Article

Title:
Ancient tree-topologies and gene-flow processes among human lineages in Africa

Short Title:
Ancient demographic histories of African genetic lineages

Authors:
Gwenna Breton 1,2*
Per Sjödin 1
Panagiotis I. Zervakis 1
Romain Laurent 3
Alain Froment 4
Agnès E. Sjöstrand 1
Barry S. Hewlett 5
Luis B. Barreiro 6
George H. Perry 7,8
Himla Soodyall 9,10
Evelyne Heyer 3
Carina M. Schlebusch 1,11,12,*
Mattias Jakobsson 1,11,12,*
Paul Verdu 3,*

Authors affiliations:
1 Human Evolution, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18C, SE-752 36 Uppsala, Sweden
2 Centre for Medical Genomics, Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital, Gothenburg, Sweden
3 UMR7206 Eco-anthropology, CNRS-MNHN-Université Paris Cité, Paris, France
4 UMR 208 Patrimoines locaux, IRD-MNHN, Paris, France
5 Department of Anthropology, Washington State University, Vancouver, Washington, USA
6 Department of Medicine, Section of Genetic Medicine, University of Chicago, Chicago, IL, USA
7 Department of Anthropology, Pennsylvania State University, University Park, PA 16801, USA
8 Department of Biology, Pennsylvania State University, University Park, PA 16801, USA
9 Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand and National Health Laboratory Service, Johannesburg, South Africa
10 Academy of Science of South Africa
11 Palaeo-Research Institute, University of Johannesburg, P.O. Box 524, Auckland Park, 2006, South Africa
12 SciLifeLab, Uppsala, Sweden

*** Authors for correspondence:**
gwenna.breton@gu.se, paul.verdu@mnhn.fr, mattias.jakobsson@ebc.uu.se, carina.schlebusch@ebc.uu.se

Keywords:
Africa; demographic history; admixture; migration; machine-learning; Approximate Bayesian Computations; Whole Genome Sequences; Rainforest Hunter-Gatherers; Agriculturalists; Khoe and San populations.

Abstract

The deep history of human evolution in Africa remains intensely debated with increasingly complex models being proposed. To investigate human evolutionary history in Africa, we sequenced and investigated 73 novel high-quality whole genomes from 14 Central and Southern African populations with diverse cultural practices. With extensive simulations and machine-learning Approximate Bayesian Computation inferences, we reconstruct jointly their complex history of divergences and migrations. We find vast genome-wide diversity within and among populations, often uncorrelated with geography and cultural practices. This shows the necessity to explicitly consider local genomic patterns, without merging samples indiscriminately into larger *a priori* categories to reconstruct human evolutionary histories. We find that tree-like population-histories with long periods of drift between short pulses of unidirectional gene-flow best explain the data, in contrast to continuous gene-flow. Without the need to invoke archaic admixture, the detailed patterns of genomic variation can be accurately fitted to a model where the lineage ancestral to Khoe-San populations diverged around 300,000 years ago from a lineage ancestral to rainforest hunter-gatherers and neighboring agriculturalists. Finally, throughout history, the short periods of gene-flow each coincide with epochs of major ecological and cultural changes in Sub-Saharan Africa.

Introduction

Unraveling genetic structure and gene flow among human lineages in Africa is crucial to the understanding of the biological evolution and diversity of *Homo sapiens* throughout the continent [1–3].

Population geneticists largely agree today that *Homo sapiens* spent a large part of its genetic evolution within Africa only, between its gradual emergence from anatomically archaic forms 600.000-200,000 years ago and the beginning of the Out-of-Africa expansions to the rest of the world around 100,000-50,000 years ago [2]. For the past 20 years, the detailed demographic and evolutionary history that shaped the genetic diversity of extant populations since the Out-of-Africa has been, and is still, the subject of numerous investigations in various regions of Africa [3–18].

However, how ancient genetic divergences, and the dynamics of admixture and/or migration events among ancient human lineages, influenced the genetic landscape observed today throughout Africa, largely remains to be assessed [19]. Classically tested demographic models [1,20–22], range from a unique ancestral *Homo sapiens* genetic population having recently and rapidly diverged into extant African populations via series of founding events and/or multifurcations (often referred to as tree-like models), to models where extant populations diverged in a remote past and remained isolated over long periods of time until relatively recently (often referred to as multiregional models). Moreover, each model may encompass possible gene exchanges among lineages via migration or admixture processes. Importantly, the timing, duration, and/or intensity of such gene-flow events may fundamentally change the most likely topology of bifurcating tree-like models compared to results obtained without gene flows, and may also create reticulations among pairs of lineages throughout history [22]. Finally, these classical albeit highly complex models have recently been enriched with possible introgression events from now extinct, unknown, or unsampled non-*Homo sapiens* or ancient "ghost" *Homo-sapiens* lineages [3,22–26]. Such latter models became plausible in Africa in analogy to the likely events of introgressions from now extinct hominid species into *Homo sapien*s lineages unveiled outside of Africa by the major advances of paleogenomics [27,28].

In this context, several recent investigations tested a variety of the above models with maximum-likelihood approaches relying on whole genome sequences from several extant populations sampled throughout the continent. They often reached highly contrasted results, with different ancient tree-topologies between different numbers of ancient lineages [2,22,24–26]. Furthermore, they found that reticulations among a limited number of ancient lineages explained the observed genetic patterns without the necessity of archaic introgressions [22]; or, instead, that admixture with other non-*Homo sapiens* or ancient *Homo sapiens* "ghost" populations most satisfactorily explained the results [23–26].

A possible source of such vast differences among obtained results is likely the variety of population sets considered in each study at a continental scale. Indeed, genetic diversity and differentiation among populations is maximal in Africa [3,29,30], and previously shown to be due to substantial differences in local populations' demographic histories (e.g. [4,6,9,11]). Therefore, representing vast regions of the continent by a single population [22,24], or even merging data from differentiated groups [26], likely influenced results obtained across studies and their interpretations.

In addition, and most importantly, all above models are highly nested, i.e. overlapping and genetically largely indistinguishable for vast spaces of model-parameter values. Indeed, depending on values of divergence times, effective population sizes, timing, duration and intensity of migration or admixture events among lineages, tree-like, reticulated, and even multiregional models, with or without ancient admixture, may be highly mimetic of one another. This essential methodological difficulty is well illustrated by the often relatively similar values of posterior likelihoods of the various competing models, thus indeed proved empirically hard to discriminate with the maximum-likelihood approaches deployed, even more so when competing models differ in the number and specifications of parameters

explored with different statistics. These fundamental statistical issues remain despite the use of high-quality whole genomes, an increasing number of populations considered at once, and increasingly powerful methods based on innovative statistics [22,24,26,31–33]. Finally, the lack of ancient DNA data older than a few thousand years in Africa, strongly hampers the empirical testing of the possible occurrence of ancient *Homo sapiens* "ghost" or non-*Homo sapiens* introgression events in *Homo sapiens* lineages in Africa [22–25,30].

In this context, we investigated 74 high coverage (>30x) whole genome sequences from an anthropologically well-characterized sample of hunter-gatherer, herder, and agricultural neighboring populations from Central and Southern Africa (**Figure1**, **Table1, TableS1**), merged together with 105 previously published high-quality whole genomes [27,34–37]. In particular, we investigated genomic diversity patterns and aimed at reconstructing ancient demographic histories among lineages having led to different groups of Khoe-San hunter-gatherer or herder populations from Southern Africa [6], and to different groups of both Eastern or Western Congo Basin hunter-gatherer populations (often historically designated "Pygmies" by Europeans) and their agriculturalist ("non-Pygmy") neighbors with whom they share, nowadays, complex socio-economic interactions [4,5,38]. Investigating detailed ancient demography and gene-flow among these three groups of populations, taking explicitly into account local differentiations and possible gene-flow among and within groups, is of particular interest as they have previously been identified to be among the most deeply diverged lineages in our species [2,3,29,39,40].

We conducted formal statistical choice of competing scenarios and subsequent joint estimation of parameters under the best scenario using machine-learning Approximate Bayesian Computation (ABC) [41–45], a methodological approach fundamentally differing from all maximum-likelihood methods previously deployed, and only rarely explored in previous studies reconstructing ancient demography in Africa [24]. ABC allows us to compare a range of complex demographic scenarios presumed to have led to observed genetic patterns in numerous population samples, and to estimate posterior parameter distributions best mimicking the data under the best scenario (e.g. [4,18,24]). Briefly, this is achieved based on informative summary-statistics computed on the observed data and on numerous explicit genetic simulations for which scenario-parameters are drawn randomly in large prior distributions set by the user. We thus aimed at inferring jointly competing tree-topologies, divergence times, effective population size changes, and timing, duration and intensity of possible asymmetric gene-flow events, to reconstruct the detailed evolutionary mechanisms underlying the genomic diversity of extant Central and Southern African populations.

Figure1: Sampling location and number of whole-genome sequenced individuals.
Individuals and populations originally sequenced in this work are indicated with filled circles. Individuals and populations from previously published data merged with the original data set for comparison purposes are indicated with open circles. Circles' diameters are proportional to the number of individuals as indicated in the top row of the legend in the bottom left corner.

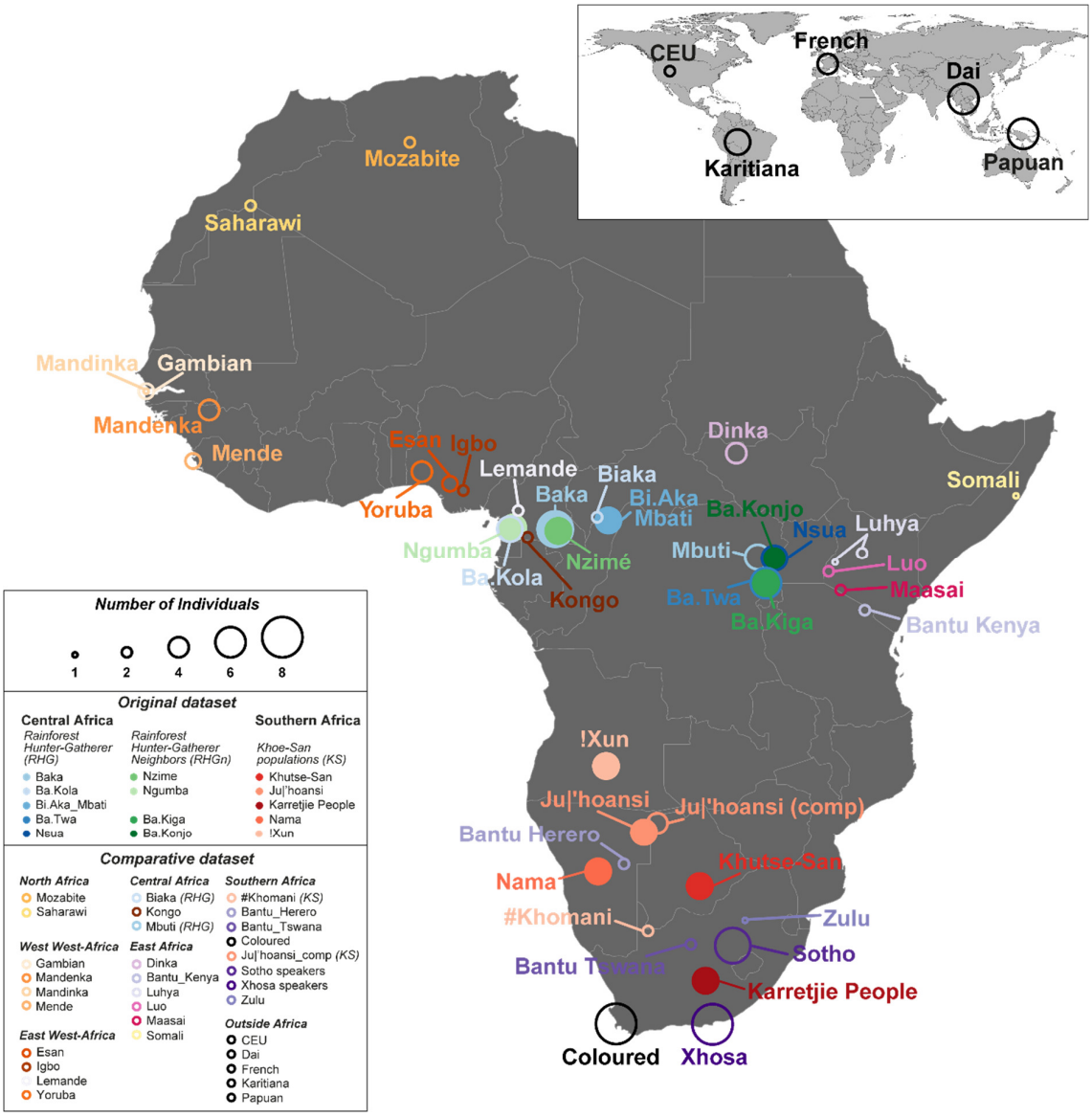


Table1: Population table

Geographical location of population samples are indicated in **Figure1**. Complementary information for genetic-geography and ANOVA analyses are provided in **TableS1**.

Population Name[1]	N[2]	Dataset[3]	Sampling Location	Language Family (language)	Close socio-economic interactions with	Samples included in the ROH-ASD-ADMIXTURE analyses (Figure2, Figure3, Figure4)	Samples included in the ABC analyses (Figure5, Figure6, Figure7, Figure8, Figure9)
Original dataset							
Baka	7	this study	Bosquet (Cameroon)	Niger-Congo non Bantu Adamawa Ubangian Gbanzili (Baka)	Nzime	Yes	wRHG, 5 individuals with highest coverage
Nzime	5	this study	Messea (Cameroon)	Bantu A842 (Nzime)	Baka	Yes	RHGn (West)
Ba.Kola	5	this study	Dispersed between Lolodorf and Kribi (Cameroon)	Bantu A80 (Kola)	Ngumba	Yes	wRHG
Ngumba	5	this study	Dispersed between Lolodorf and Kribi (Cameroon)	Bantu A80 (Ngumba)	Ba.Kola	Yes	RHGn (West)
Bi.Aka_Mbati	5	this study	Bombeketi section of Bagandou (Central African Republic)	Bantu C10 (Mbati)	Mbati farmers (not in the dataset)	Yes	wRHG
Ba.Kiga	5	this study	Mukono (Uganda)	Bantu J10 (Kiga)	Ba.Twa	Yes	RHGn (East)
Ba.Twa	6	this study	Kebiremu, Byumba, Kitariro, Mgungu, Nteko (Uganda)	Bantu J11 (Twa)	Ba.Kiga	Yes	eRHG, 5 individuals with highest coverage
Nsua	5	this study	Bundimassoli (Uganda)	Sudanic Mangbutu (Efe)	Ba.Konjo	Yes	eRHG
Ba.Konjo	5	this study	Mulimassenge (Uganda)	Bantu J40 (Konjo)	Nsua	Yes	RHGn (East)
!Xun	5	this study	Omega camp (Namibia) and Schmidtsdrift (Sout Africa)[4]	Khoisan (Ju)	*Not applicable*	Yes	nKS
Ju|'hoansi	5	this study	Tsumkwe (Namibia)	Khoisan (Ju)	*Not applicable*	Yes	nKS
Nama	5	this study	Windhoek (Namibia)	Khoisan (KhoeKhoe)	*Not applicable*	Yes	sKS
Karretjie People	5	this study	Colesberg (South Africa)	Khoisan (Tuu ancestral language) / Indo-European (Afrikaans current language)	*Not applicable*	Yes	sKS
Khutse-San	5	this study	Kutse Game reserve (Botswana)	Khoisan (Khoe Kalahari)	*Not applicable*	Yes	No
Comparative dataset							
Bantu_Herero	2	SGDP	Namibia	Bantu R30 (Herero)	*Not applicable*	Yes	No
Bantu_Kenya	2	SGDP	Kenya	*No data*	*Not applicable*	Yes	No
Bantu_Tswana	2	SGDP	South Africa	Bantu S31 (Tswana)	*Not applicable*	Yes	No
Biaka	2	SGDP	Central African Republic	Bantu C10 (Aka)	*Not applicable*	Yes	No
Coloured	8	SAHGP	South Africa	Indo-European (Afrikaans)	*Not applicable*	Yes	No
Dinka	4	SGDP, HGDP	Sudan	Nilo-Saharan (Dinka)	*Not applicable*	Yes	No
Esan	3	SGDP, KGP	Nigeria	Niger-Congo non Bantu (Esan)	*Not applicable*	Yes	No
Gambian	2	SGDP	Gambia	Niger-Congo non Bantu (Mandinka)	*Not applicable*	Yes	No
Igbo	2	SGDP	Nigeria	Niger-Congo non Bantu (Igbo)	*Not applicable*	Yes	No
Ju|'hoansi_comp	4	SGDP, HGDP	Namibia	Khoisan (Ju)	*Not applicable*	Yes	No
#Khomani	2	SGDP	South Africa	Khoisan (Tuu)	*Not applicable*	Yes	No
Kongo	1	SGDP	Cameroon	*No data*	*Not applicable*	Yes	No
Lemande	2	SGDP	Cameroon	Bantu A46 (Lemande)	*Not applicable*	Yes	No
Luhya	3	SGDP, KGP	Kenya	Bantu JE32 (Luhya)	*Not applicable*	Yes	No
Luo	2	SGDP	Kenya	Nilo-Saharan (Dholuo)	*Not applicable*	Yes	No
Maasai	2	SGDP	Kenya	Nilo-Saharan (Maasai)	*Not applicable*	Yes	No

Mandenka	4	SGDP, HGDP	Senegal	Niger-Congo non Bantu (West Maninkakan)	*Not applicable*	Yes	No
Mandinka	1	KGP	Gambia	Niger-Congo non Bantu (Mandinka)	*Not applicable*	Yes	No
Mbuti	5	SGDP, HGDP	Democratic Republic of Congo	Bantu D30 (Bambuti)	*Not applicable*	Yes	eRHG
Mende	3	SGDP	Sierra Leone	Niger-Congo non Bantu (Mende)	*Not applicable*	Yes	No
Mozabite	2	SGDP	Algeria	Afro-Asiatic (Mozabite)	*Not applicable*	Yes	No
Saharawi	2	SGDP	Western Sahara	Afro-Asiatic (Saharawi)	*Not applicable*	Yes	No
Somali	1	SGDP	Kenya	Afro-Asiatic (Somali)	*Not applicable*	Yes	No
Sotho	7	SAHGP	South Africa	Bantu S30 (Sotho)	*Not applicable*	Yes	No
Xhosa	8	SAHGP	South Africa	Bantu S41 (Xhosa)	*Not applicable*	Yes	No
Yoruba	4	SGDP, HGDP	Nigeria	Niger-Congo non Bantu (Yoruba)	*Not applicable*	Yes	No
Zulu	1	SAHGP	South Africa	Bantu S42 (Zulu)	*Not applicable*	Yes	No
French	4	SGDP, HGDP	France	Indo-European (French)	*Not applicable*	Yes	No
CEU	2	KGP	United States of America	Indo-European (English)	*Not applicable*	Yes	No
Dai	6	SGDP, KGP	China	Tai (Dai)	*Not applicable*	Yes	No
Papuan	6	SGDP, HGDP	Papua New Guinea	Papuan (no data)	*Not applicable*	Yes	No
Karitiana	5	SGDP, HGDP	Brazil	Tupian (Karitiana)	*Not applicable*	Yes	No

[1]Population Names are self-reported for the original dataset presented in this study
[2]Number of unrelated individuals considered in all analyses in this study (see **Material and Methods**)
[3]SGDP [34]; KGP [35]; HGDP [27,36]; SAHGP [37].
[4]The place of origin of the !Xun is around Menongue in Angola.

Results

We generated high-coverage whole genome sequences (>30X) in 74 individuals from 14 Central and Southern African populations for whom detailed ethno-anthropological information was gathered in the field jointly with DNA samples (**Figure1** and **Table1**). We merged this original dataset with a comparative dataset comprising 105 individuals from 27 African and five non-African populations for whom similar high-quality raw whole genome sequencing data were made available to the community [27,34–37]. After quality-control, variant-calling, and relatedness filtering procedures conducted on all 179 individuals together, we retained 73 and 104 unrelated individuals in our original and comparative datasets, respectively, for all subsequent analyses (see **Material and Methods**).

