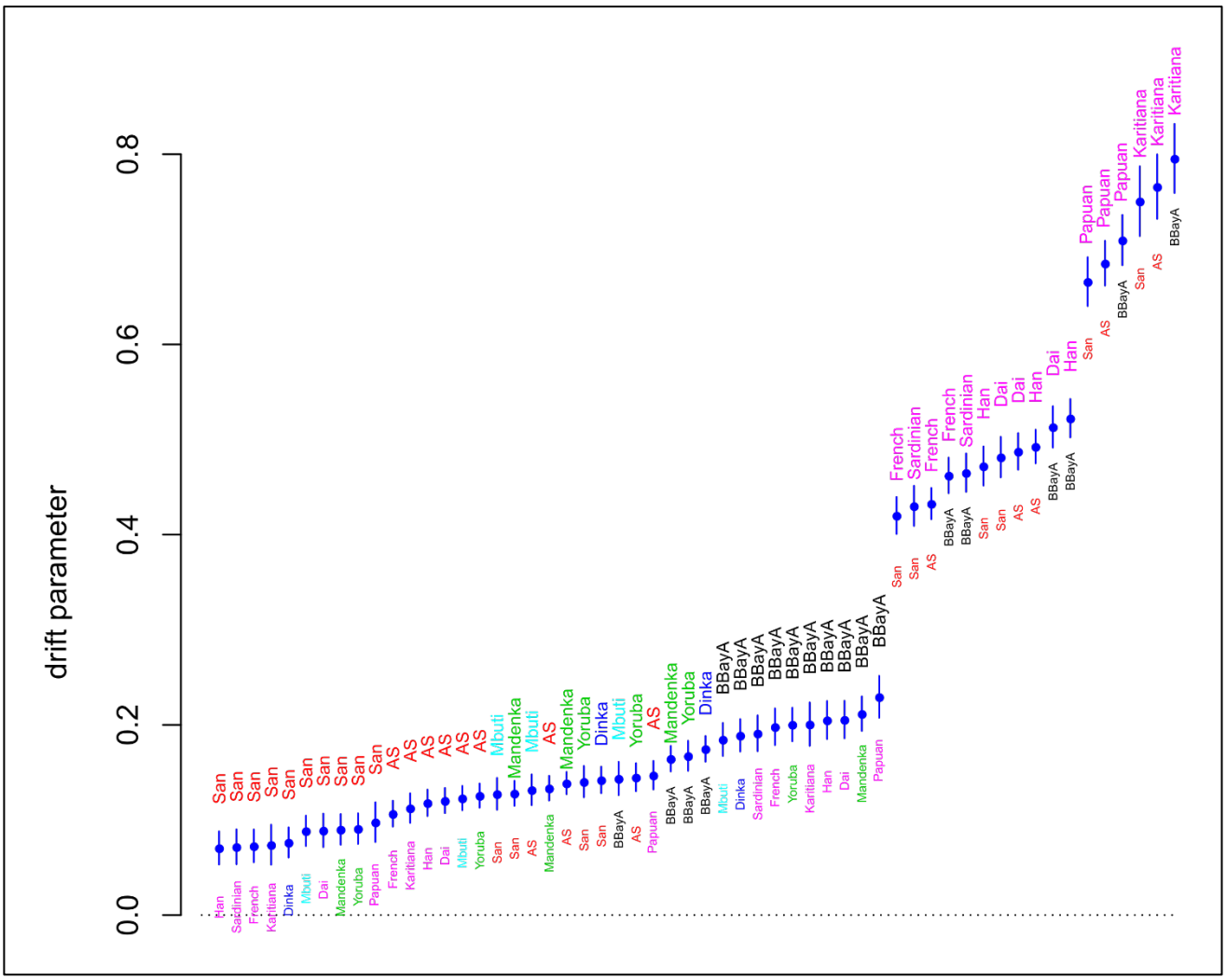

Figure S9.3: Estimation of branch specific drift until the split between Khoe-San populations and other populations. The populations above and in larger font are focal while the populations below in smaller font are the contrasting populations. In the figure, 'BBayA' refers to Ballito Bay A and 'AS' to the modeled admixed modern-day San individual.