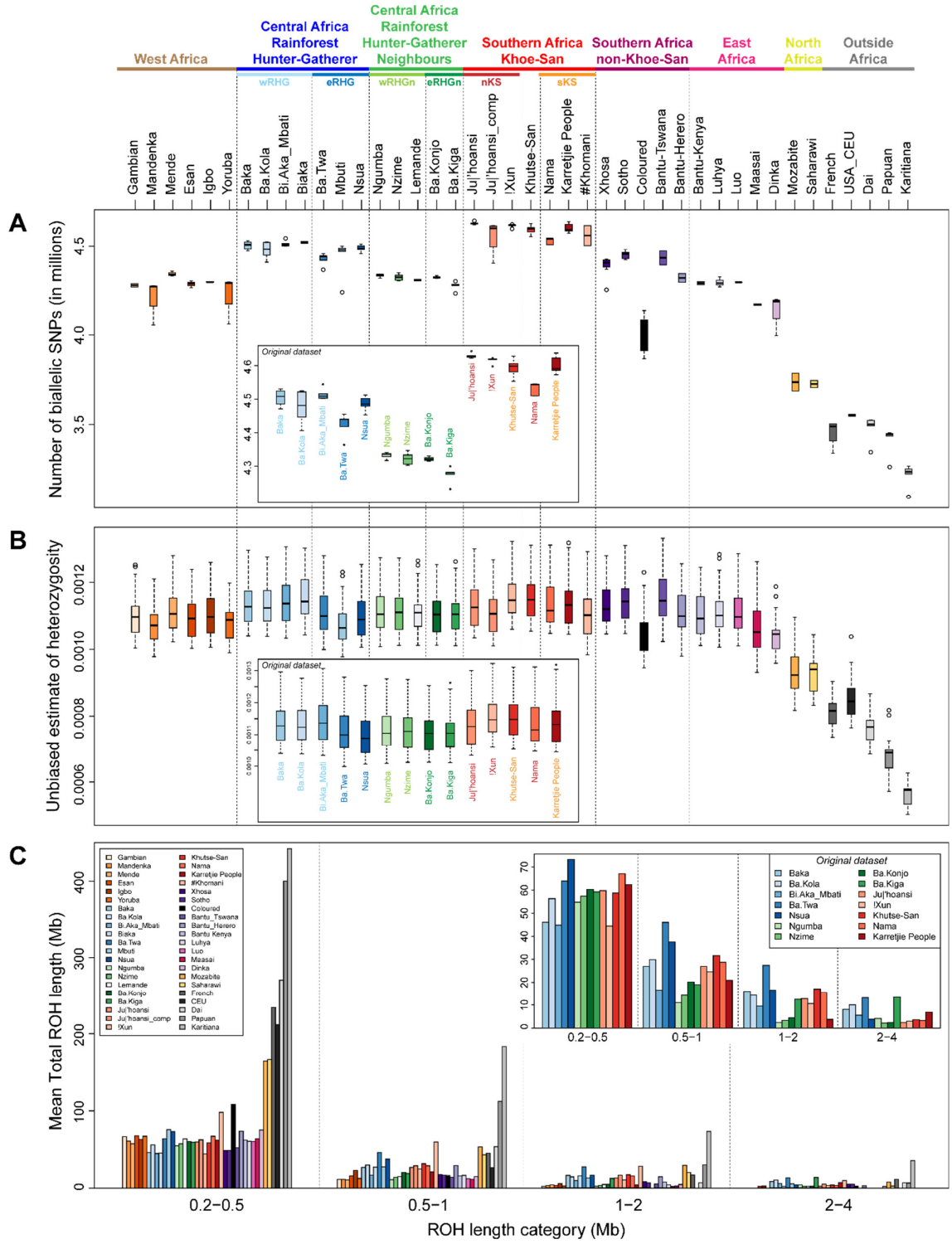
Considering only the 73 Central and Southern African unrelated individuals newly sequenced here, we identify a total of 26,780,319 biallelic SNPs (241,428 multiallelic SNPs), and 2,454,965 simple insertions/deletions (969,025 complex indels), compared to the reference sequence of the human genome GRCh38, out of which 854,114 (3.1893%) and 114,362 (4.6584%), respectively, were not previously reported in dbSNP 156 (**TableS2**). Among the 177 unrelated worldwide individuals including our original dataset, we identify 36,272,545 biallelic SNPs (257,906 multiallelic SNPs), and 3,159,306 simple insertions/deletions (977,542 complex indels), out of which 1,055,245 (2.9092%) and 189,124 (5.9863%), respectively, were not previously reported in dbSNP 156 (**TableS2**).

Furthermore, we find a large variability in the mean number of biallelic SNPs across populations (**Figure2A, TableS3**). In particular, we find substantially more biallelic SNPs within African populations and more variation of the mean number of SNPs across African populations (from mean=3,726,018; SD=27,793 across 2 individuals in the Saharawi from Western Sahara, to mean=4,628,957; SD=8,139 across 5 individuals in the Ju|'hoansi from Namibia), than within and among non-African populations (from mean=3,214,086; SD=69,241 across 5 individuals in the Karitiana from Brazil, to mean=3,551,792; SD=7,819 across 2 individuals from the USA CEU population). Overall, Southern African Khoe-San populations exhibit the highest mean number of biallelic SNPs as well as numbers of previously unreported SNPs, followed by Central African Rainforest Hunter-Gatherer populations from the Congo Basin, and then by all other African populations in our dataset. These results show that a substantial number of previously unknown variants can still be found when investigating high-quality whole genome sequences from relatively under-studied Sub-Saharan African populations, consistent with previous studies [22,26,27,34–37,40,46].

Figure2: Whole autosomal genome bi-allelic SNPs counts, unbiased heterozygosities, and distributions of Runs of Homozygosity by bins of length, in 37 African and 5 non-African populations.
(**A**) Numbers of bi-allelic SNPs across individuals within populations of more than one individual, compared to the reference sequence of the human genome GCRh38 and to previously reported variants in dbSNP 156. Detailed variant-counts are provided in **TableS2 and TableS3**. (**B**) Unbiased estimates of multi-locus heterozygosities [47], averaged across all variable (bi-allelic SNPs only) and non-variable autosomal sites with no-missing genotype within each population of more than one individual, corrected then for haploid population sample sizes. (**C**) Mean total ROH lengths in four bins of length categories for each population of more than one individual, separately. For each panel, we present a box zooming specifically on the results obtained for the original dataset of 73 unrelated individuals whole-genome sequenced from Central and Southern Africa.
See **Material and Methods** for filtering and calculation details and the software packages used. For (**A**) and (**B**), box-plots indicate the population median in between the first and third quartile of the box-limits, whiskers extending to data points no more than 1.5 times the interquartile range of the distribution, and empty circles for all more extreme points beyond this limit, if any. Population geographical location, categorization, and descriptions are provided in **Figure1**, **Table1**, and **TableS1**.

---

220 **Figure2: Whole autosomal genome bi-allelic SNPs counts, unbiased heterozygosities, and distributions of Runs of**
221 **Homozygosity by bins of length, in 37 African and 5 non-African populations.**
222



1. Large genetic diversities and differentiations in Central and Southern Africa

1.A. Heterozygosities and Runs of Homozygosity

We observe large differences across populations at a local scale in Central and Southern Africa in unbiased heterozygosities [47], calculated on both varying and non-varying autosomal sites (**Figure2A**, **Figure2B**, and **TableS3**. We find overall more variation in Rainforest Hunter-Gatherer populations from the western part of the Congo Basin than in RHG populations from the East (Wilcoxon one-sided test p-value = 0.00328 for all RHG populations in our dataset and 1.641×10^{-5} considering only those with original data). Note that the Baka and Ba.Kola western RHG populations exhibited significantly larger genetic diversities than their Nzime and Ngumba RHG neighbors (Wilcoxon one-sided test p-value = 0.04637).). Finally, observed genetic variation across Northern Khoe-San populations (!Xun, Ju|'hoansi, Ju|'hoansi_comp) and Southern KS populations (Nama, Karretjie People and #Khomani) was not significant (Wilcoxon two-sided test p-value = 0.7726).

In addition, we investigated the distributions of Runs of Homozygosity (ROH) across individuals within each of the 46 populations in our dataset. In general (**Figure2C**), individuals from Eastern RHG populations (Nsua, Mbuti, Ba.Twa), exhibit the longest total proportion of their autosomal genome in ROH across Sub-Saharan African populations, for all ROH length-classes, with the exception of #Khomani individuals from Southern Africa, and that of Coloured individuals for short ROH only. Interestingly, we find that Western RHG populations (Ba.Kola, Baka, Biaka, Bi.Aka_Mbati) have much less ROH of all classes than Eastern RHG. Furthermore, we find less short ROH, but more long ROH, in Western RHG than in all RHG neighboring populations across the Congo Basin except for the Ba.Kiga RHGn from Uganda. Finally, we find that all Northern and Southern Khoe-San populations have relatively similar total lengths of short ROH compared to that of RHG neighbors, with the notable exception of the !Xun, who have the smallest proportion of their genomes in short ROH, worldwide. Nevertheless, we find substantially more ROH of longer class in KS populations than in RHG neighbors, a relative pattern similar to what was observed for Western RHG. The distribution of ROH sizes has been shown to be highly informative about important demographic processes such as levels of inbreeding or endogamy within populations, and reproductive isolation or recent admixture among populations [18,48–50]. Thus, our ROH results highlight possibly vast differences in recent demographic processes and levels of endogamy across Central and Southern African populations.

Altogether, our heterozygosity and ROH results show the vast genetic diversity of Sub-Saharan African populations, as well as substantially diverging genomic patterns at a local geographical scale, in particular among and within groups of Rainforest Hunter-Gatherers and RHG neighbors.

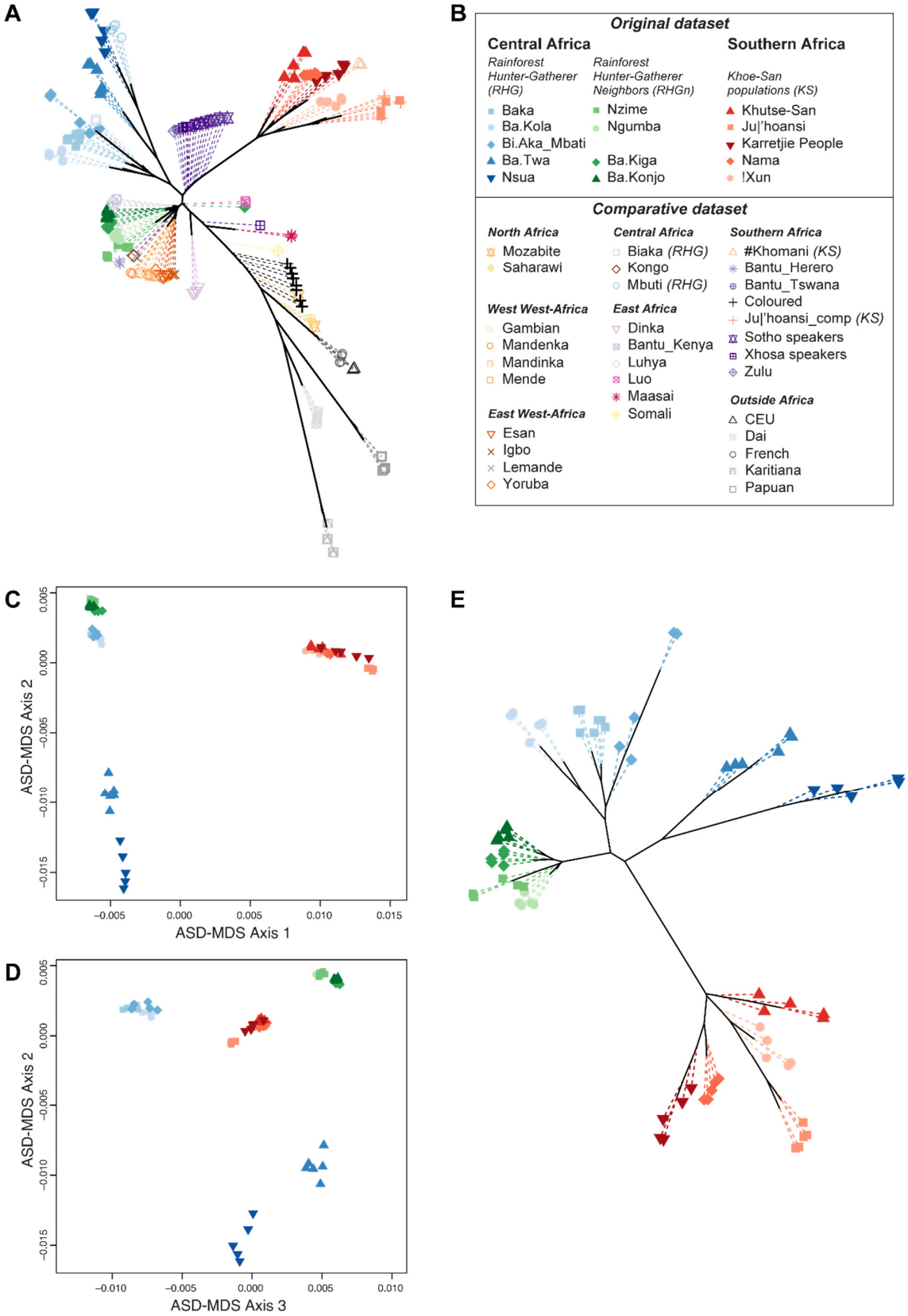
1.B. Individual pairwise genetic differentiation

We investigated individual pairwise genomic differentiation across the 177 worldwide unrelated individuals in our dataset, including the 73 unrelated novel Central and Southern African individuals, using Allele-Sharing Dissimilarities (ASD, [51]). The Neighbor Joining Tree (NJT, [52,53]) representation of the pairwise ASD matrix shows three main clusters among all African individuals, separated by a longer genomic distance from all non-African individuals (**Figure3A**). One cluster corresponds to all Southern African Khoe-San populations (represented by the pink and red symbols, see **Figure3B**), largely separated from a cluster corresponding to Rainforest Hunter-Gatherer populations from the Congo Basin (represented by the blue symbols); itself also largely separated from a third cluster grouping all other Western, Eastern, Southern and Central African populations in our data set, the latter group including Rainforest Hunter-Gatherer neighboring populations in green symbols (see, **Figure3B)**.

<u>Figure3</u>: Individual pairwise genetic differentiation patterns.
Neighbor-Joining Tree NJT [52,53] and Multi-Dimensional scaling representation of genome-wide Allele Sharing Dissimilarities [51] among pairs of individuals at worldwide and regional African scales. We considered the 14,182,615 genome-wide autosomal SNPs pruned for low LD (r^2 threshold 0.1) to calculate ASD between all pairs of individuals. (**A**) NJT computed for all 177 worldwide individuals in the ASD matrix. For easing the visualization of the internal branches of the NJT, all terminal edges are represented in dotted lines each measuring 1/10th of their true size. (**B**) Individual symbols and colors identifying the 73 Central and Southern African individuals originally whole-genome sequenced with filled symbols, and the 104 individuals from worldwide populations, including Africa, merged with the original data with open symbols. (**C**) and (**D**) First three axes of ASD-MDS computed on a subset of the full ASD matrix comprising only the individuals sequenced anew, whose symbols are provided in (**B**). (**E**) NJT computed on the same individual subset of the ASD matrix used for (**C**) and (**D**). Terminal edges of this latter tree are represented with dotted lines each measuring 1/30th of their true size.

---

284 **Figure3: Individual pairwise genetic differentiation patterns.**
285



Note, importantly, that individual pairwise dissimilarities at the genome-wide scale do not reflect geography (**Figure3**, **TableS4**), highlighting the large genetic differentiation found across African populations and the complexity of its distribution at both a local and continental scale. As can be seen in **Figure3**, from West to East of the Congo Basin throughout Central Africa, RHG populations cluster separately from RHG geographic neighbors with whom they nevertheless share close socio-economic interactions. Furthermore, Central African RHGn are more genetically resembling certain other Western, Eastern, and Southern Africans than their immediate RHG geographical neighbors. Analogously, Southern African Khoe-San populations are also clustering together and much more distant from other geographically neighboring Southern Africans.

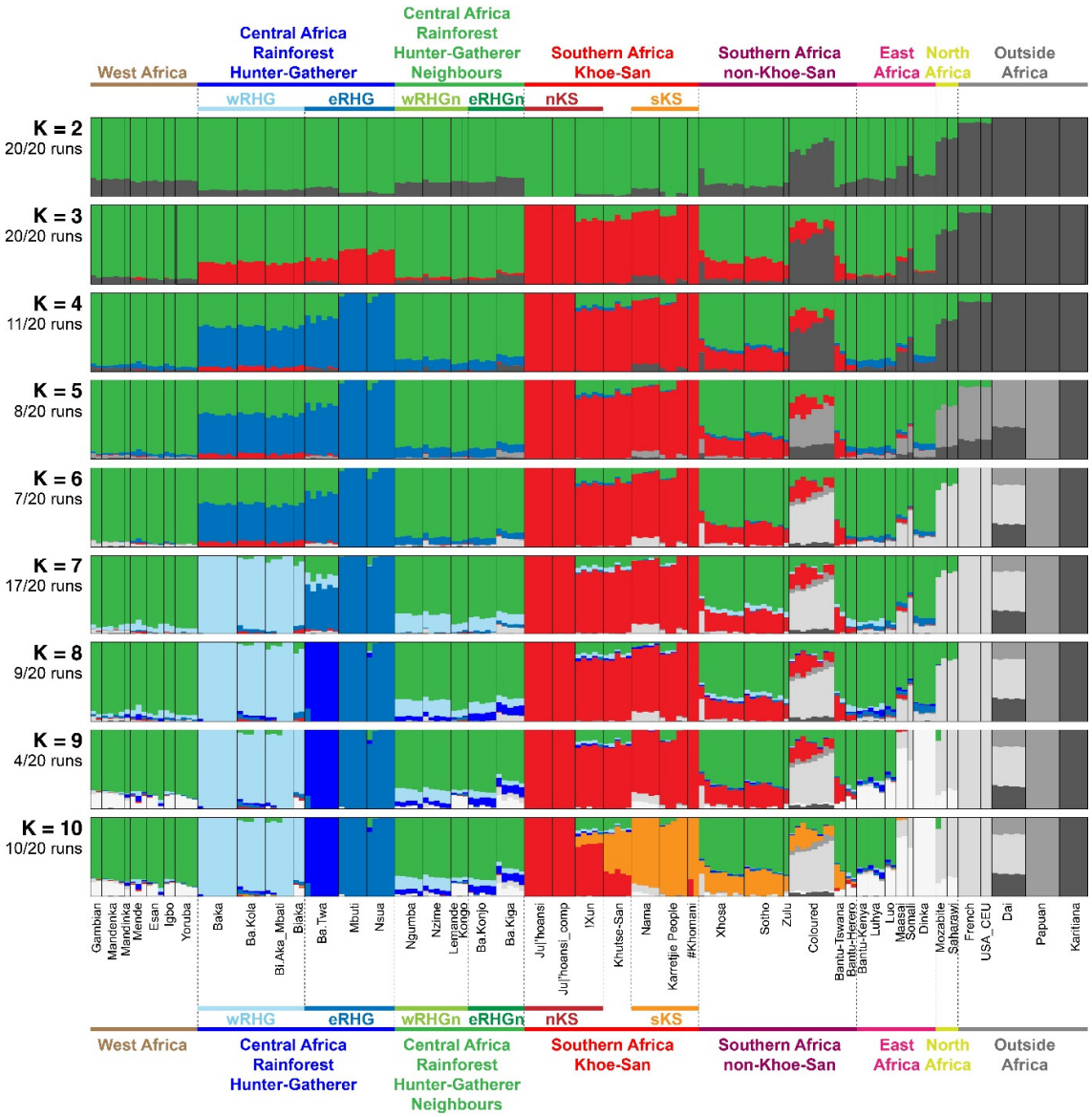
Furthermore (**TableS5**, **TableS6**, **TableS7**), ANOVA analyses show that linguistic families, sub-linguistic families, and traditional food-production strategies (**TableS1**), respectively explained a significant proportion of the variance in average population pairwise ASD throughout Africa (partial r^2 = 15.064%, 26.109%, and 9.021%, respectively, each with ANOVA permutation p <0.001), and throughout Sub-Saharan Africa only (partial r^2 = 14.811%, 22.759%, and 10.010%, respectively, each with ANOVA permutation p <0.001). However, most interestingly, we find more complex ANOVA results at a regional scale. Indeed, we find that linguistic families and sub-linguistic families do not significantly explain average population pairwise ASD patterns in Central Africa nor in the Congo Basin, while traditional food-production strategies do so in both regions (**TableS5**, **TableS6**, **TableS7**). Conversely, we find that traditional food-production strategies do not significantly explain population differentiation patterns among Southern African populations, while, here, linguistic and sub-linguistic families do. Finally, and further exemplifying the complexity of genomic differentiation patterns at a local scale in Africa, note that neither linguistic categories nor food-production strategies significantly explain average population pairwise ASD variation among Eastern African populations in our dataset.

The Multidimensional Scaling (MDS) projection of the subsampled ASD matrix for the 73 novel individuals from Central and Southern Africa, further shows substantial differentiation among groups of individuals within each one of the three clusters identified at the continental scale. Indeed, **Figure3C**, **Figure3D** and **Figure3E** show substantial differentiation between Western and Eastern Central African Rainforest Hunter-Gatherer populations, and even substantial differentiation between Ba.Twa and Nsua RHG very locally in Uganda. Conversely, we find relatively much smaller genetic differentiation among pairs of RHG neighbors from West to East of the Congo Basin. Finally, results also show substantial differentiation across Southern African Khoe-San populations, albeit these populations are relatively closer from one another than the different RHG populations.

NJT and MDS based on the ASD pairwise matrix provide an important view of the major axes of genetic differentiation and variation across samples, but do not easily allow to visually describe higher-order axes of genomic variations. To do so, we conducted an ADMIXTURE analysis [54], on the same entire worldwide dataset for increasing values of K (**Figure4**, **FigureS1**). This descriptive method is known to capture the same information as ASD-MDS and ASD-NJT, but allows to explore multiple axes of variation at the same time [18,54–59].

Figure4: Worldwide interindividual genetic diversity patterns with ADMIXTURE.
Each individual is represented by a single vertical line divided in *K* colors each proportional to an individual's genotype membership proportion assigned by ADMIXTURE [54] into the virtual cluster of that color. Two individuals showing the same relative proportions of each color at a given value of K are genetically more resembling one another than two individuals with different relative proportions of the same colors. Population and regional geographic groupings are not considered for the calculations, individuals are *a posteriori* grouped by populations and ordered by the user. Each population is separated by a thin vertical black line. Unsupervised clustering conducted with 840,031 genome-wide SNPs pruned for LD (r^2 threshold 0.1) and minor allele frequency (0.1) using ADMIXTURE for values of *K* ranging from 2 to 10, considering 177 worldwide individuals. For each value of *K* separately, we performed 20 independent runs and used PONG [60], to identify groups of resembling results (called "modes") based on Symmetric Similarity Coefficient measures of individual results. The number of runs among the 20 independent runs belonging to the same mode result (i.e. that have pairwise SSC above 0.998), is indicated below the number of K on the left of each barplot separately. Only those runs are averaged per individual to provide the barplot result for each value of K separately. For each value of K, all alternative modes identified among the 20 runs are provided in **FigureS1**.

Figure4: Worldwide interindividual genetic differentiation patterns with ADMIXTURE.



At K=2 in **Figure4**, the green virtual genetic cluster is maximized in Northern KS populations while the dark gray cluster is maximized in non-African populations, consistent with largest genetic distances in the ASD-NJT in **Figure3**, with all other individuals in our dataset presenting intermediate, unresolved, genotype membership proportions to either cluster.

At K=3, a novel red cluster is maximized in all Southern KS populations separating them from the rest of African individuals, also consistent with the ASD-NJT. At K=4, the novel blue cluster differentiates Central African RHG populations, and in particular certain Eastern RHG populations who maximize their membership proportions to this cluster.

The new gray and lighter gray clusters respectively at K=5 and K=6 differentiate Oceanian Papuan, South American Karitiana, and Western European French and North American USA-CEU individuals from one another and from all other non-African individuals. These two new clusters do not substantially affect genotype membership proportions in African individuals, except for the Southern African Coloured who show substantial membership to the light gray "European" cluster as well as reduced memberships to this cluster in Nama individuals and one Xhosa individual.

At K=7, the novel light blue cluster differentiates Eastern and Western RHG populations, as observed in **Figure3**, with Ba.Twa Eastern RHG showing largely unresolved clustering at this value of K with intermediate membership proportions to the light blue, the medium blue, and the green clusters.

At K=8, the novel dark blue cluster is now maximized in Ba.Twa Eastern RHG individuals who exhibited unresolved clustering at previous values of K. Note that this novel dark blue cluster is present only to very limited extents in some other RHG and RHGn populations.

At K=9, the novel white cluster is maximized in Dinka populations, located North-East of Central Africa and the Congo Basin, and substantially represented in other, but not all, East African populations.

Finally, at K=10, the novel orange cluster differentiates mainly among Northern and Southern Khoe-San populations, with the Central Khutse San showing substantial membership proportions to both the red and the orange clusters, respectively maximized in Northern and Southern KS. Note that, at K=10, the green cluster is maximized in all non RHG and non KS Sub-Saharan Africa populations except certain East African populations and the Southern African Coloured population, albeit note that this cluster is not fully resolved at this value of K since no individual exhibits 100% "green" genotype membership.

In summary, these ADMIXTURE results recapitulate and further dissect the pairwise genetic differentiation visually observed with ASD-MDS and ASD-NJT in **Figure3**. Most importantly, note that numerous Central and Southern African individuals retain, from K=2 to 10, intermediate genotype membership proportions in between certain clusters. This should be interpreted primarily as these individuals being at intermediate genetic distance between individuals presenting 100% membership to either cluster, which is consistent with ASD-MDS and ASD-NJT projections (**Figure3**). In turn, such intermediate distances may be either due to yet-unresolved ADMIXTURE clustering, which may appear at higher values of K, or can be due to admixture having occurred between ancestors respective to each cluster, the latter interpretation being likely but not formally tested by the ADMIXTURE method. The possible occurrence of gene-flow events across pairs of lineages and their influence on tree-topologies are incorporated explicitly in our Approximate Bayesian Computation inferences detailed below [41–45,61,62].

Altogether, our descriptive results highlight the vast genetic diversity and differentiation of African populations, maximized in certain groups of Central and Southern African populations, as previously reported (e.g. [17,22,26,27,34,36,37]). Finally, and most importantly, our results highlight in particular the vast genetic diversity and differentiation of populations at a local scale in Africa, including among immediate neighbors, not trivially correlated with geographical distances among populations (**TableS4**), linguistic family classification (**TableS5**), sub-linguistic family classification (**TableS6**),

nor traditional mode of subsistence (**TableS7**). In fact, these results anticipate that historical and demographic inferences in Africa may substantially differ when considering different population samples across the continent and even at a local scale, as well as advocate for extreme caution when grouping individuals and populations samples into larger categories.

2. Demographic and migration histories in Central and Southern African

2.A. Summary of the ABC inference design

We aimed at reconstructing the demographic and migration history that produced extant genomic patterns observed within and among Central and Southern African populations (**Figure2**, **Figure3**, **Figure4**), with machine-learning ABC scenario-choices followed by posterior-parameter estimations. To do so, we considered eight competing possible tree-topologies among five extant Northern and Southern Khoe-San, Western and Eastern Rainforest Hunter-Gatherer, and Rainforest Hunter-Gatherer neighboring populations (**Figure5**). Furthermore, for each topology, we considered two possible gene-flow processes among recent and ancient genetic lineages: instantaneous asymmetric gene-flow corresponding to instantaneous unidirectional introgression events between pairs of lineages; and recurring asymmetric gene-flow corresponding to unidirectional recurring migrations among pairs of lineages (scenarios "i" and "r" respectively, **Figure5**). Finally, we considered, for each topology and each gene-flow process, three nested gene-flow intensities for each event among pairs of lineages separately: no to very high gene-flow rates (scenarios "i1" and "r1"); no to moderate rates (scenarios "i2" and "r2"); and no to limited rates (scenarios "i3" and "r3").

This led to 8x2x3=48 competing scenarios in total which can be grouped in different ways to address formally two nested major questions:

i) which ancestral lineages to extant populations diverged first from the others and when did they do so, when considering complex gene-flow events among recent and ancient lineages?

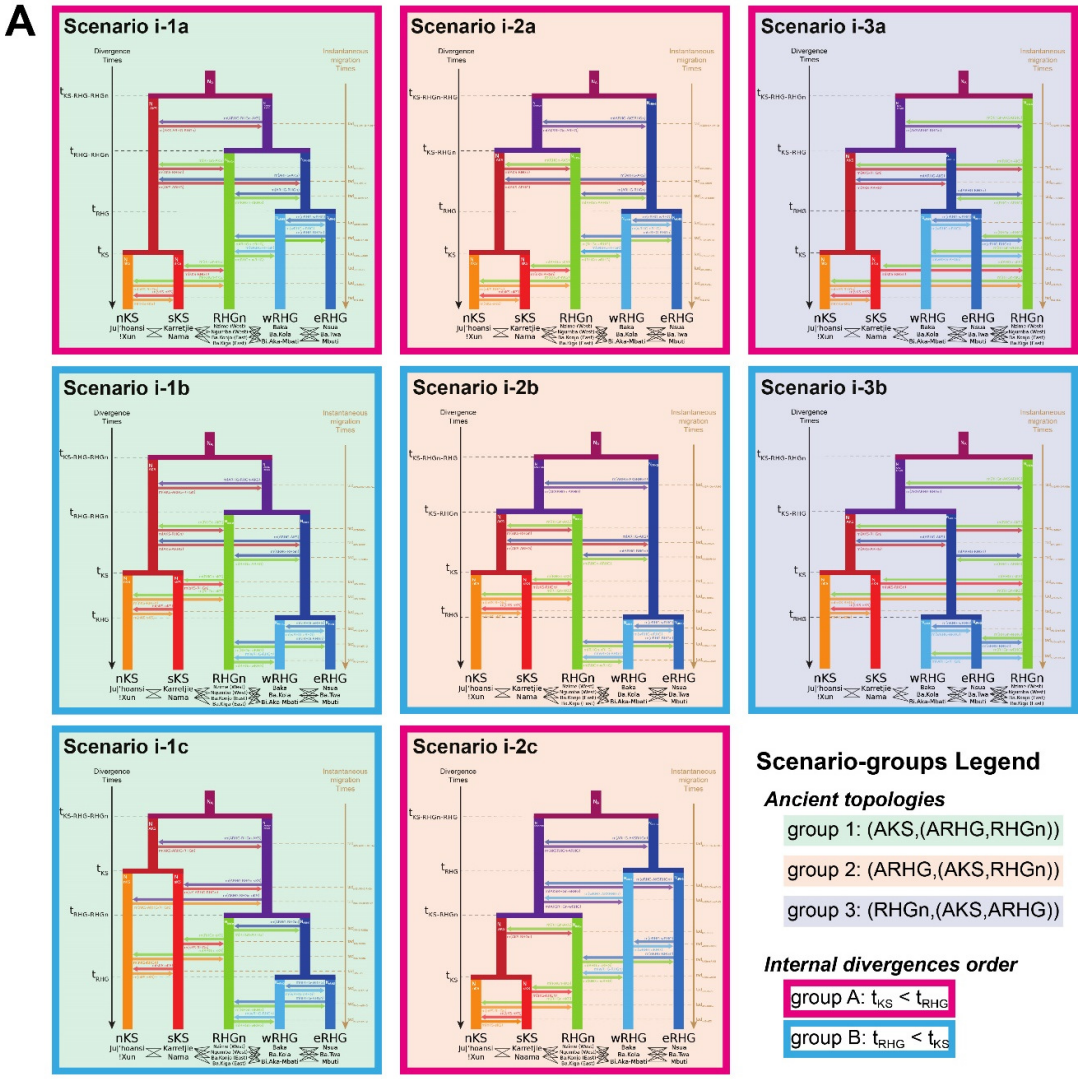
ii) did gene-flow events occur recurrently or more instantaneously among pairs of lineages during the evolutionary history of Central and Southern African populations? When did these events occur? How intense were they?

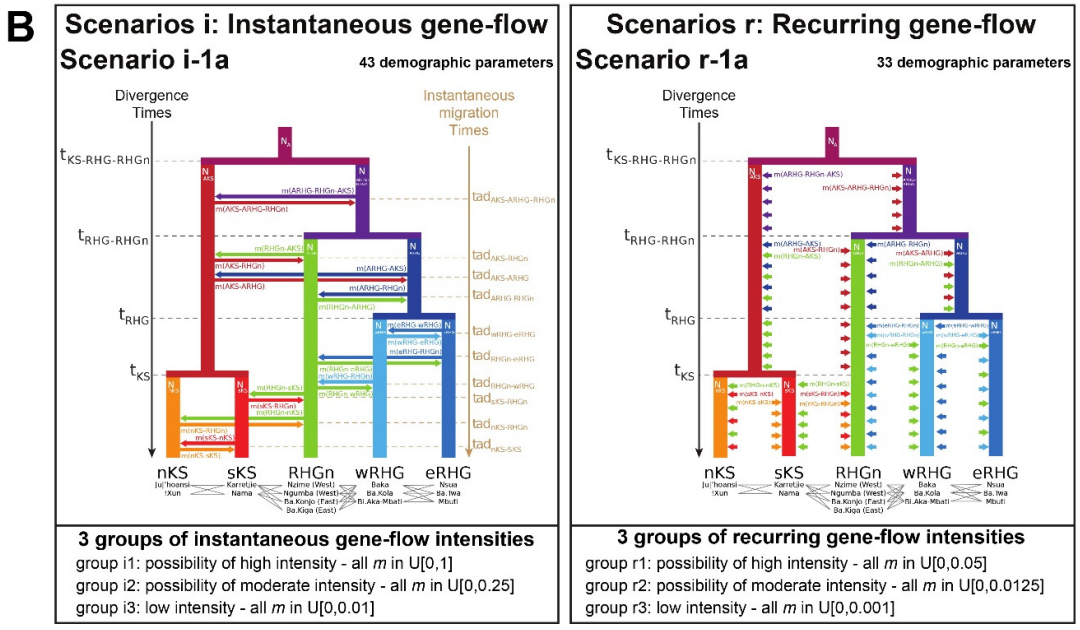
Importantly, we aimed at considering the vast genetic diversity and differentiation within and among populations and groups of populations at a local and regional scale (**Figure2**, **Figure3**, **Figure4**, and [3,29,30]). Therefore, we replicated the same ABC scenario-choice and posterior-parameter estimation procedures considering, in turn, 54 different combinations of five sampled populations of five individuals' whole autosomal genomes each (**Figure5**).

We thus conducted a total of 240,000 coalescent simulations under these 48 competing scenarios (5000 simulations per scenario), by drawing parameter values from prior distributions set by the user (**Figure5**, **TableS8**). For each simulation, we then calculated a vector of 337 summary-statistics (**TableS9**), within and among the five simulated populations; each vector of summary-statistics thus corresponds to a vector of parameter values drawn randomly from prior distributions and used for one simulation. We then employed Random Forest ABC procedures [45,63], to identify the best scenario or group of scenarios for each 54 combinations of five sampled "real" populations separately. Finally, under the best scenario, we produced 100,000 simulations, computed 202 summary-statistics for each simulation (**TableS9**), and performed Neural Network ABC posterior parameter joint-estimation [43,44], providing posterior-parameter distributions most likely to have produced observed genomic data, for each scenario-parameter.

<u>Figure5</u>: 48 competing scenarios for the history of Central and Southern African populations.
(**A**) Eight competing topologies for the demographic history of five Central and Southern African extant lineages. The Northern Khoe-San lineage (nKS) is represented by either the Ju|'hoansi or the !Xun population. The Southern Khoe-San lineage (sKS) is represented by either the Karretjie People or the Nama population. The Rainforest Hunter-Gatherer Neighbors lineage (RHGn) is represented either by the Western Congo Basin Nzime or Ngumba populations, or the Eastern Ba.Konjo or Ba.Kiga populations. The Western Rainforest Hunter-Gatherer lineage (wRHG) is represented by either the Baka, the Ba.Kola, or the Bi.Aka_Mbati population. The Eastern Rainforest Hunter-Gatherer lineage (eRHG) is represented by either the Nsua, the Ba.Twa, or the Mbuti population. As indicated in legend in the bottom right corner of the panel, the eight topologies can be grouped according to i) which ancestral lineage diverged first from the two others, or ii) whether the nKS and sKS lineages split earlier or later than the split of the wRHG and eRHG lineages. For each topology, we consider possible changes in lineages' constant effective population size, *Ne*, after each divergence event, *t*. For each topology, we considered possible gene-flow events among ancestral lineages and among recent lineages, represented as horizontal uni-directional arrows. (**B**) Example for a given topology (1a), of the fact that each one of the eight topologies are considered under two alternative competing "instantaneous" (Scenario i-1a) or "recurring" (Scenario r-1a) gene-flow processes. Instantaneous gene-flow processes consider that genes can be exchanged among pairs of ancestral or recent lineages at a single parameterized time, *tad*, with independent parameters of gene-flow intensity, *m*, from lineage "A" to "B" and from lineage "B" to "A". Recurring gene-flow processes consider instead that gene-flow occur at each generation between pairs of lineages with a constant independent rate, *m*, from lineage "A" to "B" and from lineage "B" to "A". Note that for all eight topologies and for both instantaneous and recurring gene-flow processes, we parameterized separately the gene-flow from lineage "A" into lineage "B", and the gene-flow from lineage "B" into "A", in order to allow for possibly asymmetrical gene-flow events among pairs of lineages. For both instantaneous and recurring gene-flow processes, we considered three nested intensities of gene-flow parameters indicated at the bottom of (**B**). Scenarios i1 or r1 consider that the gene-flow parameters *m* in each topology are independently drawn in Uniform distributions between 0 and 1 or 0 and 0.05, respectively, thus allowing for the possibility of no to high gene-flows from one lineage to the other. Scenarios i2 or r2 consider that the gene-flow parameters *m* in each topology are independently drawn in Uniform distributions between 0 and 0.125 or 0 and 0.0125, respectively, thus allowing only for the possibility of no to intermediate gene-flow from one lineage to the other. Finally, scenarios i3 or r3 consider that the gene-flow parameters *m* in each topology are independently drawn in Uniform distributions between 0 and 0.01 or 0 and 0.001, respectively, thus only allowing for the possibility of no to reduced gene-flow from one lineage to the other. Altogether we thus considered 8x2x3=48 competing scenarios for the demographic and migration history of Central and Southern African populations inferred with machine-learning Approximate Bayesian Computation. We conducted Random-Forest ABC scenario choice procedures among groups of these 48 scenarios in turn for 54 different sets of five observed populations each, represented as grey lines between population names, each population comprising five whole-genome sequenced individuals. Sampled-population names considered in the 54 combinations are given below each tree-leave, and combinations are limited to those pairing one RHGn with the specific eastern or western RHG population with whom they share complex socio-economic interactions, as indicated in **Table1**. See the detailed description of scenarios and their grouping, their parameters and their respective prior distributions used for ABC inference in **Material and Methods** and **TableS8**.

Figure5: 48 competing scenarios for the history of Central and Southern African populations.





483 **2.B. Are we able to mimic observed data with simulations under the 48 competing scenarios?**
484 We first checked that the data for each one of the 54 combinations of five sampled populations used for
485 all ABC inferences hereafter, which considered 100 high-quality 1 Mb-long genomic windows pruned
486 from the full genomes in our working dataset (see **Material and Methods**), reasonably echoed the
487 statistical descriptions obtained above which, instead, considered the full genomic SNP set. We thus
488 compared standardized individual pairwise ASD matrices for each one of the 54 sets of population-
489 combinations with the corresponding individual subsets of the standardized ASD matrix used for MDS
490 and NJT analyses in **Figure3**. We found that each 54 sets of population-combinations and associated
491 pruned SNP sets for ABC inferences highly resembled the genomic individual pairwise differentiation
492 patterns reported above using the entire SNP set. Indeed, we found significant (1000 Mantel
493 permutation two-sided $p < 0.002$), Pearson r correlations ranging from 0.945 to 0.974 across the 54 sets
494 of five population combinations (**TableS10**).
495 In order to conduct ABC inferences, we first checked whether we were able to simulate data, under
496 the 48 competing scenarios, for which vectors of summary statistics could mimic those obtained
497 separately from the 54 different combinations of five sampled populations each.
498 We first conducted goodness-of-fit permutation tests, separately for the 54 combinations of five
499 populations, and found that the observed vectors of 337 summary statistics were never significantly
500 different (54 goodness-of-fit permutation p-values > 0.56), from the 240,000 vectors of 337 summary
501 statistics computed from simulations under the 48 competing scenarios. Second, we computed a two-
502 dimensional PCA on the vectors of summary-statistics obtained from simulations on which we
503 projected, separately, the 54 vectors of these summary-statistics computed on observed data, and found
504 that observations each fell well within the space of simulations (**FigureS2**).
505 Both results show that we empirically were able to simulate data for which summary-statistics
506 were reasonably close to the observed ones, at least in parts of the parameter space used for simulations
507 from the 48 scenarios and for each one of the 54 separate combinations of five sampled populations
508 each. We could therefore conclude that the observed patterns of variation could be generated under
509 some of the 48 scenarios, and we proceed with Random Forest ABC scenario-choice and Neural
510 Network ABC posterior-parameter estimation procedures.
511 Note that the prior error in RF-ABC scenario-choice, using simulations as pseudo-observed data
512 and considering 5000 simulations under each one of the 48 scenarios in competition, converged towards
513 a minimum when using 1000 trees for all Random Forest ABC scenario-choice analyses conducted
514 hereafter (**FigureS3**).
515 Importantly, the *a priori* discriminatory powers of the RF, estimated as cross-validation procedures
516 based on all simulations used in-turn as pseudo-observed data, were overall good across all analyses
517 described hereafter, despite increased confusion among scenarios based on gene-flow intensity within
518 groups of instantaneous or recurring processes (**FigureS4, FigureS5, FigureS6, FigureS7**). This was
519 expected as classes of gene-flow intensities distinguishing among scenarios with, respectively, no-to-
520 low, no-to-moderate, or no-to-high gene-flow intensities, were designed to have nested parameter priors
521 (**Figure5**, **TableS8**). Note that this confusion among gene-flow intensities due to nestedness was more
522 acute for recurring gene-flow scenarios than for instantaneous gene-flow scenarios (**FigureS6**).
523 Finally, we evaluated, for each analysis below separately, how each one of the 337 summary-
524 statistics (**TableS9**), contributed *a priori* to the RF-ABC scenario-choice in the entire space of
525 simulations, by considering in-turn each simulation as pseudo-observed data and without considering
526 observed data. Most interestingly, we found that individual-pairwise ASD-MDS summary-statistics
527 describing how individuals are genetically differing from one another within and among populations
528 and population-pairwise F_{ST}, in particular, were always among the 50 most informative summary-
529 statistics for RF-ABC scenario-choice *a priori*. Alternatively, we found that all statistics summarizing
530 the Site Frequency Spectrum considered here were never among the 50 most informative statistics *a*

priori for any of the RF-ABC scenario-choice analyses (**FigureS8**, **FigureS9**, **FigureS10**, **FigureS11**, **FigureS12**, **FigureS13**, **FigureS14**, **FigureS15**, **FigureS16**, **FigureS17**).

2.C. Which historic scenario best explains genomic patterns in extant African populations?

2.C.1. Instantaneous or recurring gene-flows of which intensity among African lineages?

Figure6A shows that gene-flow processes among recent or ancient genetic lineages most likely occurred during very limited "instantaneous" periods of time, rather than recurrently, throughout the entire evolutionary history of Central and Southern African populations. Indeed, for each one of the 54 different combinations of five sampled populations, the group of 24 scenarios considering instantaneous gene-flow processes provides vectors of summary statistics systematically closest to the observed ones, whichever the tree-topology or intensity of the gene-flow considered. Note that we find that posterior probabilities of choosing the correct group of scenarios for each 54 sets of population combinations range from 0.97828 to 0.99510, compared to a prior probability of 1/2 in this particular test.

Furthermore, we find that we imperatively need these gene-flow processes to be potentially high in order to best explain the observed data for each one of the 54 combinations of population samples, whichever the tree-topology and gene-flow process (**Figure6B**); whichever the tree-topology but considering gene-flow processes separately (**Figure6C**); or even considering 24 competing scenarios under instantaneous gene-flow processes only, whichever the tree-topology (**FigureS5**). Note that, for these three particular tests, we find that posterior probabilities of choosing the correct group of scenarios for each 54 sets of population combinations range, respectively, from 0.66318 to 0.87138 (**Figure6B**), from 0.67765 to 0.83957 (**Figure6C**), and from 0.63935 to 0.87000 (**FigureS5**), compared to a prior probability of 1/3, 1/6, and 1/3 in these three particular tests, respectively.