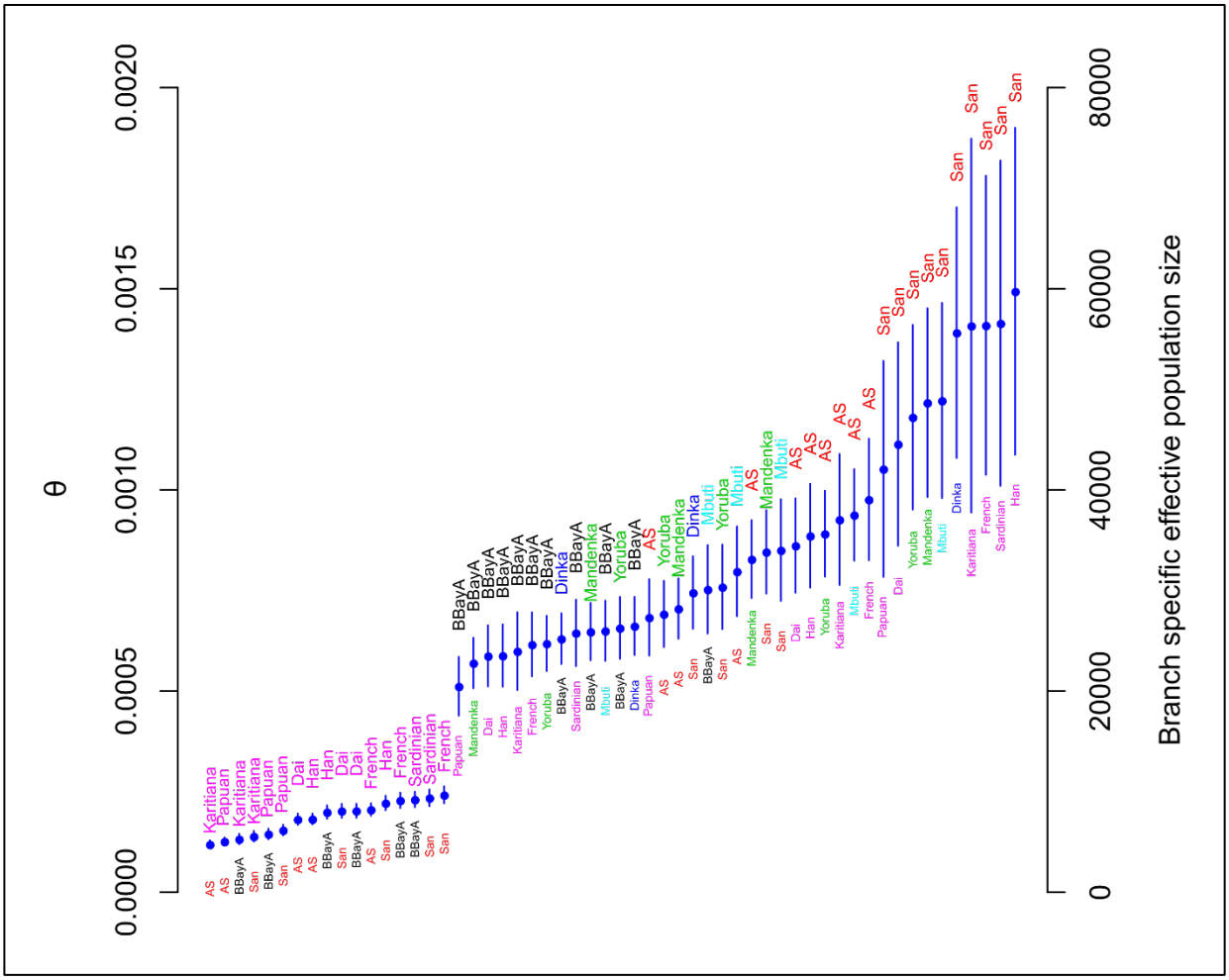

Figure S9.4: Estimation of effective size up until the split between Khoe-San populations and other populations. The populations above and in larger font are focal while the populations below in smaller font are the contrasting populations. We assume a mutation rate of 1.25×10^{-8} per site and generation, and a generation time of 30 years to translate the estimated parameter θ into a diploid effective population size. In the figure, 'BBayA refers to Ballito Bay A and 'AS' to the modeled admixed modern-day San individual.

In order to identify the modern population closest related to Ballito Bay A, and to compare the TT-method to other approaches, we calculated the genetic drift in the Ballito Bay A individual compared to several different modern-day Khoe-San individuals using the TT approach. We first compared him to the 6 Khoe-San individuals in (Mallick et al. 2016) together with the HGDP San (in total, 2 ≠Khomani and 5 Ju'|hoansi individuals). Here, only variable sites were required in order to calculate the drift parameter. Branch specific drift on the Ballito Bay A individual/branch is shown in Figure S9.5.

We also calculated genetic drift on the Ballito Bay A branch/individual when comparing to the Schlebusch et al. SNP-genotype data (Schlebusch et al. 2012). Here, because the TT-method explicitly models the mutation process, and SNP-genotype data are heavily ascertained, the TT-method is not suitable. Instead, if there is no admixture and the SNPs have been ascertained in non-Khoe-San populations, then all SNPs that are variable

in Khoe-San populations must have been present before the split between Khoe-San and other groups and the 'concordance' method described in Schlebusch et al. (Schlebusch et al. 2012) is more suitable than the TT method. Estimated drifts on the BBayA branch is shown in Figure S9.6.

Both these analyses (and supported by the outgroup-f3 analysis below) suggest that the Ballito Bay A individual shows greatest genetic affinity to southern Khoe-San groups of today (see also sections 6-8). The Ballito Bay A boy appears to be particularly closely related to the Karretjie People.

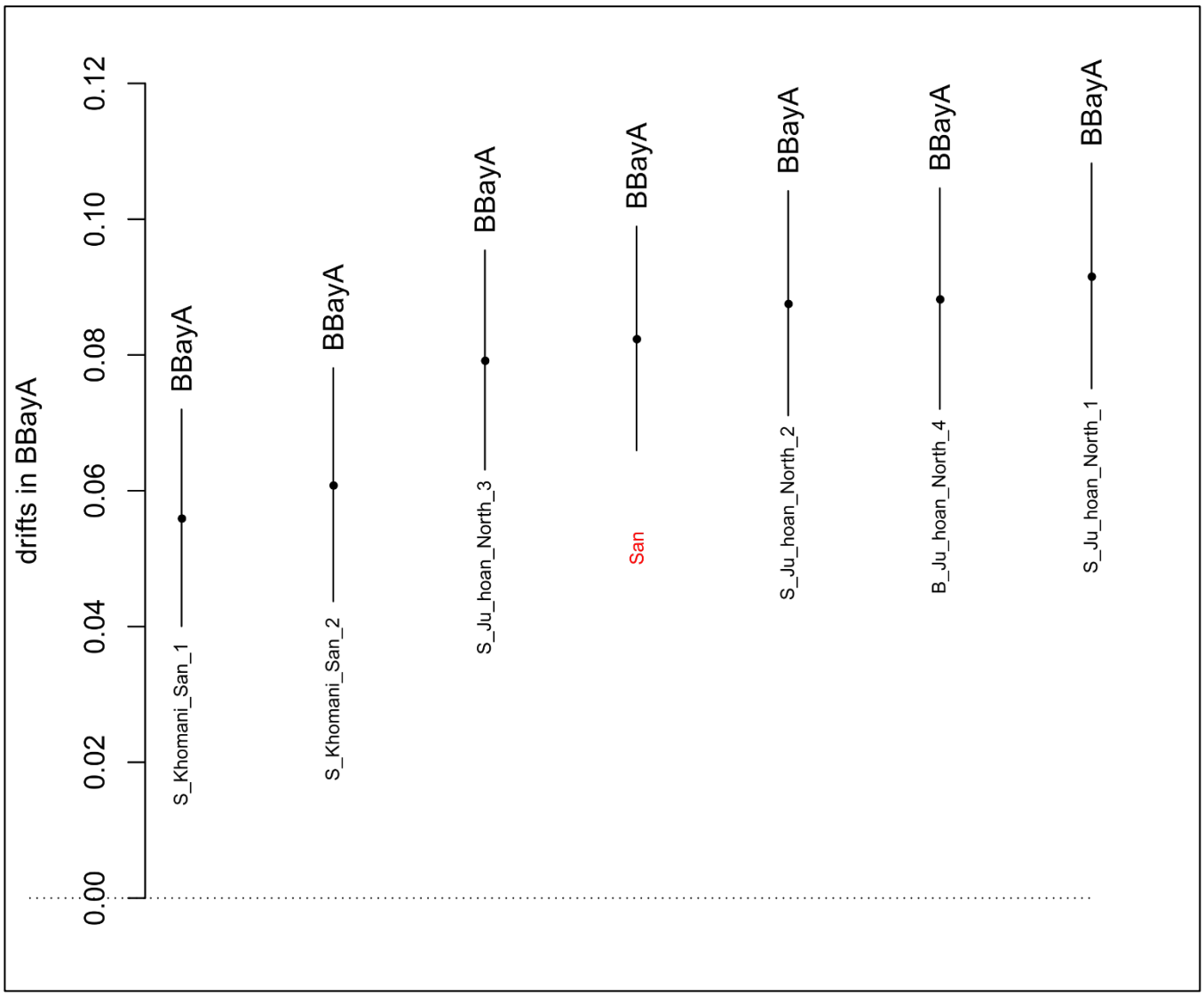

Figure S9.5: Genetic drift specific to Ballito Bay A compared to whole genome sequenced Khoe-San individuals ('San' from HGPD in red, ≠Khomani and Ju'|hoan from (Mallick et al. 2016)).