Finally, note that when considering the 48 competing-scenarios separately, a very challenging task *a priori* given the number of competing scenarios and the high level of nestedness among scenarios, the best scenarios predicted by the RF-ABC procedure is, among the 54 tests conducted, systematically found for scenarios considering an instantaneous gene-flow process allowing for possibly highly intense gene-flow rates (**FigureS6, FigureS7**).

559 ----------

560 **<u>Figure6</u>: Random-Forest ABC scenario-choices among 48 competing scenarios.**

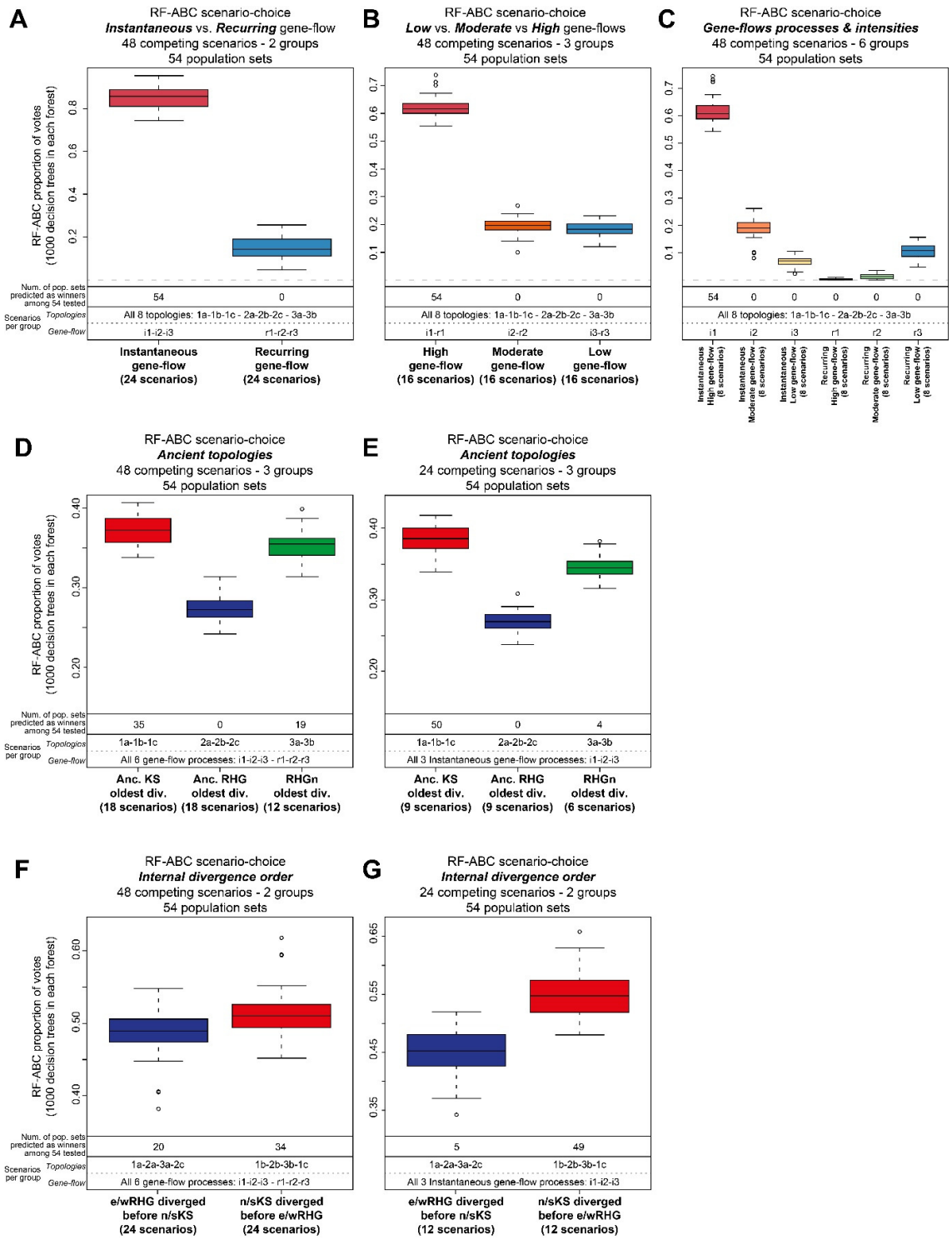
561 Random Forest Approximate Bayesian Computation scenario-choice results [45,63], are conducted, for each analysis in each
562 panel, separately for 54 different combinations of five Central and Southern African sampled populations each. Posterior
563 proportions of votes obtained for each 54 sampled-populations combinations, and for RF-ABC analysis in each panel
564 respectively, are provided as box-plots indicating the median in between the first and third quartile of the box-limits, whiskers
565 extending to data points no more than 1.5 times the interquartile range of the distribution, and empty circles for all more
566 extreme points beyond this limit, if any. In each RF-ABC analysis in each panel, proportion of votes (indicated in the y-axis)
567 are obtained using 1,000 decision trees in the random forest for each group of competing scenarios (indicated in the x-axis
568 table and labels), using 5,000 simulations per scenario and 337 summary-statistics (see **Material and Methods** and **TableS9**
569 for details). Detailed competing-scenarios grouped in each group are indicated in the two bottom lines of the x-axis table, with
570 topology and gene-flow processes codes provided in **Figure5** and explained in detail in **Material and Methods**. The top-line
571 of the x-axis table indicates, for each group of competing scenarios in each analysis, the number of times the corresponding
572 group of scenarios was predicted as the best one among the 54 different combinations of five sampled-populations. (**A**)
573 Scenario-choice results for the two groups of instantaneous or recurring gene-flow processes (24 scenarios and 120,000
574 simulations in each two competing groups), all gene-flow intensities and all topologies "being equal". Posterior probabilities
575 of choosing the correct group of scenarios for each 54 sets of population combinations ranged from 0.97828 to 0.99510,
576 compared to a prior probability of 1/2 in this particular test. (**B**) Scenario-choice results for the three groups of intensities for
577 gene-flow processes (16 scenarios and 80,000 simulations in each three competing groups), all topologies and both
578 instantaneous and recurring gene-flow processes "being equal". Posterior probabilities of choosing the correct group of
579 scenarios for each 54 sets of population combinations ranged from 0.66318 to 0.87138, compared to a prior probability of 1/3
580 in this particular test. (**C**) Scenario-choice results for the six groups of gene-flow processes and intensities separately (8
581 scenarios and 40,000 simulations in each six competing groups), all topologies "being equal". Posterior probabilities of
582 choosing the correct group of scenarios for each 54 sets of population combinations ranged from 0.67765 to 0.83957, compared
583 to a prior probability of 1/6 in this particular test. (**D**) Scenario-choice results for the three groups of scenario topologies
584 differing in the order of ancient lineages divergence (differing number of scenarios in each three competing groups are evened
585 by randomly sampling the same number of simulations, equal to 60,000, in each group, see **Material and Methods**), all gene-
586 flow processes and intensities "being equal". Posterior probabilities of choosing the correct group of scenarios for each 54 sets
587 of population combinations ranged from 0.39862 to 0.48967, compared to a prior probability of 1/3 in this particular test. (**E**)
588 Scenario-choice results for the same test as (**D**) restricted to the 24 competing scenarios considering instantaneous gene-flow
589 processes only (differing number of scenarios in each three competing groups are evened by randomly sampling the same
590 number of simulations, equal to 30,000, in each group, see **Material and Methods**), all gene-flow intensities "being equal".
591 Posterior probabilities of choosing the correct group of scenarios for each 54 sets of population combinations ranged from
592 0.43333 to 0.51857, compared to a prior probability of 1/3 in this particular test. (**F**) Scenario-choice results for the two groups
593 of scenario topologies differing in the relative order of divergences between Northern and Southern KS, and Western and
594 Eastern RHG lineages, respectively (24 competing scenarios and 120,000 simulations in each two groups), all gene-flow
595 processes and intensities "being equal". Posterior probabilities of choosing the correct group of scenarios for each 54 sets of
596 population combinations ranged from 0.62115 to 0.72015, compared to a prior probability of 1/2 in this particular test. (**G**)
597 Scenario-choice results for the same test as (**F**) restricted to the 24 competing scenarios considering instantaneous gene-flow
598 processes only (12 competing scenarios and 60,000 simulations in each two groups), all gene-flow intensities "being equal".
599 Posterior probabilities of choosing the correct group of scenarios for each 54 sets of population combinations ranged from
600 0.62033 to 0.71883, compared to a prior probability of 1/2 in this particular test.

601 ----------

602

603

Figure6: Random-Forest ABC scenario-choices among 48 competing scenarios.



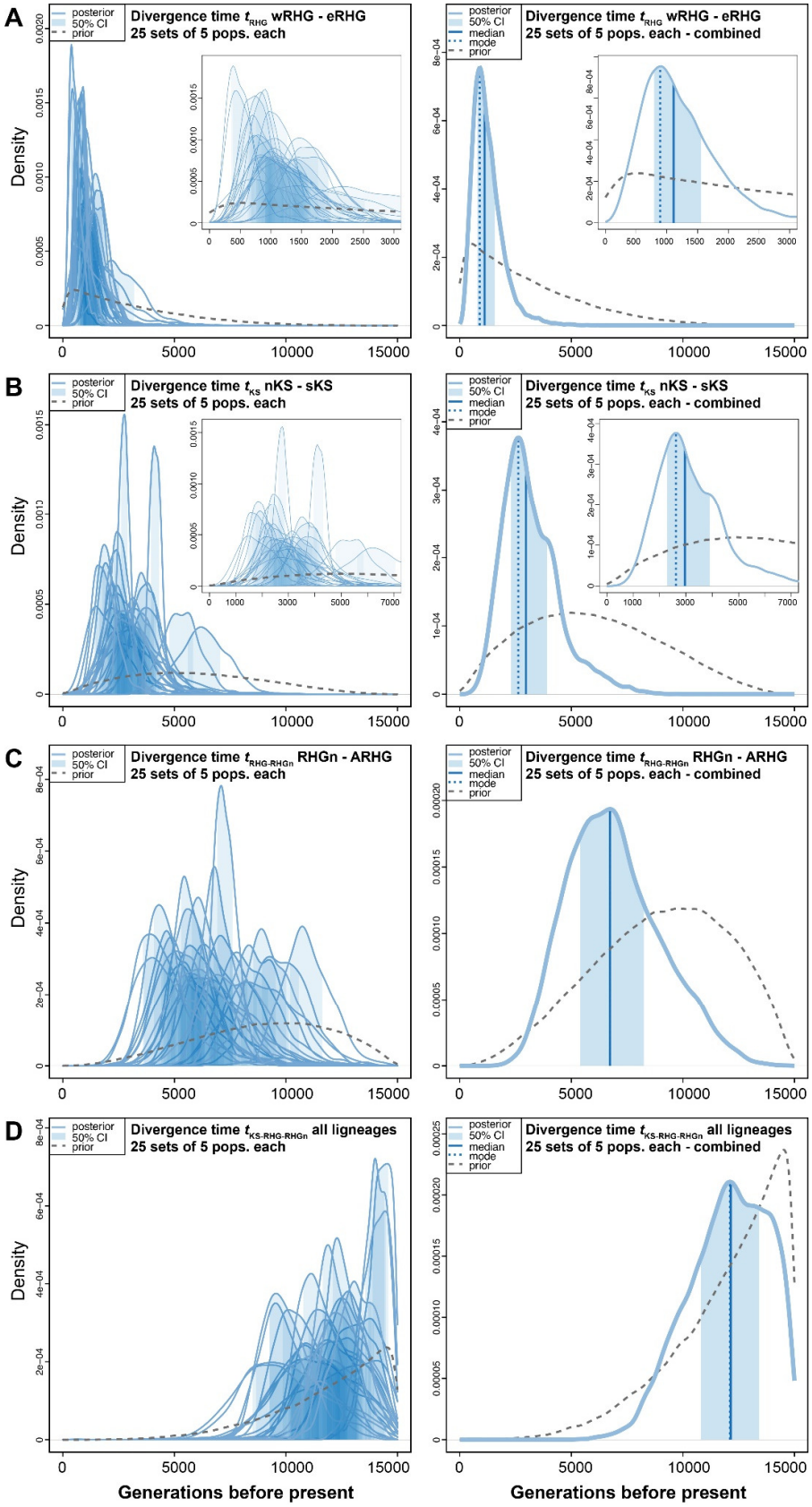
607 *2.C.2. Ancient tree-topology among African lineages?*
608 **Figure6D** shows that, overall, the RF-ABC scenario-choice procedures vote in favor of an ancient tree-
609 topology where the ancestral Khoe-San lineage diverged first followed by the divergence between the
610 ancient Rainforest Hunter-Gatherer lineage and that of their neighbors (Scenarios tree-topologies 1a-
611 1b-1c, **Figure5**), in a majority (35/54) of the 54 combinations of observed populations tested here,
612 whichever the gene-flow process and intensity. Note that we find that posterior probabilities of choosing
613 the correct group of scenarios for each 54 sets of population combinations range from 0.39862 to
614 0.48967, compared to a prior probability of 1/3 in this particular test.
615 Furthermore, we find that this majority of votes is much larger (50/54), when considering
616 instantaneous gene-flow processes only, whichever their intensities (**Figure6E**). Note that we find that
617 posterior probabilities of choosing the correct group of scenarios for each 54 sets of population
618 combinations range from 0.43333 to 0.51857, compared to a prior probability of 1/3 in this particular
619 test.
620 Finally, note also that the tree-topologies where the RHGn lineage diverge first from the lineage
621 ancestral to that of the RHG and the KS (Scenarios tree-topologies 3a-3b, **Figure5**), are predicted to be
622 the best ones in a minority of the tests (19 out of 54 combinations of sampled populations, **Figure6D**),
623 whichever the gene-flow process and/or intensity. Furthermore, they do so in an even smaller minority
624 (4/54) for instantaneous gene-flow processes only, whichever the intensities (**Figure6E**). Interestingly,
625 we found that tree-topologies where the ancient lineage for RHG populations diverges first from the
626 two others (Scenarios tree-topologies 2a-2b-2c, **Figure5**), is never favored in our analyses, whichever
627 the gene-flow process and/or intensity.
628
629 *2.C.3. Recent tree-topology among African lineages?*
630 Finally, **Figure6** shows that Northern and Southern Khoe-San extant lineages likely diverged from
631 one-another before the divergence between Eastern and Western Rainforest Hunter-Gatherer lineages
632 (Scenarios tree-topologies 1b-1c-2b-3b, **Figure5**), whichever the ancient tree-topology and gene-flow
633 process and intensity (**Figure6F**), and whichever the ancient tree-topology for instantaneous gene-flow
634 processes only, all classes of gene-flow intensities "being equal" (**Figure6G**). Interestingly, the "b"
635 group of tree-topologies is found to best-describe observed summary-statistics for 34 out of 54
636 combinations of sampled populations and this majority of RF-ABC votes is even larger (49/54) when
637 considering only the group of 24 competing scenarios for which gene-flows are instantaneous rather
638 than recurring. Note that, for these two particular tests, we find that posterior probabilities of choosing
639 the correct group of scenarios for each 54 sets of population combinations range, respectively, from
640 0.62115 to 0.72015 (**Figure6F**), and from 0.62033 to 0.71883 (**Figure6G**), compared to a prior
641 probability of 1/2 in both these two particular tests.
642
643 *2.C.4. Conclusion: intersecting scenario-choice results for the history of Africa.*
644 Altogether, when intersecting RF-ABC scenario-choice inferences in groups of gene-flow processes
645 and tree-topologies (**Figure6**), with inferences considering all competing-scenarios separately
646 (**FigureS6, FigureS7**), and among 54 combinations of five sampled populations, we find that Scenario
647 i1-1b (**Figure5**), systematically best-describes the observed data in all conformations of scenario-choice
648 tests for 25 combinations of populations (**TableS11**), while the second scenario best explaining the data
649 (Scenario i1-3b) does so consistently across all tests for only four combinations of population out of 54.
650 Therefore, our results point to a scenario for the history of Central and Southern African
651 populations where the ancestral KS lineage diverged first, followed by the divergence between the
652 ancestral RHG and the RHGn lineages, with a divergence between the extant Northern and Southern
653 KS lineages occurring independently before that of the extant Eastern and Western RHG lineages. Most
654 importantly, the best scenario necessarily involves gene-flow events among all pairs of lineages

655 throughout history which occurred during relatively short "instantaneous" periods of time, rather than
656 recurrently over longer periods of time. Furthermore, these instantaneous gene-flow events must each
657 have been possibly of high intensity, rather than all limited to moderate or low intensities, and possibly
658 asymmetrical between pairs of lineages, to best explain the observed data. Finally, our results also
659 highlight that considering different combinations of extant populations separately in the analyses may
660 provide contrasted results where alternative evolutionary scenarios may best explain observations.
661 For conservativeness and to further consider the genetic diversity of extant Central and Southern
662 African populations at a local or regional scale, we will henceforth conduct all Neural-Network ABC
663 posterior-parameter estimations separately for the 25 combinations of sampled populations
664 systematically providing the Scenario i1-1b as the most-likely one among the 48 competing scenarios.
665
666 **2.D. Which scenario-parameters best explain genomic patterns in extant African populations?**
667 *2.D.1. Divergence times.*
668 Divergence times are most often well estimated by our Neural Network ABC posterior parameter
669 inferences (**Figure7** and **TableS12**), with relatively narrow 90% Credibility Intervals (CI), posterior
670 distributions very often largely departing from the priors, and low cross-validation posterior parameter
671 estimation errors in the vicinity of the observed data (**TableS13**). Interestingly, we find relatively
672 consistent posterior estimates across the 25 different combinations of five populations for, separately,
673 the Eastern and Western RHG divergence between 383 generations before present (gbp) (mode point
674 estimate, 90%CI=[220-934]) and 2174 gbp (90%CI=[1,226-4,373], **TableS12**), and for the Northern
675 and Southern KS divergence between 1,486 gbp (mode point estimate, 90%CI=[844-3,798]) and 6,201
676 gbp (90%CI=[4,465-8,126]). Combining the posterior distributions for these two parameters
677 (**Figure7A, Figure7B**, **TableS14**), respectively, provides, synthetically, a modal point-estimate
678 divergence time between eRHG and wRHG of 892 gbp (90%CI =[422-2,531]), and one between nKS
679 and sKS, largely more ancient, having occurred some 2,623 gbp (90%CI=[1,481-5,763]).
680 In a more remote past, the divergence time between the lineage ancestral to all RHG populations
681 and the RHG neighbors' lineage is also most often well estimated, based on CI-width and departure
682 from prior distributions, and cross-validation posterior parameter estimation errors in the vicinity of the
683 observed data, for each 25 combinations of populations respectively, albeit results are more variable
684 across sets of population combinations than for more recent divergence times (**Figure7C**, **TableS12,**
685 **TableS13**). Indeed, ancient RHG and RHGn lineages diverged between 3,904 gbp (mode point
686 estimate, 90%CI=[2,828-6,636]) and 10,727 gbp (90%CI =[9,268-12,736], **TableS12**) across
687 population combinations, and synthetically combining results together 6,726 gbp (90%CI=[3,876-
688 10,748], **Figure7C** and **TableS14**).
689 Finally, we also find relatively variable posterior estimates for the most ancient divergence time in
690 our tree-topology, between the ancestral lineages to all extant KS lineages and the ancestral lineages to
691 all RHG and the RHG neighbors; with, again, posterior-parameter distributions across 25 combinations
692 of sampled populations very often satisfactorily estimated, with relatively reduced 90%CI and
693 substantial departure from the priors, and low cross-validation posterior parameter estimation errors in
694 the vicinity of the observed data (**Figure7D**, **TableS12, TableS13**). Overall, we find that the original
695 most ancient divergences in our tree-topology occurred some 12,117 gbp (90%CI=[8,875-14,538]),
696 when all results are combined together synthetically (**Figure7D** and **TableS14**).
697 Altogether, the relatively large variation in divergence times posterior estimates further back in
698 time across combinations of sampled populations empirically highlights that demographic and historical
699 reconstructions largely depend on the specific samples considered when investigating highly
700 differentiated Sub-Saharan populations. Our results thus explicitly advocate for caution when
701 summarizing results obtained across population sets and further likely explain the, sometimes, apparent

discrepancies that arise across studies considering different sample sets and/or artificially merging several samples from relatively differentiated populations (e.g. [22,26]).

Figure7: Neural Network ABC posterior parameter distributions of divergence times.
Neural Network Approximate Bayesian Computation posterior parameter joint estimations [43,44], of topological divergence-times (in generations before present) for 25 sets of five Central and Southern African populations for which the best scenario identified by RF-ABC was Scenario i1-1b (**Figure5**). NN ABC posterior parameter estimation procedures were conducted using 100,000 simulations under Scenario i1-1b, each simulation corresponding to a single vector of parameter values drawn randomly from prior distributions provided in **TableS8**. We considered 42 neurons in the hidden layer of the NN and a tolerance level of 0.01, corresponding to the 1,000 simulations providing summary-statistics closest to the observed one. NN posterior estimates are based on the logit transformation of parameter values using an Epanechnikov kernel between the corresponding parameter's prior bounds (see **Material and Methods** and **TableS8**). Posterior parameter densities are represented with solid blue lines. 50% Credibility Intervals are represented as the light blue area under the density. The median and mode values are represented as a solid and dotted blue vertical line, respectively. Parameter prior distributions are represented as dotted gray lines. For all panels, the left plots represent the NN-ABC posterior parameter distributions for each 25 sets of five Central and Southern African populations best explained under Scenario i1-1b, separately (**TableS11** and **TableS12**). For all panels, the right plots represent a single parameter posterior distribution obtained from combining the 25 posterior distributions together. (**A**) Results for parameter t_{RHG} corresponding to the split time between the Western and Eastern Rainforest Hunter-Gatherer (RHG) lineages (**Figure5**). (**B**) Results for parameter t_{KS} corresponding to the split time between Northern Khoe-San (nKS) and Southern Khoe-San (sKS) lineages (**Figure5**). (**C**) Results for parameter $t_{RHG\text{-}RHGn}$ corresponding to the split time between the Rainforest Hunter-Gatherer neighboring population lineage (RHGn) and the lineage ancestral to Western and Eastern RHG (**Figure5**). (**D**) Results for parameter $t_{KS\text{-}RHG\text{-}RHGn}$ corresponding to the split time between the lineage ancestral to KS populations and the lineage ancestral to all RHG and RHGn lineages; this event thus corresponding to the split time of all lineages in the history of Central and Southern African populations (**Figure5**). Results for all left panels are summarized in **TableS14** and cross-validation posterior parameter estimation errors in the vicinity of the observed data in **TableS13**.

Figure7: Neural Network ABC posterior parameter distributions of divergence times.



2.D.2. Effective population sizes.

Effective population sizes (Ne) for all recent Northern and Southern KS, Western and Eastern RHG, and RHG neighbors' lineages are often reasonably well estimated with relatively reduced 90%CI, substantial departure from their priors, and low cross-validation posterior parameter estimation errors in the vicinity of the observed data (**TableS13**), for each 25 sets of sampled population combinations (**FigureS18**, **TableS12**). Combining posterior distributions among the 25 separate tests, we find Ne posterior estimates ranging from modal point-estimates of 9,461 diploid effective individuals (90%CI=[5,074-67,317]) in the wRHG to 18,022 (90%CI=[9,150-77,157]) in the RHGn extant lineages (**FigureS18**, **TableS14**). Importantly, we were unable to satisfactorily estimate, for almost all 25 combinations of sampled populations, the three effective population sizes for, respectively, the ancient KS lineage, the ancient RHG lineage, and the lineage ancestral to all RHG and RHG neighbors' extant lineages (**TableS14**, **FigureS18**, **TableS12, TableS13**).

Conversely, our posterior estimates of the effective size of the lineage most ancestral to all our populations was satisfactorily estimated, with relatively narrow 90%CI, substantial departure from the prior distributions, and among the lowest cross-validation posterior parameter estimation errors in the vicinity of the observed data (**TableS13**), for almost all 25 sets of population combinations. Combining posterior distributions, despite a noteworthy variation across the 25 tests (**FigureS18**, **TableS12**), we find an ancestral effective size for the lineage ancestral to all our Central and Southern African extant populations of 18,507 diploid effective individuals (90%CI=[9,399-60,616]), the largest posterior estimate across all recent and ancient lineages in our analyses (**TableS14**). Note that we considered large priors (Uniform[100-100,000] diploid individuals) for constant effective population sizes in all lineages with possible changes at each divergence time, for simplicity. Therefore, the posterior estimates here found may be different in future procedures considering more complex effective demographic regimes likely to have occurred in certain lineages, such as possible bottlenecks and/or population expansions (e.g. [8,40,64]).

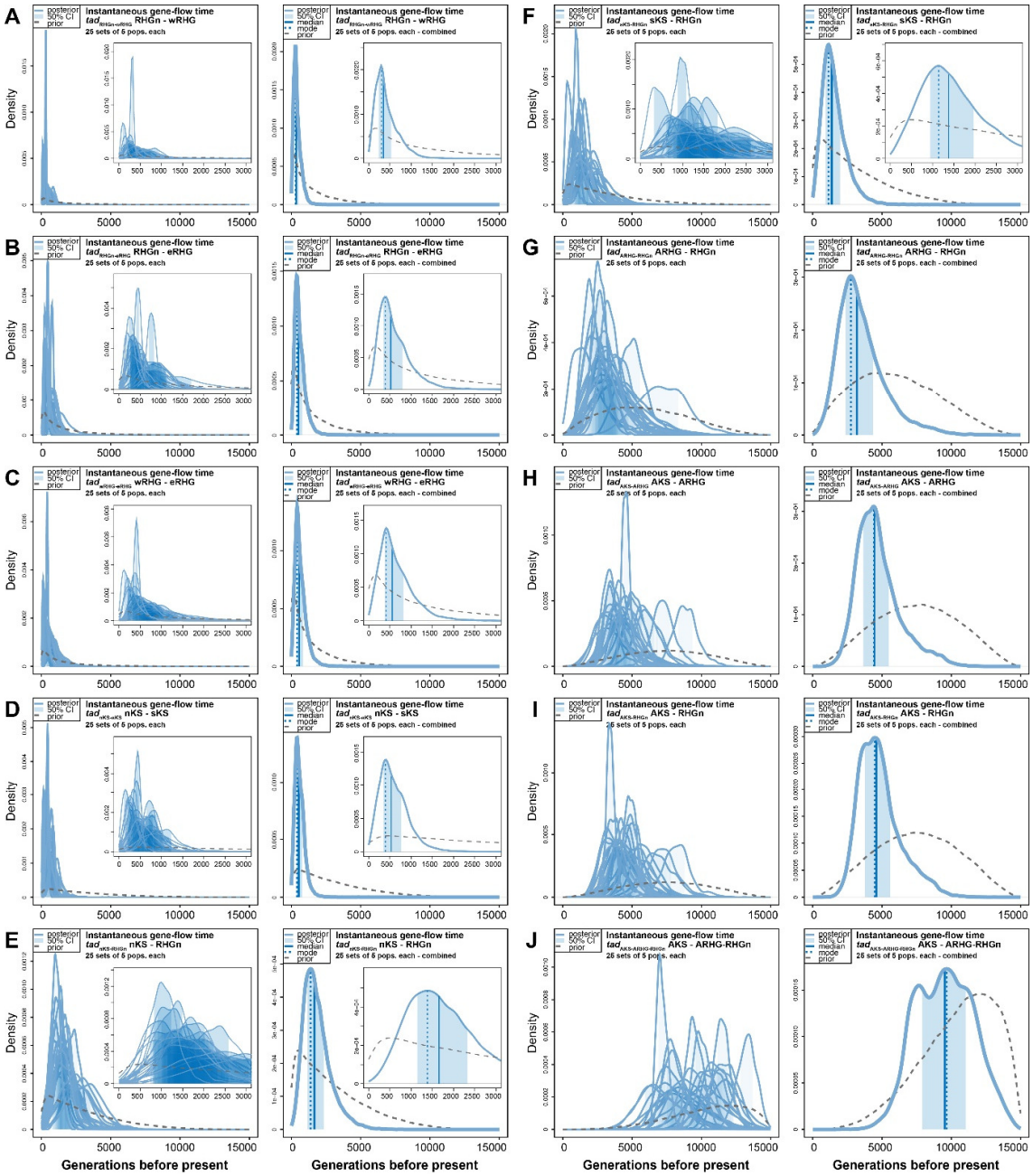
2.D.3. Instantaneous gene-flow times.

In between each lineage divergence times, we estimated the time of occurrence of instantaneous and potentially asymmetric gene-flow exchanges across pairs of lineages (**Figure5**), for each 25 combinations of sampled populations separately. For four out of the six recent gene-flow times between pairs of extant lineages ($tad_{nKS\text{-}sKS}$, $tad_{wRHG\text{-}eRHG}$, $tad_{wRHG\text{-}RHGn}$, $tad_{eRHG\text{-}RHGn}$, **Figure5**, **TableS8**), we find relatively consistent posterior estimates across the 25 tests, almost all with relatively reduced 90%CI, substantial departure from the priors, and low cross-validation posterior parameter estimation errors in the vicinity of the observed data (**Figure8A – Figure8D**, **TableS12, TableS13**, **TableS14**). Note that, while we also obtain similarly satisfactory posterior distributions among the 25 tests for gene-flows among Northern KS and RHGn lineages ($tad_{nKS\text{-}RHGn}$), and among Southern KS and RHGn lineages ($tad_{sKS\text{-}RHGn}$), respectively, we found much larger variance across the 25 different population sets (**Figure8E – Figure8F**, **TableS12**).

Interestingly, we also obtain satisfactory posterior estimates of gene-flow timing among ancient lineages for almost all 25 sets of population combinations ($tad_{ARHG\text{-}RHGn}$, $tad_{AKS\text{-}RHGn}$, $tad_{AKS\text{-}ARHG}$, $tad_{AKS\text{-}ARHG\text{-}RHGn}$, **Figure5**, **TableS8**). Nevertheless, note that posterior estimates are increasingly variable from one set of sampled populations to the other as the estimated parameter is further back in time in the tree-topology (**Figure8G – Figure8J**, **TableS14**).

These results further illustrate that different sample sets used in genetic inferences can provide substantially differing demographic inferences, sometimes for even the most recent gene-flow events, but most often for more ancient gene-flow events across ancestral lineages.

<u>Figure8</u>: ABC posterior parameter distribution of gene-flow times.
Neural Network Approximate Bayesian Computation [43,44], posterior parameter joint estimations of gene-flow instantaneous times *tad* (in generations before present) for 25 sets of five Central and Southern African populations for which the best scenario identified by RF-ABC was Scenario i1-1b (**Figure5**, **TableS11**). Methodological details are provided in **Figure7** caption and detailed in **Material and Methods**.



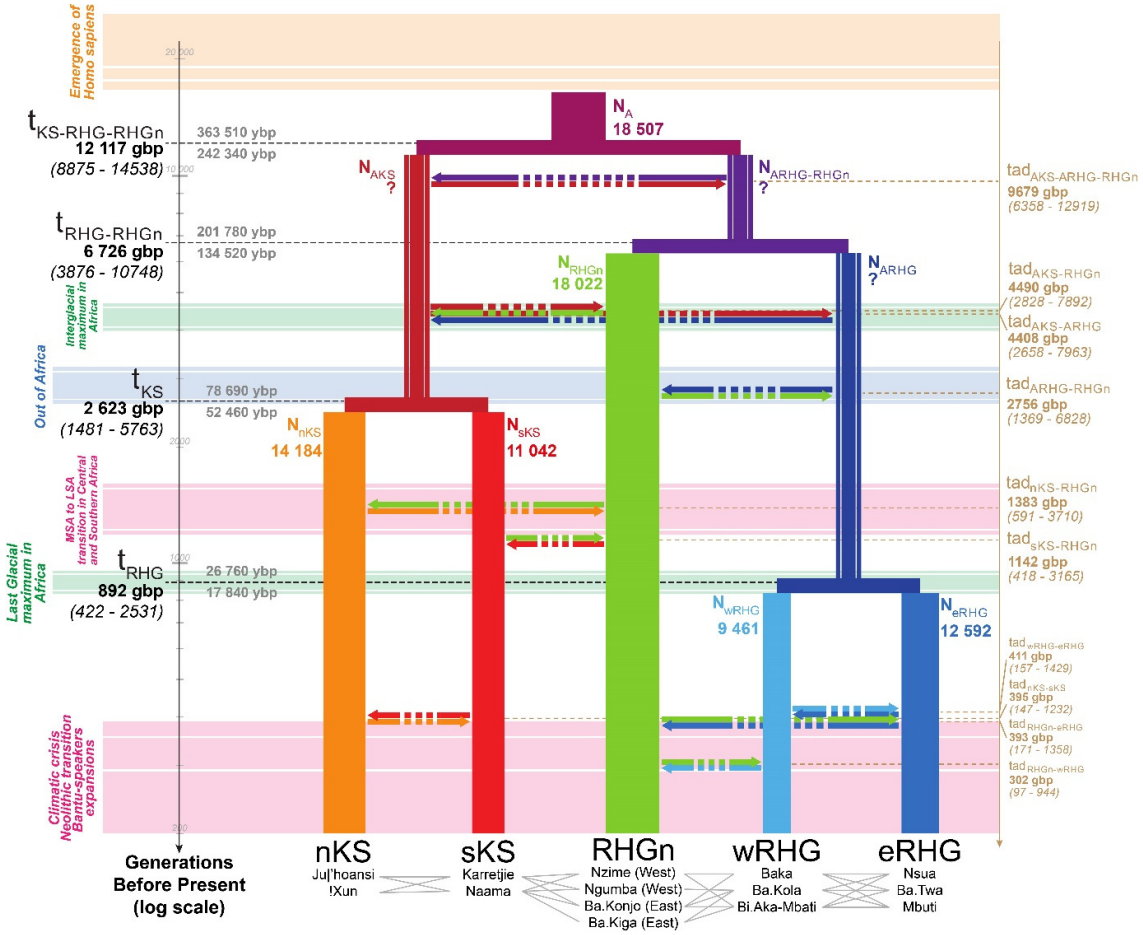
2.D.4. Gene-flow intensities.

We obtain overall unsatisfactory posterior estimates of the 20 separate unidirectional gene-flow rates for the 10 separate instantaneous gene-flow events, for a majority of the 25 sets of population combinations (**FigureS19, TableS12**), with relatively large 90%CI, reduced posterior distributions' departure from priors, and high cross-validation posterior parameter estimation errors in the vicinity of the observed data (**TableS13**). Considering only the several satisfactory posterior estimates for each gene-flow parameter separately, we notice their very large variability across the sets of population combinations for each parameter, making it unreasonable to combine posterior distributions to obtain a synthetic value (**TableS14**). The strong parameter-estimation limitation encountered here is extensively discussed in the **Discussion** section below.

2.D.5. A synthesis of the demographic history of Central and Southern African populations.

We propose, in **Figure9**, a schematic synthesis of all the above inference results for the demographic history of 25 out of 54 combinations of five Central and Southern African populations, for which all our RF-ABC scenario-choice procedures provided systematically consistent results.

<u>Figure9</u>: Schematic inferred demographic and migration history of Central and Southern African populations. Schematic representation of Scenario i1-1b (**Figure5**), and Neural Network ABC posterior parameter mode estimates summarizing results obtained separately for 25 sets of five Central and Southern African populations (**TableS11**), represented by the gray lines in between population names. For the time of each divergence and gene-flow event, mode point estimates are provided in generations before present (gbp) in bold, and 90% credibility intervals are provided between parentheses (**TableS14**). We provide two estimates of the divergence times estimates in years before present (ybp), one (upper) corresponding to 30 years per generation and the other (lower) to 20 years per generation. Mode point estimates of effective population sizes *Ne* are provided in numbers of diploid effective individuals and width of lineages are proportional to the estimated *Ne* (**TableS14**, **FigureS18**). NN-ABC posterior distributions for the effective population sizes of the ancestral Khoe-San lineage (AKS), for the ancestral RHG lineage (ARHG), and for the lineage ancestral to RHGn and ARHG, were all three poorly distinguished from their respective prior distributions and with high cross-validation posterior parameter estimation errors (**FigureS18**, **TableS13, TableS14**), as indicated by the question marks. NN-ABC posterior distributions for instantaneous asymmetric gene-flow rates were overall poorly departing from their priors and with high cross-validation posterior parameter estimation errors (**FigureS19**, **TableS12, TableS13**), as indicated by the doted arrows. All posterior distributions are shown graphically in **Figure7**, **Figure8, FigureS18, FigureS19**, and detailed in **TableS12** and **TableS14**, with cross-validation posterior errors for each parameter in **TableS13**.



Discussion

We aimed at inferring jointly the tree-topology, instantaneous or recurring gene-flow processes, and asymmetric gene-flow intensities across pairs of recent and ancient lineages that most likely produced genome-wide genetic patterns in 54 different combinations of five Central and Southern African populations. Therefore, we explicitly considered the large genetic differentiation observed across African populations at regional and local scales, and the possible confounding effects of complex gene-flow processes on tree-topologies' predictions among 48 complex scenarios in formal statistical competition.

First, our results altogether show that, in fact, considering different sets of populations at a regional and local scale in Africa highlights different aspects of the demographic histories of African populations, sometimes substantially divergent. This demonstrates in practice that apparently discrepant results obtained across previous studies may not necessarily be absolutely reconciled, but rather interpreted as separate illustrations of the large diversity of demographic histories experienced by our species in Africa since its emergence in a remote past. This further strongly advocates for avoiding artificially merging individual samples in un-discriminated *a priori* categories based on geography, subsistence-strategy, and/or linguistics criteria. Instead, we recommend developing novel approaches, probably at a large computational cost as experienced here, to explicitly take into account African genetic diversity at the continental, regional, and/or local scales, and *a posteriori* interpreting the diversity of results obtained separately.

Short periods of gene-flow rather than recurring migrations in the ancient history of Africa

As pointed out here and in numerous previous theoretical or empirical studies (e.g. [22–24,26,31,65,66]), whether ancient gene-flow processes occurred recurrently or more instantaneously across lineages has fundamental consequences on the biologic and cultural evolution of our species. Indeed, recurrent gene-flow processes allow for continuous allelic exchanges across populations -thus never reproductively isolated throughout entire periods of evolution-, which may strongly influence the relative influence of drift and selection across populations and the notion of ancient tree-like evolution in human populations in Africa. Conversely, we found that instantaneous gene-flow processes systematically vastly outperformed recurring migration processes, whichever the tree-topology for ancient and recent Central and Southern African lineages divergences, and for all combinations of five observed populations. Therefore, our results unambiguously favor an evolutionary history of African lineages ancestral to a variety of Central and Southern African populations where *Homo sapiens* populations experienced long periods of isolation and drift, followed by short periods of possibly asymmetric gene-flow, which may in turn have induced some reticulations among lineages [22,67]. Notably, this scenario further allows for differential selection and adaptation processes across lineages (e.g. [6,10,23]), and for ancient admixture-related selection processes within Africa, analogous to previous findings of such processes having occurred more recently throughout the continent [7,9,68,69]).

Nevertheless, it is certain that *Homo sapiens* evolutionary history of ancient and recent gene-flow in Africa has involved more complex processes among pairs of lineages than the ones here identified. Previous studies considered arguably more realistic models to explain their data, with either a conjunction of recurring symmetric migrations among pairs of lineages with possible instantaneous introgressions among certain pre-specified recent and/or ancient lineages [22,24], or only pre-specified instantaneous introgressions without recurring gene-flows [23,26], including specific events involving the contributions from *Homo sapiens* or non-*Homo sapiens* unsampled lineages. In this context, our results highlight that complex instantaneous introgression scenarios should probably be preferred as a starting-

873 point from where one could then increase complexity, for instance by considering more than a single
874 instantaneous gene-flow event among pairs of ancient lineages.
875 Moreover, our results advocate for the possibility of at least some such introgressions to be
876 relatively intense, as scenarios considering low or moderate gene-flows, whether instantaneous or
877 recurring, are systematically poorly mimetic of observed genomic patterns. Echoing this result, a
878 previous study identified such high levels of instantaneous introgressions across African lineages [26],
879 albeit the authors considered very different maximum-likelihood approaches based on different
880 statistics and scenario specifications, as well as different population samples, sometimes artificially
881 merged in larger categories at a continental scale.
882 Nevertheless, our approach failed to provide satisfactory posterior estimates of gene-flow
883 intensities, a serious limitation which we discuss extensively in the following methodological limits
884 and perspectives section below. Importantly, previous studies investigating conjunctions of recurring
885 and more instantaneous migrations overall estimated that symmetric recurring gene-flows were weak,
886 whether recent or ancient (e.g. [22]). Although our results are difficult to compare with previous studies
887 due to strong differences in model specifications, statistics used and methodological approaches as well
888 as population samples, this latter previous result may be reflected in ours that disfavor recurring
889 migration models as being poorly explicative of our data.
890
891 **Inferred demographic history in Central and Southern Africa**
892 Extensive previous work agrees that divergences among lineages ancestral respectively to extant Khoe-
893 San, Rainforest Hunter-Gatherers, and Rainforest Hunter-Gatherer neighbors were among the most
894 ancient in *Homo sapiens* evolution (e.g. reviewed in [2,3]). However, the relative order of their divergences
895 as well as their timing has been a matter of extensive debate [25,26,40]. This is due in part to split-time
896 inferences conducted on tree-like topologies considering gene-flow or no gene-flow, using ancient
897 DNA data or not, and to differing methods, statistics, and population samples.
898
899 *Most ancient divergences in Africa*
900 Here, we formally tested, with Random-Forest ABC scenario-choice, which ancient tree-topology best
901 explained extant genomic patterns, whichever the gene-flow processes and intensities across pairs of
902 recent and ancient lineages, for 54 different combinations of five different Khoe-San, Rainforest
903 Hunter-Gatherer, and Rainforest Hunter-Gatherer neighboring populations. It is the first time to our
904 knowledge that such formal comparison of numerous competing-scenarios is conducted systematically
905 for a variety of sets of population samples at a regional and local scale, rather than comparing
906 maximum-likelihood values from vastly differing models, each obtained separately using differing
907 population samples [22,25,26]. Importantly, note that [24] also explored highly complex demographic
908 scenarios in Africa with ABC, focusing on varying unidirectional introgression processes among certain
909 lineages for a fixed tree-shape among, in particular, three extant African populations at the continental
910 scale.
911 Our results show that whichever the gene-flow processes considered, the lineage ancestral to
912 extant KS populations diverged first from the lineage ancestral to RHG and RHGn, for almost all
913 combinations of sampled populations. Furthermore, we estimated with Neural Network ABC posterior-
914 parameter inferences that this original divergence likely occurred in a remote past ~300,000 years ago,
915 followed by the divergence between ancestral RHG and RHGn lineages ~165,000 years ago,
916 considering between 20 and 30 years per generations.
917 Both estimates fell within the upper bound of previously obtained results considering a variety of
918 similar population samples, genetic data, inference methods, and statistics, albeit scenarios may not
919 have always been specified in analogous ways [4–6,8,25,26,40,70]. Note however, that considering different
920 combinations of populations at the regional and local scales allowed us to identify a small minority