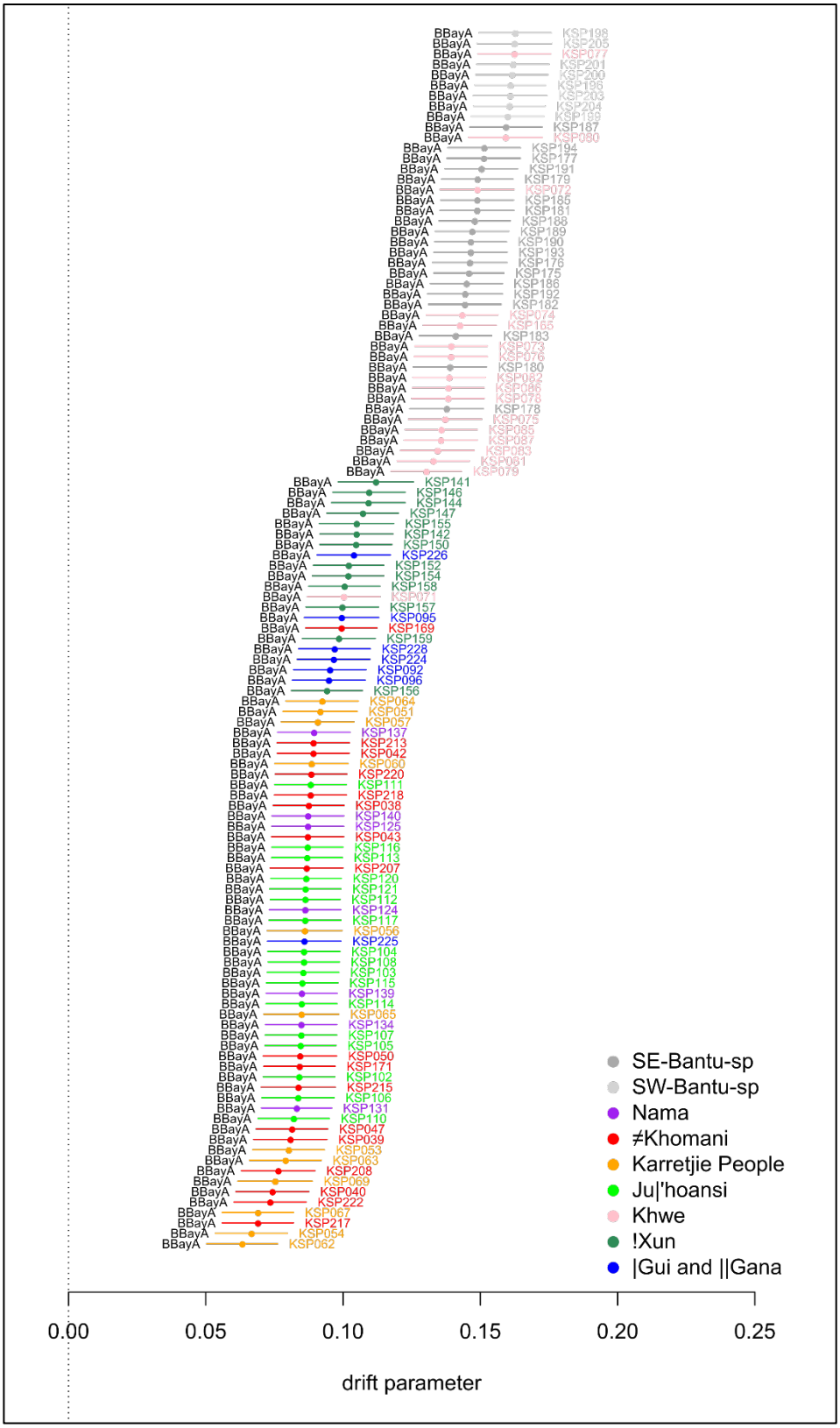

Figure S9.6: Genetic drift specific to Ballito Bay A when comparing to Khoe-San and Bantu speakers from Schlebusch et al. (2012), based on SNP-genotype data and the 'concordance' method of Schlebusch et al. (2012) to estimate branch-specific genetic drift.

9.4 Other measures of drift (F_{ST} and outgroup f3)

For reference, we estimated pairwise F_{ST} between Ballito Bay A and the 11 HGDP individuals with the same filtered data as for the TT analyses. We also estimated pairwise F_{ST} between Ballito Bay A and the Schlebusch et al. SNP-genotype data (Schlebusch et al. 2012). See Figures S9.7 and S9.8.