921 (4/54) of such combinations for which an alternative scenario (Scenario i1-3b), where RHGn lineages
922 diverged first from a lineage ancestral to all RHG and KS extant populations, somewhat analogous to
923 those proposed in [26], would better explain the data consistently across our analyses, again whichever
924 the gene-flow processes and range of intensities considered. This latter result may explain apparently
925 discrepant results across some previous studies, which would then be due to differing population
926 samples used for demographic inferences and/or to artificial merging of differentiated populations into
927 larger groups [22,24–26], thus advocating for further accounting for the vast genetic diversity among African
928 populations even at a local scale in future studies.
929 Note that, in [22], the fundamental divergence between their Nama Khoe-San sample and other Sub-
930 Saharan populations was dated to ~110,000-135,000 years ago under their two best-fitting models, thus
931 strongly discrepant with our findings. However, these important results are very difficult to compare
932 with ours, since the authors did not consider any Central African Rainforest Hunter-Gatherers nor their
933 neighbors in their inferences, and since they investigated highly complex models specified completely
934 differently from those here envisioned. In particular, Ragsdale and colleagues included in their models
935 possible very ancient genetic structures, long before *Homo sapiens* emergence, a feature that is
936 unspecified in our scenarios which considered simply a single ancestral population in which all extant
937 lineages ultimately coalesce. Nevertheless, note that substructure and reticulation within the ancestral
938 population is not *per se* incompatible with our scenarios. In fact, it may be compatible with our posterior
939 estimates of a large effective population ancestral to all extant populations here investigated, the largest
940 among all inferred ancient and recent Central and Southern African effective population sizes.
941 Therefore, it will be reasonable in future work to complexify the scenarios here proposed to evaluate
942 whether very ancient substructures and reticulations within our ancestral population, prior to the
943 original divergence between Southern and Central African populations, may improve the fit to the
944 observed genomic data, as proposed by [22].
945
946 *More recent divergences in Central and Southern Africa*
947 More recently during the evolutionary history of Sub-Saharan Africa, we found that the divergence
948 among Northern and Southern Khoe-San populations largely pre-dated the divergence of Western and
949 Eastern Congo Basin Rainforest Hunter-Gatherers, a question rarely addressed to our knowledge. First,
950 we found that KS divergence dated sometime between 50,000 and 80,000 years ago, thus substantially
951 more recently than estimates previously proposed [2]. Beyond vast differences in models, methods, and
952 statistics used to provide either inference, note that our results for this divergence time were
953 substantially variable across pairs of sampled populations used in each analysis, some specific sets of
954 populations providing posterior estimates consistent with previous result. Thus, it further highlights that
955 complex inferences in Africa imperatively need to explicitly consider population variation at a local
956 scale.
957 Interestingly, our synthesized estimates for the Northern and Southern Khoe-San population
958 divergence were relatively synchronic to the genetic onset of the Out-of-Africa (e.g. [2]). Population
959 genetics inferences only provide possible mechanisms to explain observed genetic patterns and are, in
960 essence, not addressing the possible causes underlying the inferred mechanisms. In this context, we
961 may hypothesize that the global climatic shifts inducing massive ecological changes that have occurred
962 in Africa at that time (e.g. [71]), sometimes proposed to have triggered ancient *Homo sapiens* movements
963 Out-of-Africa, may also have triggered, independently, the genetic isolation among ancestral Khoe-San
964 populations. Nevertheless, where the ancestors of extant Khoe-San populations lived at that time
965 remains unknown and is nevertheless crucial to further elaborate possible scenarios for the causes of
966 the genetic divergence here inferred.
967 Long after this divergence, we found that Rainforest Hunter-Gatherer populations across the
968 Congo Basin diverged roughly between 17,000 and 27,000 years ago, relatively consistently across

969 pairs of sampled populations used for inferences; estimates highly consistent with previous studies
970 [8,70,72], despite major differences in gene-flow specifications across RHG groups between studies.
971 Interestingly, this divergence time is relatively synchronic with absolute estimates for the Last-Glacial
972 Maximum in Sub-Saharan Africa (e.g. [73]). The fragmentation of the rainforest massif during this period
973 in the Congo Basin may have induced isolation between Eastern and Western RHG extant populations,
974 as plausibly previously proposed [70]. However, similarly as above for the Northern and Southern Khoe-
975 San populations divergence, where the ancestors of extant Eastern and Western Rainforest Hunter-
976 Gatherers lived remains unknown, which prevents us from formally testing this hypothesis [74].
977 Altogether, these results show that Central African Rainforest Hunter-Gatherer and Southern
978 African Khoe-San populations have had, respectively, extensive time for selection processes, including
979 adaptive introgression processes, to have influenced independently both groups of populations as well
980 as populations within each group separately (e.g. [6,7,9,10]).
981
982 *Ancient and recent instantaneous gene-flow times in Africa*
983 We obtained reasonably well estimated instantaneous asymmetric gene-flow times among almost all
984 pairs of ancient and recent lineages, albeit the variation of estimates across sets of Central and Southern
985 African sampled populations increased substantially with most ancient times. In this context, we deem
986 it hard to confidently try to interpret the most ancient event of instantaneous gene-flow between the two
987 most ancestral lineages in our tree-topology. However, other instantaneous gene-flow time estimates
988 throughout the topology were more consistently estimated overall, and showed relative synchronicity
989 in some cases, which has never been reported before to our knowledge, even if they were specified
990 independently in our models and drawn from large distribution *a priori*.
991 Interestingly, we found strong indications for almost synchronic events of introgressions having
992 occurred during the Last Interglacial Maximum in Africa [67], between ~90,000 and ~135,000 years ago
993 (when considering 20 or 30 years per generation). They involved gene-flow between lineages ancestral
994 to Khoe-San populations and ancestors of Rainforest Hunter-Gatherer neighbors on the one hand and,
995 on the other hand, between lineages ancestral to Khoe-San populations and the lineage ancestral to all
996 Rainforest Hunter-Gatherers. An increase in material-based culture diversification and innovation,
997 possibly linked to climatic and environmental changes locally, has previously been observed during
998 this period of the Middle Stone Age in diverse regions of continental Africa; prompting a long-standing
999 debate as to its causes if human populations were subdivided and isolated biologically and culturally at
1000 the time [75–78].
1001 In this context, our results instead may suggest that population movements at that time among
1002 previously isolated populations may itself have triggered the observed increased cultural diversification
1003 locally, even if obvious signs of pan-African cultural spread at the time are difficult to assess [77]. In turn,
1004 it would interestingly echo the known effect of within-population genetic diversity increase induced by
1005 genetic admixture between previously isolated populations (e.g. [18,31,65,79]).
1006 Then, note that we estimated that the instantaneous gene-flow event between the ancestral
1007 Rainforest Hunter-Gatherers lineage and that of their extant neighbors seemingly occurred
1008 synchronically to the genetic Out-of-Africa ([71]; see above). This would imply that possible climatic and
1009 ecological shifts at that time may not have only induced population divergences and displacement, but
1010 may also have triggered population gene-flow.
1011 Relatively more recently, around 30,000 years ago, we found two loosely synchronic gene-flow
1012 events between ancestors to extant Central African Rainforest Hunter-Gatherer neighbors' lineages and,
1013 separately, Northern and Southern Khoe-San lineages. This corresponds to the end of the Interglacial
1014 Maximum and a period of major cultural changes and innovations during the complex transition from
1015 Middle Stone Age to Late Stone Age in Central and Southern Africa [75,78,80–82]. Nevertheless, connecting
1016 the two lines of genetic and archaeological evidence to conclude for increased population movements

at the time and their possible causes should be considered with caution. Indeed, in addition to genetic-dating credibility-intervals being inherently much larger than archaeological dating, this period remains highly debated in paleoanthropology mainly due to the scarcity and complexity of the material-based culture records, and that of climatic and ecological changes locally, across vast regions going from the Congo Basin to the Cape of Good Hope [75,78,80–82].

Finally, we found strong signals for multiple instantaneous gene-flow events having occurred between almost all five recent Central and Southern African lineages between 6000 and 12,000 years ago, during the onset of the Holocene in that region, shortly before or during the beginning of the last Post Glacial Maximum climatic crisis in Western Central Africa [83], the emergence and spread of agricultural techniques [84], and the demic expansion of now-Bantu-speaking populations from West Central Africa into the rest of Central and Southern Africa [9,17,85]. These results are consistent with previous investigations that demonstrated the determining influence of Rainforest Hunter-Gatherer neighboring populations' migrations through the Congo Basin in shaping complex socio-culturally determined admixture patterns [4,5,8,70], including admixture-related natural selection processes [9,10,72,86]. As our estimates for introgression events are in the upper bound of previous estimates for the onset of the so-called "Bantu expansion" throughout Central and Southern Africa, we may hypothesize here that major climatic and ecological changes that have occurred at that time may have triggered increased population mobility and gene-flow events between previously isolated populations, rather than consider that the Bantu-expansions themselves were the cause for all the gene-flow events here identified.

Finally, we did not find signals of more recent introgression events from Bantu-speaking agriculturalists populations into Northern or Southern Khoe-San populations, in particular among the !Xun, albeit such events have been identified in several previous studies (see [2]). This is likely due to the fact that we considered only a limited number of individual samples from each population, and therefore may lack power to detect these very recent events with our data and approach.

Conceptual, methodological, and empirical limitations and perspectives for inferring ancient histories from observed genomic data

All previous attempts at reconstructing the human evolutionary histories which led to genomic patterns observed today across African populations have faced major conceptual, methodological, and empirical challenges. Conceptually, large amounts of more-or-less nested scenarios can be envisioned *a priori* to explain extant genetic diversity, based on previous results from paleo-anthropology, population genetics, and paleogenomics. These scenarios may range from tree-like models without gene-flow events among ancient or recent lineages to complex networks of weakly differentiated populations exchanging migrants over large periods of time, with or without the contribution of ancient *Homo* or non-*Homo* now extinct or unsampled lineages. Systematically exploring all possible models is often methodologically out of reach due to differing fundamental scenario-specifications or, when scenarios are specified and parameterized in analogous ways, due to scenarios' nestedness and un-identifiability in certain parts of their parameter spaces (e.g. [87]). Empirically, formal scenario comparisons are first hampered by necessarily limited amounts of genomic data representative, at continental, regional, and local scales, of the known diversity and differentiation of human populations in Africa; in addition to yet limited amounts of ancient DNA data throughout the continent. Finally, empirical limitations also emerge from the use of different statistics to explore genomic diversity patterns, which possibly each capture different facets of human evolutionary histories, thus providing discrepant results and interpretations only in appearances. While machine-learning ABC procedures provide significant advantages over other maximum-likelihood approaches, in particular concerning the formal exploration of competing scenarios' fit to observed data across numerous highly complex sometimes nested scenarios and using numerous summary-statistics [43,45,63,87], the above challenges are also largely faced in this study.

While most divergence and gene-flow times, as well as effective sizes, were inferred satisfactorily in our results, the lack of satisfactory posterior estimates for all 20 gene-flow rates parameters consistent among the 25 sets of population combinations likely stems from different limitations. First, considering only five individual genomes per population is inherently limiting when trying to estimate gene-flow rates from inter-population summary-statistics in ABC [88]. In future studies, increasing sample sizes and adding statistics based on the inter and intra-individual distribution of admixture fractions [31,32,65,88], will likely improve the posterior estimation of these parameters using machine-learning ABC approaches [24]. Note, however, that these statistics require non-trivial computation time in ABC frameworks comprising hundreds of thousands of simulations [89,90]. Furthermore, in all cases, considering multiple sets of different and substantially genetically diverse populations as the ones here considered from sub-Saharan Africa may inevitably lead to large variation across gene-flow rates' posterior estimates depending on the specific population-combinations.

Second, for simplicity, we considered a single instantaneous time for the two separate unidirectional gene-flow events, for each gene-flow event between pairs of lineages separately. Therefore, while our RF-ABC scenario-choice results strongly support instantaneous asymmetric gene-flow events rather than recurring asymmetric gene-flows to best explain extant genomic patterns, it is plausible that scenarios where each unidirectional gene-flow event may occur at a different time [24], will more realistically explain the observed data. If this is the case, it is possible that the choice of a unique time for two separate gene-flow events rendered the corresponding gene-flow rates harder to identify with our joint NN-ABC posterior parameter estimation procedure. Furthermore, we explored two extreme gene-flow processes, establishing an open competition between only instantaneous gene-flow and only recurring gene-flow processes. While we demonstrated that only instantaneous gene-flow processes vastly outperformed only recurring gene-flow processes, whichever the range of intensities for each event, we only established here a starting point for the future complexification of evolutionary scenarios to be tested. For instance, it will be of natural interest to consider, next, more than a single "pulse" of gene-flow between any two ancient or recent lineage.

Third, the neural-networks used here may have been unable to satisfactorily identify the 20 gene-flow parameters among the 43 jointly-estimated parameters, due to their lack of complexity when considering a unique layer of hidden neurons. For instance, and as a first step, considering Multilayer-Perceptrons in the future, *i.e.* more complex feed-forward neural-networks with additional layers of hidden neurons, may help posterior estimations of these parameters (e.g. [91]), at the known cost of non-trivial parameterization of the neural-networks themselves (e.g. [90,92,93]).

Altogether, joint posterior estimation of numerous gene-flow rates parameters and the timing of their occurrence under highly complex demographic scenarios remains one of the most challenging tasks in population genetics. It will unquestionably benefit from future analytical theoretical developments [94,95], and the improvement of machine-learning-based inference procedures (e.g. [93,96–98]).

Archaic admixture with *Homo sapiens* or non-*Homo sapiens* unsampled lineages?

Our results demonstrate that explicitly considering archaic admixture from unsampled populations is not necessary to satisfactorily explain the observed genomic diversity of extant Central and Southern African populations when considering jointly the 337 relatively classical population genetics summary-statistics used here for demographic inferences, consistently with a previous study [22], and conversely to others [3,25,26]. We did not explore possible contributions from unsampled lineages, whether from non-*Homo sapiens* or from ancient "ghost" human populations, and therefore cannot formally reject such possibilities. However, by parsimony, our results show that the simpler explanation (without significant archaic admixture signature at the genomic level), may possibly be more likely.

Our set-up formally comparing competing-scenarios rather than comparing posterior likelihoods of highly complex yet vastly differing models, provide a clear and reasonable starting point for future

complexification of scenarios comprising possible contributions from ancient or ghost unsampled populations; where the burden of proof lies on showing that scenarios with archaic admixture fit the data significantly better than scenarios without such archaic admixture (despite the increased complexity and more parameters). Such endeavors will also benefit from the explicit use of additional novel summary-statistics ([22,26,32]; see also above).

In any case, the complexification of scenario-specifications to account for possible past "archaic" or "ancient" introgressions will not fundamentally solve the issue of the current lack of reliable ancient genomic data older than a few hundreds or thousands of years from Sub-Saharan Africa [3,99,100]. Indeed, analogously to archaic admixture signals that were unambiguously identified outside Africa only when ancient DNA data were made available for Neanderthals and Denisovans (e.g. [27,28]), we imperatively need to overcome this lack of empirical ancient DNA data in Africa to formally test whether, or not, ancient human or non-human now extinct lineages have significantly contributed to shaping extant African diversity.

Material and Methods

Population samples

Central and Southern Africa dataset

We investigated high-coverage whole genomes newly generated for 74 individual samples (73 after relatedness filtering, see below), from 14 Central and Southern African populations (**Figure1**, **Table1, TableS1**). Based on extensive ethno-anthropological data collected from semi-directed interviews with donors and their respective communities, we grouped *a posteriori* the 73 individual samples from 14 populations in three larger categories.

The Baka, Ba.Kola, Bi.Aka_Mbati, Ba.Twa, and Nsua individuals were categorized as Rainforest Hunter-Gatherers (RHG), based on several criteria including self-identification and relationships with other-than-self, socio-economic practices and ecology, mobility behavior, and musical practices [4,38].

The Nzime, Ngumba, Ba.Kiga, and Ba.Konjo individuals were categorized as Rainforest Hunter-Gatherer neighbors (RHGn), based on the same set of categorization-criteria as RHG. Indeed, RHG and RHGn populations in Central Africa are known to identify themselves separately and to live in different ways in the Central African rainforest, whilst often sharing languages locally as well as complex socio-economic relationships and interactions that may include intermarriages [5]. Historically, RHG populations have been designated as "Pygmies" by European colonists, an exogenous term from ancient Greek that is sometimes used derogatorily by the RHGn.

Finally, the Nama, Ju|'hoansi, Karretjie People, !Xun, and Khutse_San individuals were categorized as Khoe-San populations (KS) based primarily on self-identification and relationships with other-than-self locally, lifestyle, and languages [101]. Note that research was approved by the South African San Council. Individuals from these populations were whole-genome sequenced anew with improved methods (see below), compared to previous work [40].

Note that the three groups can be further subdivided based on geography into Western and Eastern RHG and RHGn, and into Northern, Central, and Southern KS (**Figure1**, **Table1**).

Comparative dataset

The 74 genomes were analyzed together with 105 (104 after relatedness filtering, see below) high-coverage genomes from 32 worldwide populations (**Figure1**, **Table1, TableS1**). 64 samples from the Simon's Genome Diversity Project (SGDP) [34]; 9 HGDP samples [27]; 7 samples from the 1000 Genomes project (KGP) [35]; one Karitiana sample [36]; 24 samples from the South African Human Genome Project (SAHGP) [37]. Selection criteria were: Illumina paired-end reads, high-coverage (>30X), access to raw data and no (known) population substructure.

Note that, although samples from these datasets have not been collected by us for the purpose of this study, based on previous knowledge and publications, the Biaka and Mbuti HGDP samples can be categorized as RHG and the #Khomani and Ju|'hoansi_comp samples as KS (**Figure1**, **Table1**).

Data generation

DNA extraction

DNA was extracted from saliva with the DNA Genotek OG-250 kit for the Baka, Nzime, Bi.Aka Mbati, Nsua, and Ba.Konjo samples. DNA was extracted from buffy coats with DNeasy Blood&Tissue spin-

column Qiagen™ kits for the Ba.Kola, Ngumba and Ba.Kiga samples. Both extraction methods were used for the Ba.Twa samples. For the Nama, Ju|'hoansi, Karretjie People, !Xun, and Khutse_San samples, DNA was extracted from EDTA-blood using the salting-out method [102].

Library preparation and sequencing

Library preparation and sequencing was performed by the SciLifeLab SNP&SEQ Technology platform in Uppsala, Sweden. Libraries were prepared with TruSeq DNA preparation kits, and paired-end sequencing (150 bp read length), was performed on Illumina HiSeqX machines with v2 sequencing chemistry, to a coverage of 30X or more. For the Southern African samples, we included paired-end data (100 bp read length) sequenced on Illumina HiSeq2000 machines obtained previously for the same libraries [40].

Sequencing data quality-control processing and relatedness filtering

The processing pipeline used in the present study is adapted from the Genome Analyses Toolkit (GATK) "Germline short variant discovery (SNPs + Indels)" Best Practices workflow [103–105]. It is described and compared to the original GATK Best Practices workflow in [106]. **FigureS20.** gives an overview of the pipelines used for the processing; all template codes with accompanying detailed explanations for all processing steps are provided below and in the corresponding GitHub repository (https://github.com/Gwennid/africa-wgs-descriptive).

Reads were mapped to a decoy version of the human reference genome, GRCh38 (1000 Genomes Project version), with BWA-MEM [107] from BWAKIT v0.7.12 using a strategy appropriate for ALT contigs. The input was either FASTQ or BAM files that were reverted to unmapped BAM or to FASTQ prior to mapping. Mapped reads were selected with SAMtools version 1.1 and Picard version 1.126 (https://broadinstitute.github.io/picard/).

Duplicate reads were marked (at the lane level) with Picard version 1.126 MarkDuplicates. Realignment around indels was performed (at the lane level) with GATK version 3.5.0 RealignerTargetCreator and IndelRealigner. We then performed a "triple mask base quality score recalibration (BQSR)", where the sample's variation is used together with dbSNP, as described in [40,106]. This step includes merging the reads from a given sample with SAMtools version 1.1; calling variants with GATK version 3.5.0 HaplotypeCaller with "--genotyping_mode DISCOVERY"; and (in a parallel track) standard BQSR with GATK version 3.5.0 BaseRecalibrator and PrintReads, and dbSNP version 144. Following triple mask BQSR (performed at the lane level), the recalibrated reads from a given sample are merged with SAMtools version 1.1, sorted and indexed with Picard version 1.126 SortSam. Duplicate marking and realignment around indels were performed again, at the sample level.

Variant calling was performed independently for the autosomes and the X chromosomes, due to the difference in ploidy. SNP and short indels were first called with GATK version 3.7 HaplotypeCaller in each genome, with the "--emitRefConfidence BP_RESOLUTION" option. Multi-sample GVCFs were then generated with CombineGVCF, and variants were finally jointly called with GATK GenotypeGVCFs. Genotypes were emitted for each site (option "allSites").

Variants were filtered with GATK version 3.7 Variant Quality Score Recalibration (VQSR) using recommended resources from the GATK bundle: HapMap (version 3.3), 1000 Genomes Project Illumina Omni 2.5M SNP array, 1000 Genomes Project phase 1 high confidence SNPs, and dbSNP version 151 for the SNPs; and [108] gold standard indels and dbSNP version 151 for the indels. For the X chromosome, autosomal variants were included for training the VQSR model (VariantRecalibrator step).