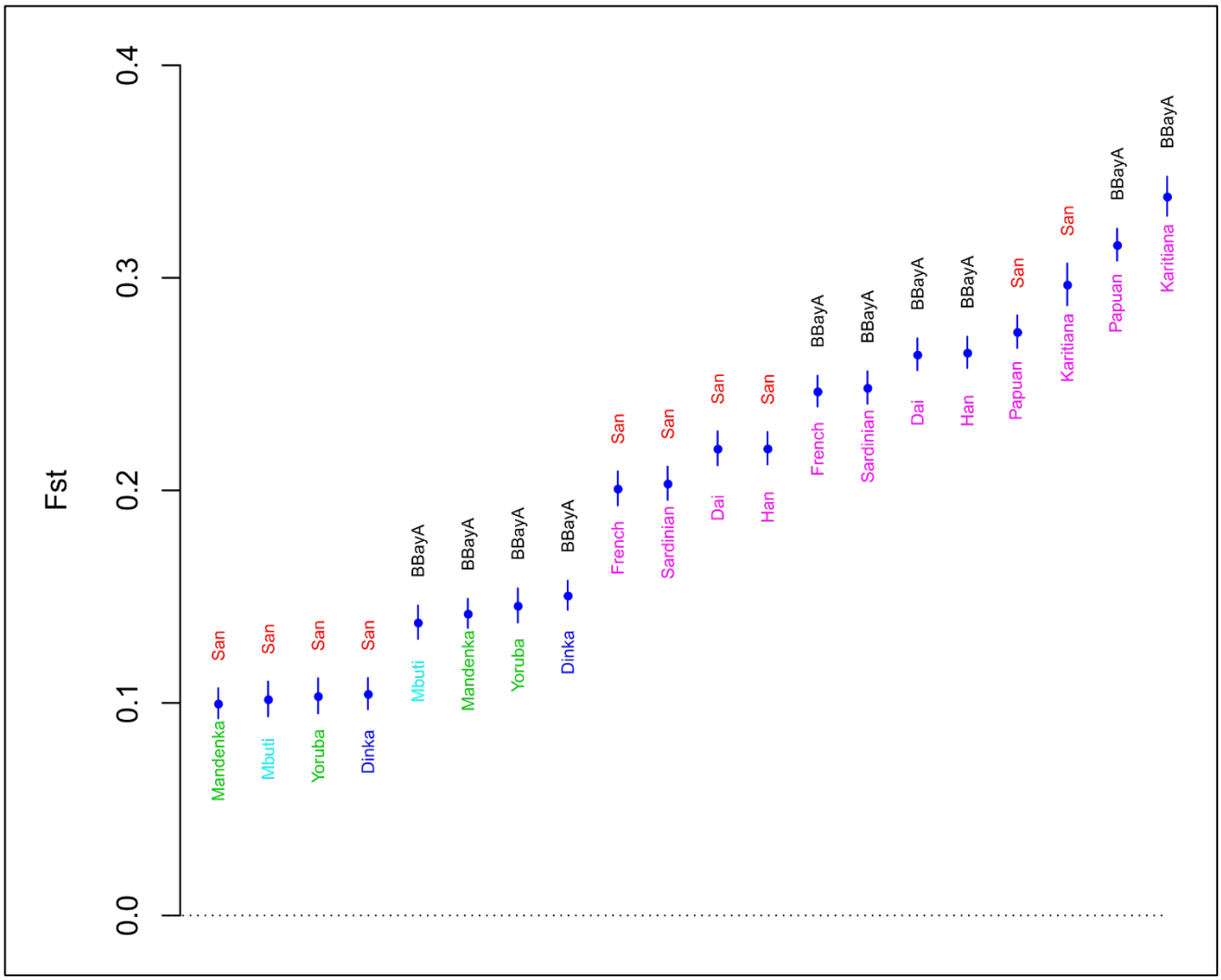

Figure S9.7: Pairwise F_{ST} between Ballito Bay A and the 11 HGDP genome-sequenced individuals.