We controlled the dataset for related individuals with KING [109] "--kinship" and plink version 1.90b4.9 [110] "--genome --ppc-gap 100". The dataset was pruned for linkage disequilibrium (LD) before

1212 relatedness estimation with plink, using plink "--indep-pairwise" with sliding windows of 50 SNPs,
1213 shifting by five SNPs, and a r^2 threshold of 0.5. Both methods identified two pairs of first-degree
1214 relatives. We excluded the sample with greatest missingness from each pair to produce the family
1215 unrelated working datasets henceforth used.
1216 The callset was then further refined; two samples with a first degree relative in the dataset were
1217 excluded with GATK version 3.7 SelectVariants; a "FAIL" filter status was set on sites that are
1218 ambiguous in the reference genome (base "N") or had greater than 10% missingness, identified with
1219 VCFtools version 0.1.13 "--missing-site" [111]. Moreover, for the autosomes, variants heterozygous in all
1220 samples were marked as failed (Hardy Weinberg Equilibrium -HWE- filter, identified with VCFtools
1221 version 0.1.13 "--hardy").
1222 Coverage was computed with QualiMap version 2.2 [112] including or not duplicates ("-sd" option
1223 for the latter). The average coverage (without duplicates) across individuals within populations is
1224 provided in **TableS3**.

1225 Descriptive analyses

1226 Variant counts

1227 We obtained per-sample and aggregated metrics of variant counts (**Figure2**), with the tool
1228 *CollectVariantCallingMetrics* of the software package Picard v2.10.3
1229 (https://broadinstitute.github.io/picard/), applied to the full 177 worldwide individuals' dataset after
1230 VQSR, relatedness, HWE and 10% site missingness filtering. We conducted variant counts procedures
1231 for chromosomes 1 to 22. We used dbSNP 156 as a reference; we downloaded
1232 "GCF_000001405.40.gz" and modified the contig names to match contig names in the VCFs.
1233 Separately, we applied the same variant counts pipeline as above to the Central and Southern African
1234 73 individuals' original dataset extracted from the full dataset with BCFtools version 1.17
1235 (https://github.com/samtools/bcftools), using view with the option "-S". All scripts are provided in the
1236 corresponding GitHub repository (https://github.com/Gwennid/africa-wgs-descriptive).

1237 Genome-wide Heterozygosity

1238 We calculated observed and expected heterozygosities, for the autosomes and for the X-chromosome
1239 separately using custom Python, Bash, and R scripts (full pipeline available at GitHub
1240 https://github.com/Gwennid/africa-wgs-descriptive), for the 177 worldwide individuals after variant-
1241 counts filtering described above. For autosomes, we considered the main contigs ("chr1" to "chr22").
1242 For the X chromosome, we excluded the Pseudo-Autosomal Regions with coordinates from GRCh38
1243 (PAR1: 10,000 to 2,781,480, PAR2: 155,701,382 to 156,030,896).
1244 For each population with more than one individual and for each autosome and X-chromosome
1245 separately, we counted the number of variable and non-variable sites and excluded multi-allelic sites,
1246 indels, and sites with missing genotypes in at least one individual in the population for simplicity and
1247 conservativeness. For each autosome and each individual, we then counted separately the numbers of
1248 observed homozygous and heterozygous sites of each configuration compared to the reference (0/0;
1249 0/1; 1/0; 1/1). Based on these counts, for each population and each chromosome separately, we first
1250 calculated observed heterozygosities simply as the average proportion of heterozygous individuals per
1251 variable sites with no-missing data in the population sample. We also computed unbiased expected
1252 multi-locus heterozygosities for all variable loci with no missing genotypes in the population as in
1253 equations (2) and (3) in [47]. We averaged this value across all sites with no missing information in the
1254 population sample including both variable and non-variable sites, and finally corrected it for haploid

population sample sizes (**Figure2**). We compared unbiased heterozygosities by chromosomes across groups of populations using Wilcoxon signed-rank tests. All scripts are provided in the corresponding GitHub repository (https://github.com/Gwennid/africa-wgs-descriptive).

Runs of homozygosity

We identified runs of homozygosity (ROH) using the *homozyg* tool from PLINK version 1.90b4.9 [110]. We selected autosomal biallelic SNPs from the variant-counts pipeline described above with GATK version 3.7 SelectVariants with options "-selectType SNP -restrictAllelesTo BIALLELIC -excludeFiltered". We converted the VCF to TPED and then binary plink fileset with VCFtools version 0.1.13 "-plink-tped" and plink. For the 177 individuals worldwide, we considered only ROHs measuring more than 200 Kb (*--homozyg-kb* 200), containing at least 200 SNPs (*--homozyg-snp* 200), containing at least one variable site per 20 Kb on average (*--homozyg-density* 20), and with possible gaps of up to 50 Kb (*--homozyg-gap* 50). Default values were considered for all other parameters: *--homozyg-window-het* 1 *--homozyg-window-snp* 50 *--homozyg-window-threshold* 0.05. Finally, the total mean ROH lengths were calculated for each population separately using the ".hom" output-file for each of four length categories: 0.2 to 0.5 Mb, 0.5 to 1 Mb, 1 to 2 Mb, and 2 to 4 Mb (**Figure2**). Corresponding pipelines are provided in the corresponding GitHub repository (https://github.com/Gwennid/africa-wgs-descriptive).

Individual pairwise genome-wide genetic differentiation

We explored genome-wide genetic differentiation between pairs of individuals with Neighbor-Joining Tree (NJT) [52,53], and Multi-Dimensional Scaling (MDS) approaches based on the pairwise matrix of Allele-Sharing Dissimilarities (ASD) [51] computed using 14,182,615 genome-wide autosomal SNPs pruned for low LD, with PLINK version 1.90b4.9 [110] *indep-pairwise* function (*--indep-pairwise 50 5 0.1*). We considered at first all 177 unrelated individuals in our dataset to calculate the ASD matrix, and then subset this ASD matrix for the 73 Central and Southern African unrelated individuals original to this study before computing NJT and MDS analyses anew (**Figure3**).

We computed the ASD matrix considering, for each pair of individuals, only those 14,182,615 SNPs without missing data, using the *asd* software (v1.1.0a; https://github.com/szpiech/asd; Szpiech, 2020). We computed the unrooted NJT using the *bionj* function of the R package *ape* and setting the branch-length option to "true". We computed the MDS using the *cmdscale* function in R.

We compared ASD values with populations' geographic locations by computing the standardized average ASD across all pairs of individuals from two different populations, for all pairs of African populations separately (**Figure1**, **Table1, TableS1**), and compared these values with the logarithm of geographic distances among pairs of populations (setting 0 km distances to "1"), using Spearman ρ and Pearson r correlations for different continental, sub-continental, and regional sets of populations. We tested these correlations using 1000 Mantel permutation tests (**TableS4**), with the *partial.mantel.test* function of the *ncf* package in R. We computed the shortest geographic distance between two populations' GPS coordinates on an Earth-like ellipsoid (with an earth radius of 6378.137 km by default in WGS84), using the *distm* function with *distGeo* method from the *geosphere* package in R.

We conducted permutational ANOVA analyses to test, separately, whether linguistic families, sub-linguistic families, or traditional food-producing strategies (**TableS1**), significantly explain average population-pairwise ASD patterns, using the *adonis2* function from the *vegan* package in R with 1000 permutations for each analysis conducted on subsets of the above ASD matrix (**Figure3**) at varying geographical scales (**TableS5**, **TableS6**, **TableS7**).

The clustering software ADMIXTURE [54] allows researchers to further explore inter-individual genome-wide levels of dissimilarity and resemblance. Indeed, while it is tedious and often cognitively

difficult to explore multiple combinations of dimensions of genetic variation using an MDS or a NJT approach, ADMIXTURE instead allows exploring K higher such dimensions at once [54–58]. However, note that the visual representation of relative "distances" across pairs of individuals is lost in the classical bar-plot representation of ADMIXTURE, hence showing the complementarity of this descriptive analysis to the above MDS and NJT.

We considered here only 840,031 genome-wide autosomal SNPs pruned for LD (r^2 threshold 0.1) and with Minimum Allele Frequency above 0.1, for the 177 worldwide unrelated individuals, using PLINK *indep-pairwise* and *maf* functions. We computed unsupervised ADMIXTURE version 1.3.0 clustering with values of K ranging from 2 to 10, considering 20 independent runs for each value of K separately. We then calculated ADMIXTURE results symmetric-similarity-coefficient SSC for each value of K separately in order to find the groups of runs providing highly similar results (SSC>99.8%). Individual's genotype membership proportions to each K cluster were then averaged, per individual, across such highly resembling runs, and then plotted. SSC calculations, averaging results across similar runs, and producing barplots were conducted using the software PONG with the "greedy" algorithm [60]. Corresponding pipelines are provided in the corresponding GitHub repository (https://github.com/Gwennid/africa-wgs-descriptive). For each value of K separately, results representing the majority of the 20 independent runs are presented in **Figure4**, and all other results are presented in supplementary **FigureS1**.

Machine-Learning Approximate Bayesian Computation scenario-choice and posterior parameter estimation

We reconstructed the complex demographic history of Central and Southern African populations using machine-learning Approximate Bayesian Computation [41–45]. In principle, in ABC, researchers first simulate numerous genetic datasets under competing scenarios by drawing randomly a vector of parameter values for each simulation in distributions set a priori by the user. For each simulation separately, we then calculate a vector of summary statistics, thus corresponding to a vector of parameter values used for the simulation. The same set of summary statistics is then computed on the observed data. ABC scenario-choice then allows the researcher to identify which one of the competing scenarios produces the simulations for which summary-statistics are closest to the observed ones. Under this best scenario, ABC posterior-parameter inference procedures allow the researcher to estimate the posterior distribution of parameter values most likely underlying the observed genetic patterns.

48 competing scenarios for the demographic history of Central and Southern African populations

We designed 48 competing demographic-history scenarios possibly underlying the genetic patterns observed in five Central and Southern African populations of five individuals each (**Table1**, **Figure5**); one Eastern Rainforest Hunter-Gatherer population (eRHG: Nsua, Ba.Twa, or Mbuti), one Western Rainforest Hunter-Gatherer population (wRHG: Baka, Ba.Kola, or Aka Mbati), one Eastern or Western RHG neighboring population (eRHGn: Ba.Kiga or Ba.Konjo; wRHGn: Nzimé or Ngumba), one Northern Khoe-San population (nKS: Ju|'hoansi or !Xun), and one Southern Khoe-San population (sKS: Karretjie People or Nama). Note that we did not include the Khutse-San population in ABC inferences for simplicity, as this population is located at intermediate geographic distances between the nKS and sKS groups of populations.

The 48 competing scenarios differed in their ancient tree-topologies and relative timing of the more recent divergence events, combined with different (duration and intensities) possibly asymmetric gene-

flow processes among ancient and recent lineages (**Figure5**). This design, while complex, allowed us, for the first time to our knowledge, to consider explicitly the expected confounding effects of gene-flow processes on otherwise different tree-topologies, which may also possibly induce certain reticulations among ancient lineages, while jointly estimating divergence times, gene-flow events' timing and intensities, and effective population size changes over time.

Importantly, we relied on extensive previous findings having formally demonstrated the common origin of Congo Basin RHG populations [4,70], and that of KS populations [6,40], respectively, and thus did not consider all possible tree-topologies for five tree-leaves. Moreover, we considered only a single RHGn sample in each combination, in turn from the East or the West of Central Africa, in order to simplify already highly complex scenarios, as we were not interested here in the detailed demographic history of RHGn which has been extensively studied previously (e.g. [9,17]). This simplification was deemed reasonable as RHGn populations throughout the Congo Basin and into Southern Africa have previously been shown to be strongly more genetically resembling one another, compared to RHG and KS neighboring populations, or compared to much more genetically dissimilar RHG and KS populations, respectively (e.g. [4,5,8,9,17,72,86]); a result that we also obtained here (**Figure5**).

Eight competing topologies

We investigated eight competing topologies starting with a single ancestral population and resulting in five different sampled populations in the present (**Figure5**). The eight topologies differ in which lineage ancestral to modern KS, RHG, or RHGn diverged first from the two others, and in the relative order of divergence events internal to KS (nKS-sKS divergence) and internal to RHG (wRHG-eRHG divergence), in order to consider the known variable demographic histories of Sub-Saharan populations at a regional scale.

Note that divergence and introgression times (all *t* and *tad*, see **Figure5**, **TableS8**), were each randomly drawn in uniform prior distributions between 10 and 15,000 generations, thus effectively setting an upper limit for the most ancient divergence times among lineages at roughly 450,000 years ago considering an upper-bound of 30 years for human generation duration [113]. This was vastly anterior to the current estimates for the genetic or morphological emergence of *Homo sapiens* [114,115], which allowed us to estimate *a posteriori* the most ancient divergences among our populations, without constraining our assumptions based on previous results obtained with different methods, data, and models. Furthermore, these prior-distribution boundaries allowed for gene-flow events among ancient lineages (see next section), to influence, potentially, the timing of the earliest divergence events among human lineages previously estimated [2,22,26]. Note that we retained for simulations only those vectors of randomly-drawn parameter-values that satisfied the chronological order of lineage divergences set for each eight topologies, respectively (**Figure5**, **TableS8**)

In each eight topologies, we incorporated the possibility for changes in effective population sizes, *N* (**Figure5**), during history along each lineage separately by defining constant diploid effective population sizes parameters separately for each tree-branch, each drawn randomly in U[10-100,000] (**Figure5**, **TableS8**).

Asymmetric instantaneous or recurring gene-flows and their intensities

Migration of individuals between populations is ubiquitous in human history (for an overview in Africa, see e.g. [2,3]). Here we aimed at disentangling the nature of migration processes that may have occurred across pairs of lineages throughout the history of Central and Southern African populations.

In particular, we first aimed at determining whether gene-flow events during history occurred relatively instantaneously, leading to reticulations in tree-topologies, or, conversely, occurred recurrently over longer periods of time leading to un- or weakly-differentiated lineages throughout history [1,21,22]. To do so, for each eight topologies described above (**Figure5**), we simulated gene-flow

either as single-generation gene-flow pulses across pairs of lineages, or as constant recurring gene-flow across pairs of lineages in-between each lineage-divergence event. This design resulted in 16 competing scenarios encompassing eight different possible topologies, contrasting instantaneous gene-flow events with recurring gene-flows. Note that, for simplicity, we did not consider possible gene-flows between KS and RHG recent lineages, as genetic signatures of such recent migrations have never been identified in previous genetic studies nor in historical records to our knowledge.

Importantly, we aimed at determining whether gene-flow occurred symmetrically, or not, across pairs of lineages, in particular in the past. Indeed, previous studies already identified recent asymmetric admixture processes between RHG and RHGn [4,5,8,70], and such possible asymmetries have not been explored among ancient lineages in previous studies [22,26], although they may be influencing ancient tree-topologies or reticulations across lineages and the subsequent past evolution of extant populations.

To do so, for each gene-flow event in each 16 competing scenarios (**Figure5**), we parameterized separately the introgression of lineage A into lineage B, from that of lineage B into lineage A, by drawing randomly the corresponding parameter values independently in the same prior distribution (**TableS8**). ABC posterior-parameter estimations would thus reveal possible asymmetries in gene-flows across pairs of lineages for each event separately, if identifiable from genomic data.

Finally, for each one of the 16 competing combinations of topologies and gene-flow processes, we considered three classes of gene-flow intensities by setting different boundaries of gene-flow parameters' uniform prior-distributions (**TableS8**): no to low possible gene-flow (U[0, 0.001] for recurring processes, and U[0, 0.01] for instantaneous processes), no to moderate gene-flow (U[0,0.0125] for recurring processes, and U[0, 0.25] for instantaneous processes), or no to intense gene-flow events (U[0,0.05] for recurring processes and U[0,1] for instantaneous processes). Note that, for each three classes of gene-flow intensities, each gene-flow parameter from one lineage to another at each point or period in time was drawn independently in the above prior-distributions' boundaries, and thus may differ across events within a scenario, and across scenarios (**Figure5**, **TableS8**).

Altogether, this design resulted in 16x3 = 48 competing scenarios for the complex demographic history of Central and Southern African populations. Importantly, note that for any given class of gene-flow process or intensity, the eight topologies are highly nested in certain parts of the space of parameter values [87]. In particular, despite the fact that we did not consider competing scenarios with explicitly trifurcating ancient tree-topologies, scenarios in which the parameter values for the oldest and second oldest divergence times are similar are expected to provide results highly resembling those obtained with ancient trifurcation scenarios.

All scenario-parameters are represented schematically in **Figure5**, and their prior distributions and constraints are indicated in **TableS8**.

Simulating genomic datasets

We performed simulations under the coalescent using *fastsimcoal2* [116,117], for each one of the 48 scenarios described above, separately. For each simulation, we generated a vector of parameter values randomly drawn in prior-distributions and satisfying the topological constraints as described above and in **TableS8**, with custom-made Python and Bash scripts available in the GitHub repository for this article (https://github.com/Gwennid/africa-wgs-abc). The genetic mutation model is based on the model deployed in [90]. We simulated 100 independent loci (or "chromosomes") with the same structure. Each such "chromosome" corresponds to a linkage block, with the following properties: the type of marker was "DNA", the length of the loci 1 Mb, the recombination rate was 1×10^{-8} per base pair, the mutation rate 1.25×10^{-8} per base pair, and there was no transition bias (transition rate of 0.33).

We performed 5,000 such simulations for each 48 competing scenarios separately in order to conduct Random Forest ABC scenario-choice, hence producing 240,000 separate simulated dataset

each corresponding to a single vector of parameter values randomly drawn in prior distributions (**TableS8**). Then we performed an additional 95,000 separate simulations under the best scenario obtained with RF-ABC for each 54 separate combinations of observed population samples (see below). We thus reached 100,000 simulations under each best scenario identified for each 54 observed datasets respectively, to be used for Neural Network ABC posterior parameter inference.

Building observed genome-wide data-sets for 54 sets of five Central and Southern African populations

We considered 54 separate combinations of four eRHG, wRHG, nKS, sKS, and one eRHGn or wRHGn populations, each with five unrelated individuals (**Table1**). We prepared a call-set of high-quality regions by applying the 1000 Genomes phase 3 accessibility mask and filtering out indels. We pieced high quality windows together to create 1 Mb-long windows. We selected 100 of these regions for which the total length from start to end is less than 1.2 Mb. We then extracted these 100 independent autosomal loci of 1 Mb each from the 25 individuals, hence mimetic of the simulated data.

We checked that these pruned datasets provided individual pairwise genomic differentiation patterns resembling those obtained with the entire SNP set, by computing standardized ASD matrices for each one of the 54 combinations of five populations separately. We compared these ASD matrices with the same individual subsets of the entire ASD matrix computed as above and used for MDS and NJT analyses in **Figure3,** using Spearman ρ and Pearson r correlations tested using 1000 Mantel permutation tests (**TableS10**), with the *partial.mantel.test* function of the *ncf* package in R.

Calculating 337 summary statistics

For each simulated genetic dataset, we computed 337 summary statistics. 41 summary statistics were computed within each five populations separately (hence 205 statistics for five populations), 132 summary statistics were computed across the five populations, for each one of the 54 combinations of five populations separately. In brief, all summary statistics were calculated with plink v1.90b4.9 [110], R v3.6.1 [118], Python v2.7.15, bash, awk[1], and scripts developed by [90], available at https://gitlab.inria.fr/ml_genetics/public/demoseq/-/tree/master, and using the software *asd* (https://github.com/szpiech/asd). The list of calculated summary statistics is provided in **TableS9**. The same statistics were computed for each one of the 54 observed data-sets separately, using the same computational tools and pipeline. Custom-made Python and Bash scripts for computations of all summary-statistics for each simulation and for the observed data are available in the GitHub repository for this article in the "fsc-simulations/code" folder (https://github.com/Gwennid/africa-wgs-abc).

We used the complete set of 337 statistics to perform Random-Forest ABC scenario-choice, as this method is relatively fast and is unaffected by correlations among statistics [45]. We then considered a subset of 202 statistics among the 337 for Neural-Network ABC posterior-parameter estimation, in order to diminish computation time (**TableS9**).

Prior-checking simulations' fit to the observed data

Before conducting ABC scenario-choice and posterior parameter estimation inferences, we checked that summary-statistics calculated on the observed data fell within the range of values of summary-statistics obtained from our simulations. First, we visually verified that each observed vector of summary-statistics computed separately from 54 combinations of five populations each, clustered with the 240,000 vectors of summary-statistics obtained from simulations under the 48 competing scenarios, using two-dimensional PCA calculated with the *prcomp* function R (**FigureS2**). Second, we used the *gfit* function in R to perform goodness-of-fit 100-permutations tests between the 240,000 vectors of

summary-statistics obtained from simulations under the 48 competing scenarios and, in turn, each vector of summary-statistics obtained from the 54 observed datasets respectively.

Results both showed that our simulation design could successfully mimic summary-statistics observed in our five-population sample sets, for each 54 combinations of five populations separately. Hence, ABC inferences could be confidently conducted a priori based on the simulations and observed data here considered.

Random Forest ABC grouped scenario-choice

We conducted series of Random Forest ABC scenario-choice procedures for different groups of scenarios [45,63], elaborated specifically to address our different questions of interest regarding gene-flow processes, their intensities, and topological features of the history of Central and Southern African populations. Random Forest-ABC scenario-choice has proven to be performing efficiently and satisfactorily with a significantly lower number of simulations compared to any other ABC scenario-choice procedure. Moreover, RF-ABC scenario choice is insensitive to correlations among summary-statistics [45,63].

Each RF-ABC scenario-choice procedure presented below was conducted using the *predict.abcrf* and the *abcrf* functions with the "group" option of the *abcrf* package in R [45], with 1000 decision trees to train the algorithm after checking that error rates were appropriately low with the *err.abcrf* function (**FigureS3**), separately for each 54 combinations of five sampled populations.

We conducted cross-validation procedures considering in turn each one of the simulations as pseudo-observed data and all remaining simulations to train the algorithm, for each RF-ABC analysis separately (**FigureS4, FigureS5, FigureS6, FigureS7**). While not necessarily predicting the outcome of the scenario-choice for the observed data, these cross-validation procedures provide us with a sense of the discriminatory power, *a priori*, of RF-ABC for our set of competing-scenarios, as well as empirical levels of nestedness among scenarios or groups of scenarios [87,88]. Finally, we estimated the relative contribution of each one of the 337 summary-statistics to each one of the *a priori* RF-ABC scenario-choice *a priori* procedure, i.e. the relative importance of each statistics to discriminate among competing scenarios for the entire simulation space using out-of-bag cross-validation on all simulations only without observed data, using the "n.var" option in the *abcrf* plotting function. For simplicity and readability, we provide results for the 50 most important such summary-statistics for each RF-ABC scenario-choice procedure performed in **FigureS8** to **FigureS17**.