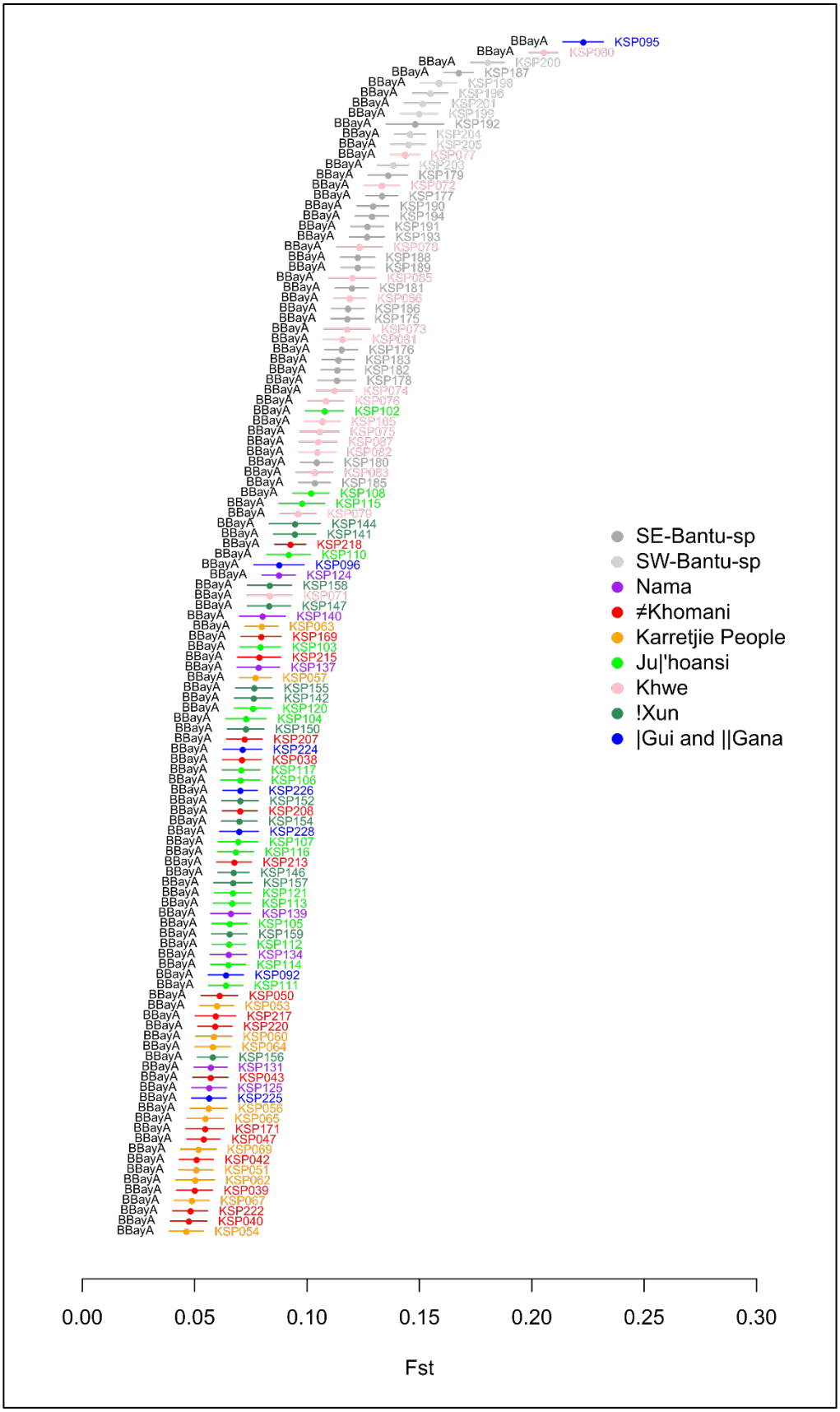

Figure S9.8: Pairwise F_{ST} between Ballito Bay A and the individuals in Schlebusch *et al.,* (2012).

Finally, we estimated shared drift as measured by outgroup f3 values between Ballito Bay A and the southern African dataset (Schlebusch et al. 2012) (Figure S9.9).

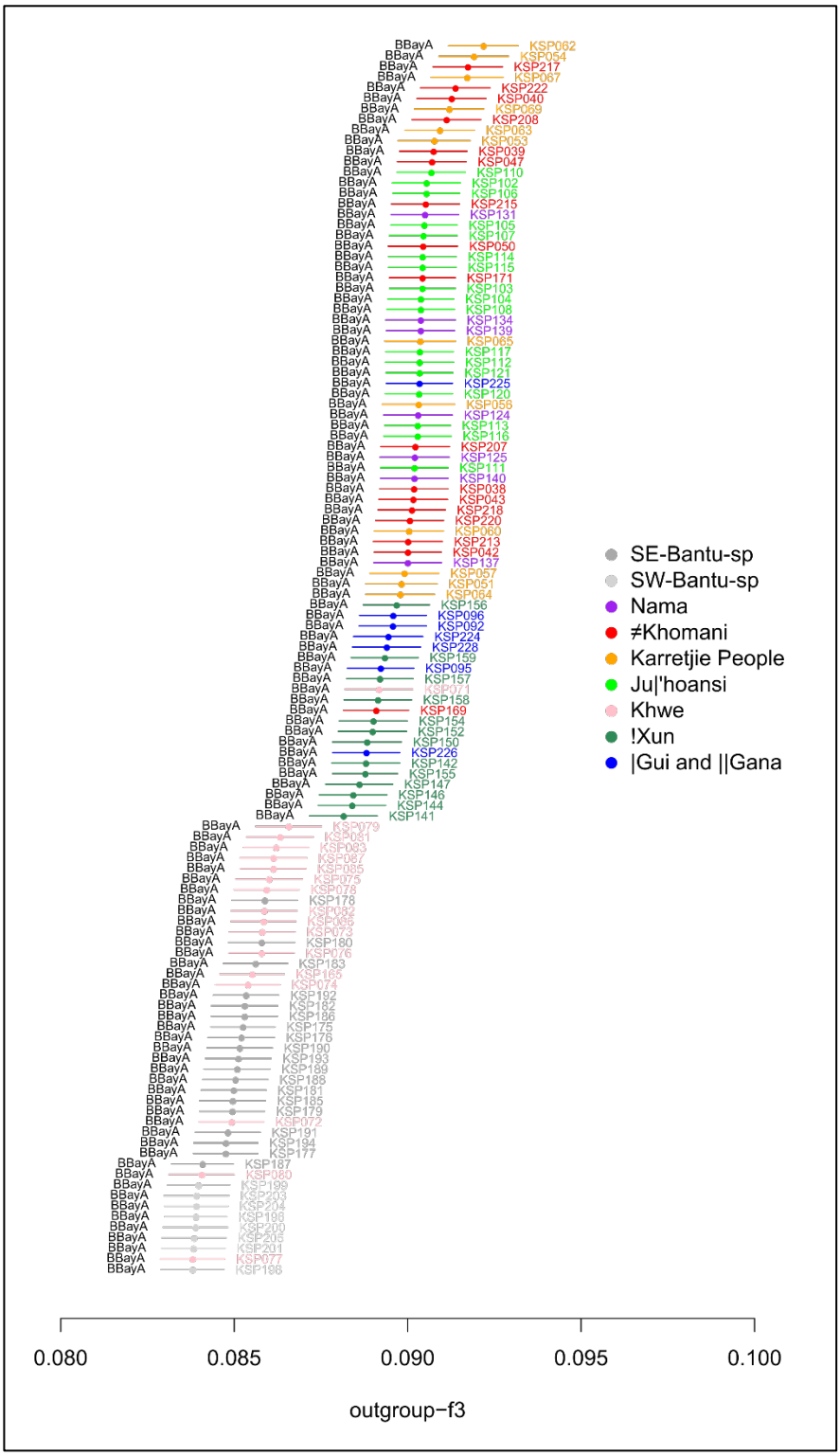

Figure S9.9: Outgroup-f3 between Ballito Bay A and the individuals in Schlebusch *et al.*, (2012) (Schlebusch et al. 2012).

The F_{ST} and outgroup f3 analyses comparing Ballito Bay A to the individuals in the Schlebusch et al. (2012) data both suggest that the Ballito Bay A individual is closer related to modern day southern Khoe-San individuals than he is to modern day northern Khoe-San individuals. Moreover, that F_{ST} between non-Khoe-San individuals and Ju'|hoansi (from the HGDP, figure S9.7) is lower compared to between the non-Khoe-San individuals and Ballito Bay A is consistent with admixture into Ju'|hoansi (from non-Khoe-San individuals), an admixture that is not present in the Ballito Bay A boy.

References

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, et al. (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538:201-206

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E (2017) Tracing the peopling of the world through genomics. Nature 541:302-310

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43-49

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, Li S, De Jongh M, Singleton A, Blum MG, Soodyall H, Jakobsson M (2012) Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. Science 338:374-379

Skoglund P, Gotherström A, Jakobsson M (2011) Estimation of population divergence times from non-overlapping genomic sequences: examples from dogs and wolves. Mol Biol Evol 28:1505-1517

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

Supplementary Information section 10 - Genomic regions of interest and selection

10.1 Variants of specific phenotypic interest

In order to investigate SNP variants associated with particular traits, we scanned the literature for specific sites and determined the alleles at these sites in the ancient samples using the samtools mpileup function (v1.3) (Li et al. 2009). Genes coding for traits of particular interest in African populations were analyzed (Fan et al. 2016), including the following genes/regions: i) the *MCM6* gene containing regulatory functions for the physically nearby *LCT* gene that produces lactose and is strongly associated with lactase persistence in adulthood (Fan et al. 2016), ii) *DARC*, *HBB*, *G6PD*, *ATP2B4*, and *APOL1* genes for resistance to malaria and African sleeping sickness (Genovese et al. 2010; Howes et al. 2011; Bedu-Addo et al. 2013; McManus et al. 2017), and iii) the *SLC24A5/A2*, *HERC2* and *OCA2* pigmentation genes (White and Rabago-Smith 2011; Sturm and Duffy 2012; Beleza et al. 2013). Either the OMIM or NCBI SNP directory was used to obtain the rs number, chromosome position and reference allele for each associated site in (or nearby) the genes. Chromosome positions in according to the hg19 reference sequence were used in the analysis.

All ancient southern African individuals (that had enough data) exhibited the reference SNP call for all lactase persistence genes (Table S10.1), and none of the samples displayed any variants that were linked to lactase persistence.

For malaria resistance, the alternative variants were found in Eland Cave (possibly heterozygote C/T), Mfongosi (possibly heterozygote C/T), and Newcastle (homozygote C) for the malarial resistance Duffy null allele (Table S10.1). Interestingly, all three ~300-500-year-old individuals (that have enough data) carry at least one Duffy null allele that has a strong protective effect against malaria (McManus et al. 2017), while the older samples do not carry the Duffy null allele. For the Duffy FY*A/B locus, the FY*B alleles were found in Champagne Castle (at least one allele is FY*B), ELA (possibly homozygote FY*B), MFO (homozygote FY*B), and NEW (homozygote FY*B). One of the ~2,000-year-old individuals that had enough data displayed the FY*A allele. The FY*A allele potentially has some protective effect against malaria compared to the FY*B allele (Howes et al. 2011), but this locus likely has less impact on malaria resistance than the variants at the Duffy null locus (McManus et al. 2017). The Duffy null allele is usually found on a FY*B background and therefore the high frequency of FY*B among the more recent individuals is not surprising. For the ATP2B4 gene variant, both alleles appear in both the old (~2,000) and the young (300-500) set of individuals. Taken together, these observations points to strong malaria protective variants existing in migrant Iron Age farmers (of West African origin) in contrast to southern African Stone Age hunter-gatherers.

Having at least one G allele for the *APOL1* gene SNP rs73885319 confers resistance to African sleeping sickness (Genovese et al. 2010). Eland Cave is heterozygous for this polymorphism and Newcastle is homozygous for the alternative variant (Table S10.1). This suggests that the protective variant was present in moderate frequency among southern African Iron Age farmers.

The *SLC24A5* G allele is near fixation in African populations (Sturm and Duffy 2012) and all individuals with enough data exhibit the alternative G variant for the *SLC24A5* gene SNP rs1426654, which codes for

darker skin color (Sturm and Duffy 2012). All individuals with enough data exhibit the ancestral C variant for the *SLC24A5* gene SNP rs16891982, which is also associated with darker skin pigmentation. All individuals (with enough data) present the ancestral allele for the *OCA2* and *HERC2* genes associated with eye color (Table S10.1), and the individuals were likely brown eyed (Sturm and Duffy 2012; Beleza et al. 2013).

Table S10.1 Variants associated with traits and the number of sequence reads (shown in parentheses after the allele) presenting the different alleles in study individuals. Non-reference alleles are highlighted in boldface.

Trait	Gene	SNP name	Position (chr:hg19pos)	Ref allele (hg19)	Allele in BBayA (depth)	Allele in BBayB (depth)	Allele in CHA (depth)	Allele in ELA (depth)	Allele in MFO (depth)	Allele in NEW (depth)
Lactase Persistence Middle East	MCM6	rs41525747	2:136608643	G	G (11)	NA	NA	G(8)	G(8)	G(11)
Lactase Persistence European	MCM6	rs4988235	2:136608646	G	G (11)	NA	NA	G(8)	G(8)	G(11)
Lactase Persistence Middle East	MCM6	rs41380347	2:136608651	A	A (11)	NA	NA	A(8)	A(7)	A(9)
Lactase Persistence East Africa	MCM6	rs145946881	2:136608746	C	C (11)	C(2)	C(1)	C(2)	C(5)	C(9)
Malaria resistance (Duffy null)	DARC	rs2814778	1:159174683	T	T (19)	T(1)	NA	T(5), **C(2)**	T(1), **C(4)**	**C(18)**
Malaria resistance (Duffy FY*A or B)	DARC	rs12075	1:159175354	G (FY*A)	G (10) (FY*A)	NA	**A(1)** (FY*B)	**A(5)** (FY*B)	**A(8)** (FY*B)	**A(8)** (FY*B)
Malaria resistance Sickle cell anemia	HBB	rs334	11:5226502	T	NA	NA	NA	NA	T(3)	NA
Malaria resistance	G6PD	rs1050828	X:153764217	C	C (4)	C(3)	NA	C(3)	C(7)	C(21)
Malaria resistance	ATP2B4	rs10900585	1:203654024	G	**T(15)**	G(1)	NA	G(3), **T(1)**	G(5), **T(2)**	**T(9)**
Resistance to African sleeping sickness	APOL1	rs73885319	22:36661906	A	A(16)	A(3)	A(1)	A(3), **G(4)**	A(15)	**G(18)**
Resistance to African sleeping sickness	APOL1	rs60910145	22:36662034	T	T(18)	T(2)	NA	T(3), **G(1)**	T(8)	**G(10)**
Resistance to African sleeping sickness	APOL1 (insertion/ deletion allele	rs71785313	22: 36662046: 36662051	- /TTATA A	I	I	I	I	I	I
Skin pigmentation	SLC24A5	rs1426654	15:48426484	A	**G(11)**	**G(5)**	NA	**G(7)**	**G(9)**	A(1),**G(18)**
Skin pigmentation	SLC45A2	rs16891982	5: 33951693	C	C(16)	C(1)	NA	C(10)	C(2)	C(12)
Eye color (brown)	HERC2	rs12913832	15:28365618	A	A(19)	A(1)	NA	A(4)	A(8)	A(17)
Eye color	OCA2	rs1800407	15: 28230318	C	C(22)	C(1)	NA	C(10)	C(2)	C(12)

References

Bedu-Addo G, Meese S, Mockenhaupt FP (2013) An ATP2B4 polymorphism protects against malaria in pregnancy. J Infect Dis 207:1600-1603

Beleza S, Johnson NA, Candille SI, Absher DM, Coram MA, Lopes J, Campos J, Araujo, II, Anderson TM, Vilhjalmsson BJ, Nordborg M, Correia ESA, Shriver MD, Rocha J, Barsh GS, Tang H (2013) Genetic architecture of skin and eye color in an African-European admixed population. PLoS Genet 9:e1003372

Fan S, Hansen MEB, Lo Y, Tishkoff SA (2016) Going global by adapting local: A review of recent human adaptation. Science 354:54-59

Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, Bowden DW, Langefeld CD, Oleksyk TK, Uscinski Knob AL, Bernhardy AJ, Hicks PJ, Nelson GW, Vanhollebeke B, Winkler CA, Kopp JB, Pays E, Pollak MR (2010) Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science 329:841-845

Howes RE, Patil AP, Piel FB, Nyangiri OA, Kabaria CW, Gething PW, Zimmerman PA, Barnadas C, Beall CM, Gebremedhin A, Menard D, Williams TN, Weatherall DJ, Hay SI (2011) The global distribution of the Duffy blood group. Nat Commun 2:266

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079

McManus KF, Taravella AM, Henn BM, Bustamante CD, Sikora M, Cornejo OE (2017) Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. PLoS Genet 13:e1006560

Sturm RA, Duffy DL (2012) Human pigmentation genes under environmental selection. Genome Biol 13:248

White D, Rabago-Smith M (2011) Genotype-phenotype associations and human eye color. J Hum Genet 56:5-7