Instantaneous or recurring gene-flows in the history of Central and Southern African populations?

The 5,000 simulations performed under each one of the 48 competing scenarios were first gathered into two groups of 24 scenarios each (**Figure5**), corresponding to either instantaneous asymmetric gene-flow processes or to recurring constant asymmetric gene-flow between each pair of lineages, respectively. Both groups thus contained all simulations from the eight competing topologies and all three sets of possible gene-flow intensities (low, moderate, or high, see above), and only differed in the process of gene-flow itself. For each 54 combinations of five sampled populations separately, we performed such RF-ABC scenario-choice to determine the best group of scenarios (**Figure6A**).

Low, moderate, or high gene-flow intensities?

For each 54 combinations of five sampled populations separately, we performed RF-ABC scenario-choice in order to determine which class of intensities of gene-flows best explained the data, everything else (topology and gene-flow process) being-equal. We thus considered three groups of scenarios in our RF-ABC scenario-choice, each corresponding to "low", "moderate", or possibly "high" intensities, and

each grouping the 16 scenarios corresponding to eight topologies and instantaneous or recurring gene-flow processes indiscriminately (**Figure6B**).

Then, we conducted, for each 54 population combinations separately, the same RF-ABC scenario-choice procedure to disentangle groups of gene-flow intensities "all topologies being equal", considering here only the 24 scenarios from the best group among instantaneous or recurring gene-flow processes obtained from the above scenario-choice procedure (**FigureS5**).

Finally, we performed, for each 54 combinations of five populations separately, RF-ABC scenario choice for the 48 competing scenarios grouped in six different groups (encompassing eight scenarios each), combining instantaneous or recurring processes with the three classes of intensities respectively (**Figure6C**), thus "intersecting" the two corresponding group analyses (see above).

Which ancestral Central or Southern African lineage diverged first?

The eight topologies considered in the 48 competing scenarios can also be grouped according to their ancient topologies, by considering which ancestral lineage diverged first from all others at the oldest divergence event in the tree (**Figure5**). The ancient lineage which separated first from the two others can either be the lineage ancestral to Northern and Southern Khoe-San populations (AKS, scenario topologies 1a, 1b, 1c in **Figure5**), the lineage ancestral to eastern and western Rainforest Hunter-Gatherer populations (ARHG, scenario topologies 2a, 2b, 2c in **Figure5**), or the lineage leading to Rainforest Hunter-Gatherer Neighboring populations (RHGn, scenario topologies 3a, 3b in **Figure5**).

For each 54 combinations of five populations separately, we conducted RF-ABC scenario-choice across these three groups of scenarios, randomly drawing 2/3 of the 5,000 simulations per scenario for the scenarios 1a, 1b, and 1c, and for the scenarios 2a, 2b, and 2c, respectively, and kept all simulations for the scenarios 3a and 3b, in order to even the total number of simulations in competition among the three groups of scenarios. We thus performed RF-ABC scenario-choice across the three groups of topologies "all gene-flow processes and intensities being equal" (**Figure6D**).

Then we restricted the RF-ABC scenario-choice of the three-competing groups of topologies only for the 24 scenarios from the best instantaneous or recurring gene-flow processes obtained above, all gene-flow intensities being equal (**Figure6E**).

Northern and Southern Khoe-San populations diverged before or after the divergence between Eastern and Western RHG?

The eight topologies considered in the 48 competing scenarios can alternatively be grouped according to the relative order of divergence-time between the divergence event among Northern and Southern KS populations, and that among Eastern and Western RHG populations; an important question to understand the relative duration of separate evolution of each groups of populations.

To address this specific question we thus conducted, for the 54 combinations of five populations separately, RF-ABC scenario-choice procedures by grouping the 48 competing scenarios in two separate groups of 24 scenarios each. One group corresponded to KS populations diverging from one-another before the RHG divergence (scenario topologies 1b, 2b, 3b, and 1c in **Figure5**), and the other group where RHG diverged from one another before the KS divergence (scenario topologies 1a, 2a, 3a, and 2c in **Figure5**), instantaneous or recurring gene-flow and all "low", "moderate", or "high" gene-flow intensities being equal among the two groups (**Figure6F**). Last, we conducted the same RF-ABC scenario-choice procedure between two groups of four topologies each, but restricted to the 24 best scenarios among the instantaneous or recurring gene-flow processes as determined from above procedures (**Figure6G**).

Finally, we conducted RF-ABC scenario choice procedures considering, respectively, all 48 competing scenarios separately (**FigureS6**), and all 24 best scenarios among the instantaneous or recurring gene-flow processes as determined above (**FigureS7**).

Neural Network ABC posterior-parameter estimation

We intersected all the RF-ABC scenario-choice results considering the different groups of scenarios or all scenarios independently, as described above, in order to determine, which single scenario was explaining data most accurately in the majority of the 54 combinations of five Central and Southern African populations. We found that the Scenario i1-1b conservatively produced simulations whose observed summary-statistics most resembled those obtained in 25 out of the 54 possible population combinations here tested (see **Results**, the detailed list of 25 population combinations is provided in **TableS11**). The second most often best-scenario, Scenario i1-3b, only succeeded for four population combinations among the 54 tested.

Based on this *a posteriori* Scenario i1-1b, we performed an additional 95,000 simulations considering the same prior distributions and constraints among parameters, in order to obtain 100,000 simulations for further ABC posterior parameter inferences, separately for each one of the 25 population combinations for which it was identified confidently as the winner, separately. As it remains difficult to estimate jointly the posterior distribution of all parameters of the complex scenarios here explored with RF-ABC [62], we instead conducted Neural Network ABC joint posterior parameter inferences using the *abc* package in R [43,44], following best-practices previously bench-marked for highly complex demographic scenarios [88,90]. There are no rules of thumb in order to determine a priori the best number of neurons and tolerance rate to be set for training the neural network in NN-ABC [90].

There are no evident criteria to choose *a priori* the best tolerance level and numbers of neurons in the neural network's hidden layer for parameterizing the NN-ABC posterior-parameter estimation procedure [90,93]. As the total number of parameters in the best Scenario i1-1b was large (43, **Figure5** and **TableS8**), and as the number of summary-statistics considered was also large (202, **TableS9**), we chose to conduct the 25 NN-ABC posterior-parameter inferences for the 25 combinations of sampled populations considering, in-turn, 7, 14, 21, 28, 35, 42, or 43 neurons in the hidden layer (43 being the assumed number of dimensions equal to the number of parameters for this scenario [43,90]), and a tolerance level of 0.01, thus considering the 1,000 simulations closest to each observed dataset out of the 100,000 performed, respectively. We found *a posteriori* that, considering 42 neurons in the hidden layer provided overall posterior parameter distributions departing from their priors, and in particular for divergence-times parameters which we were highly interested in, and therefore decided to provide all posterior parameter distributions' results using this parameterization (tolerance = 0.01, number of hidden neurons = 42) for the training of each one of the neural networks.

We thus performed 25 separate NN-ABC joint parameter posterior inferences, using the "neuralnet" method option in the function *abc* of the *abc* package in R [44], with logit-transformed ("logit" transformation option), an "epanechnikov" kernel, parameter values within parameter-priors' boundaries, a tolerance level of 0.01, and 42 neurons in the hidden layer. Adjusted posterior parameter distributions obtained with this method were then plotted for each parameter separately using a gaussian kernel truncated at the boundaries of the parameter prior-distributions. We then estimated the mode, median, mean, and 50% and 95% Credibility Intervals of these distributions *a posteriori*. We provide all posterior-parameter distributions together with their priors in figure-format (**Figure7**, **Figure8**, **FigureS18, FigureS19**), and in table-format (**TableS14**, **TableS12**).

Neural Network ABC posterior-parameter estimation cross-validation errors

For each one of the 25 combinations of sampled population separately, we randomly chose 50 simulations (5%) out of the 1000 simulations closest to each observed dataset as determined above. We then conducted NN-ABC joint posterior parameter inference with the same hyper-parameters as above and considering, for each chosen simulation considered as pseudo-observed, the other 99,999

simulations in the reference table, for each one of these 50 pseudo-observed simulations and for each 25 combinations of sampled populations separately (1250 separate NN-ABC joint posterior parameter inferences in total). We then calculated, for each 43 parameter and each 25 populations' combination, the cross-validation posterior parameter estimation error as the mean-squared error of the median NNB-ABC posterior estimate $(\hat{\theta}_i)$ scaled by the true-parameter's (θ_i) variance (**TableS13**), as $\sum_1^{50}(\hat{\theta}_i - \theta_i)^2 / (50 \times Variance(\theta_i))$ [44].

REFERENCES

1. Henn, B. M., Steele, T. E. & Weaver, T. D. Clarifying distinct models of modern human origins in Africa. *Current Opinion in Genetics & Development* **53**, 148–156 (2018).
2. Schlebusch, C. M. & Jakobsson, M. Tales of Human Migration, Admixture, and Selection in Africa. *Annu. Rev. Genom. Hum. Genet.* **19**, 405–428 (2018).
3. Pfennig, A., Petersen, L. N., Kachambwa, P. & Lachance, J. Evolutionary Genetics and Admixture in African Populations. *Genome Biology and Evolution* **15**, evad054 (2023).
4. Verdu, P. *et al.* Origins and Genetic Diversity of Pygmy Hunter-Gatherers from Western Central Africa. *Current Biology* **19**, 312–318 (2009).
5. Verdu, P. *et al.* Sociocultural Behavior, Sex-Biased Admixture, and Effective Population Sizes in Central African Pygmies and Non-Pygmies. *Molecular Biology and Evolution* **30**, 918–937 (2013).
6. Schlebusch, C. M. *et al.* Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science* **338**, 374–379 (2012).
7. Breton, G. *et al.* Lactase Persistence Alleles Reveal Partial East African Ancestry of Southern African Khoe Pastoralists. *Current Biology* **24**, 852–858 (2014).
8. Patin, E. *et al.* The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nat Commun* **5**, 3163 (2014).
9. Patin, E. *et al.* Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**, 543–546 (2017).
10. Perry, G. H. *et al.* Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proc. Natl. Acad. Sci. U.S.A.* **111**, (2014).
11. Busby, G. B. *et al.* Admixture into and within sub-Saharan Africa. *eLife* **5**, e15266 (2016).
12. Pierron, D. *et al.* Genomic landscape of human diversity across Madagascar. *Proc. Natl. Acad. Sci. U.S.A.* **114**, (2017).
13. Semo, A. *et al.* Along the Indian Ocean Coast: Genomic Variation in Mozambique Provides New Insights into the Bantu Expansion. *Molecular Biology and Evolution* **37**, 406–416 (2020).
14. Lucas-Sánchez, M., Serradell, J. M. & Comas, D. Population history of North Africa based on modern and ancient genomes. *Human Molecular Genetics* **30**, R17–R23 (2021).
15. Sengupta, D. *et al.* Genetic substructure and complex demographic history of South African Bantu speakers. *Nat Commun* **12**, 2080 (2021).
16. Fortes-Lima, C. *et al.* Demographic and Selection Histories of Populations Across the Sahel/Savannah Belt. *Molecular Biology and Evolution* **39**, msac209 (2022).
17. Fortes-Lima, C. A. *et al.* The genetic legacy of the expansion of Bantu-speaking peoples in Africa. *Nature* **625**, 540–547 (2024).
18. Laurent, R. *et al.* A genetic and linguistic analysis of the admixture histories of the islands of Cabo Verde. *eLife* **12**, e79827 (2023).
19. Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E. M. L. & Skoglund, P. Origins of modern human ancestry. *Nature* **590**, 229–237 (2021).
20. Stringer, C. Modern human origins: progress and prospects. *Phil. Trans. R. Soc. Lond. B* **357**, 563–579 (2002).
21. Hollfelder, N., Breton, G., Sjödin, P. & Jakobsson, M. The deep population history in Africa. *Human Molecular Genetics* **30**, R2–R10 (2021).
22. Ragsdale, A. P. *et al.* A weakly structured stem for human origins in Africa. *Nature* **617**, 755–763 (2023).
23. Lachance, J. *et al.* Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. *Cell* **150**, 457–469 (2012).
24. Lorente-Galdos, B. *et al.* Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. *Genome Biol* **20**, 77 (2019).

25. Lipson, M. *et al.* Ancient DNA and deep population structure in sub-Saharan African foragers. *Nature* **603**, 290–296 (2022).
26. Fan, S. *et al.* Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* **186**, 923-939.e14 (2023).
27. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**, 222–226 (2012).
28. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
29. Tishkoff, S. A. *et al.* The Genetic Structure and History of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
30. Skoglund, P. & Mathieson, I. Ancient Genomics of Modern Humans: The First Decade. *Annu. Rev. Genom. Hum. Genet.* **19**, 381–404 (2018).
31. Gravel, S. Population Genetics Models of Local Ancestry. *Genetics* **191**, 607–619 (2012).
32. Ragsdale, A. P. & Gravel, S. Models of archaic admixture and recent history from two-locus statistics. *PLoS Genet* **15**, e1008204 (2019).
33. Kamm, J., Terhorst, J., Durbin, R. & Song, Y. S. Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. *Journal of the American Statistical Association* **115**, 1472–1487 (2020).
34. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
35. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
36. Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014).
37. Choudhury, A. *et al.* Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat Commun* **8**, 2062 (2017).
38. *Hunter-Gatherers of the Congo Basin: Cultures, Histories and Biology of African Pygmies*. (Routledge, Taylor and Francis Group, London New York, 2017).
39. Li, J. Z. *et al.* Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* **319**, 1100–1104 (2008).
40. Schlebusch, C. M. *et al.* Khoe-San Genomes Reveal Unique Variation and Confirm the Deepest Population Divergence in Homo sapiens. *Molecular Biology and Evolution* **37**, 2944–2954 (2020).
41. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
42. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
43. Blum, M. G. B. & François, O. Non-linear regression models for Approximate Bayesian Computation. *Stat Comput* **20**, 63–73 (2010).
44. Csilléry, K., François, O. & Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC): *R package: abc*. *Methods in Ecology and Evolution* **3**, 475–479 (2012).
45. Pudlo, P. *et al.* Reliable ABC model choice via random forests. *Bioinformatics* **32**, 859–866 (2016).
46. Breton, G., Fortes-Lima, C. & Schlebusch, C. M. Revisiting the demographic history of Central African populations from a genetic perspective. *Hum Popul Genet Genom* 1–29 (2021) doi:10.47248/hpgg2101010004.
47. Nei, M. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583–590 (1978).
48. Pemberton, T. J. *et al.* Genomic Patterns of Homozygosity in Worldwide Human Populations. *The American Journal of Human Genetics* **91**, 275–292 (2012).
49. Mooney, J. A. *et al.* Understanding the Hidden Complexity of Latin American Population Isolates. *The American Journal of Human Genetics* **103**, 707–726 (2018).

50. Szpiech, Z. A. *et al.* Ancestry-Dependent Enrichment of Deleterious Homozygotes in Runs of Homozygosity. *The American Journal of Human Genetics* **105**, 747–762 (2019).
51. Bowcock, A. M. *et al.* Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proceedings of the National Academy of Sciences* **88**, 839–843 (1991).
52. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* (1987) doi:10.1093/oxfordjournals.molbev.a040454.
53. Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14**, 685–695 (1997).
54. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664 (2009).
55. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
56. Rosenberg, N. A. Genetic Structure of Human Populations. *Science* **298**, 2381–2385 (2002).
57. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
58. Lawson, D. J., van Dorp, L. & Falush, D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat Commun* **9**, 3258 (2018).
59. Peter, B. M. A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis. *Phil. Trans. R. Soc. B* **377**, 20200413 (2022).
60. Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P. & Ramachandran, S. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32**, 2817–2823 (2016).
61. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**, 1791–1798 (1999).
62. Raynal, L. *et al.* ABC random forests for Bayesian parameter inference. *Bioinformatics* **35**, 1720–1728 (2019).
63. Estoup, A., Raynal, L., Verdu, P. & Marin, J.-M. Model choice using Approximate Bayesian Computation and Random Forests: analyses based on model grouping to make inferences about the genetic history of Pygmy human populations. *Journal de la Société Française de Statistique* 167–190 (2018).
64. Seidensticker, D. *et al.* Population collapse in Congo rainforest from 400 CE urges reassessment of the Bantu Expansion. *Sci. Adv.* **7**, eabd8352 (2021).
65. Verdu, P. & Rosenberg, N. A. A General Mechanistic Model for Admixture Histories of Hybrid Populations. *Genetics* **189**, 1413–1426 (2011).
66. Harris, K. & Nielsen, R. The Genetic Cost of Neanderthal Introgression. *Genetics* **203**, 881–891 (2016).
67. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S. & Chikhi, L. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity* **116**, 362–371 (2016).
68. Hamid, I., Korunes, K. L., Beleza, S. & Goldberg, A. Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde. *eLife* **10**, e63177 (2021).
69. Cuadros-Espinoza, S., Laval, G., Quintana-Murci, L. & Patin, E. The genomic signatures of natural selection in admixed human populations. *The American Journal of Human Genetics* **109**, 710–726 (2022).
70. Patin, E. *et al.* Inferring the Demographic History of African Farmers and Pygmy Hunter–Gatherers Using a Multilocus Resequencing Data Set. *PLoS Genet* **5**, e1000448 (2009).

71. Beyer, R. M., Krapp, M., Eriksson, A. & Manica, A. Climatic windows for human migration out of Africa in the past 300,000 years. *Nat Commun* **12**, 4889 (2021).
72. Lopez, M. *et al.* The demographic history and mutational load of African hunter-gatherers and farmers. *Nat Ecol Evol* **2**, 721–730 (2018).
73. Bartlein, P. J. *et al.* Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis. *Clim Dyn* **37**, 775–802 (2011).
74. Perry, G. H. & Verdu, P. Genomic perspectives on the history and evolutionary ecology of tropical rainforest occupation by humans. *Quaternary International* **448**, 150–157 (2017).
75. Ziegler, M. *et al.* Development of Middle Stone Age innovation linked to rapid climate change. *Nat Commun* **4**, 1905 (2013).
76. Scerri, E. M. L. *et al.* Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? *Trends in Ecology & Evolution* **33**, 582–594 (2018).
77. Gosling, W. D., Scerri, E. M. L. & Kaboth-Bahr, S. The climate and vegetation backdrop to hominin evolution in Africa. *Phil. Trans. R. Soc. B* **377**, 20200483 (2022).
78. Thomas, D. S. G. *et al.* Lacustrine geoarchaeology in the central Kalahari: Implications for Middle Stone Age behaviour and adaptation in dryland conditions. *Quaternary Science Reviews* **297**, 107826 (2022).
79. Long, J. C. The genetic structure of admixed populations. *Genetics* **127**, 417–428 (1991).
80. Cornelissen, E. [No title found]. *Journal of World Prehistory* **16**, 197–235 (2002).
81. Mesfin, I., Oslisly, R. & Forestier, H. Technological analysis of the quartz industry of Maboué 5 – Layer 3 (Lopé National Park, Gabon): Implications for the Late Stone Age emergence in western Central Africa. *Journal of Archaeological Science: Reports* **39**, 103130 (2021).
82. Bader, G. D., Mabuza, A., Price Williams, D. & Will, M. Rethinking the Middle to Later Stone Age transition in southern Africa - A perspective from the highveld of Eswatini. *Quaternary Science Reviews* **286**, 107540 (2022).
83. Lézine, A.-M., Izumi, K., Kageyama, M. & Achoundong, G. A 90,000-year record of Afromontane forest responses to climate change. *Science* **363**, 177–181 (2019).
84. Phillipson, D. W. *African Archaeology*. (Cambridge University Press, 2005). doi:10.1017/CBO9780511800313.
85. Bostoen, K. *et al.* Middle to Late Holocene Paleoclimatic Change and the Early Bantu Expansion in the Rain Forests of Western Central Africa. *Current Anthropology* **56**, 354–384 (2015).
86. Lopez, M. *et al.* Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest. *Current Biology* **29**, 2926-2935.e4 (2019).
87. Robert, C. P., Mengersen, K. & Chen, C. Model choice versus model criticism. *Proceedings of the National Academy of Sciences* **107**, E5–E5 (2010).
88. Fortes-Lima, C. A., Laurent, R., Thouzeau, V., Toupance, B. & Verdu, P. Complex genetic admixture histories reconstructed with Approximate Bayesian Computation. *Mol Ecol Resour* **21**, 1098–1117 (2021).
89. Boitard, S., Rodríguez, W., Jay, F., Mona, S. & Austerlitz, F. Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. *PLoS Genet* **12**, e1005877 (2016).
90. Jay, F., Boitard, S. & Austerlitz, F. An ABC Method for Whole-Genome Sequence Data: Inferring Paleolithic and Neolithic Human Expansions. *Molecular Biology and Evolution* **36**, 1565–1579 (2019).
91. Wang, H., Czerminski, R. & Jamieson, A. C. Neural Networks and Deep Learning. in *The Machine Age of Customer Insight* (eds. Einhorn, M., Löffler, M., De Bellis, E., Herrmann, A. & Burghartz, P.) 91–101 (Emerald Publishing Limited, 2021). doi:10.1108/978-1-83909-694-520211010.
92. Leung, F. H. F., Lam, H. K., Ling, S. H. & Tam, P. K. S. Tuning of the structure and parameters of a neural network using an improved genetic algorithm. *IEEE Trans. Neural Netw.* **14**, 79–88 (2003).

93. Huang, X., Rymbekova, A., Dolgova, O., Lao, O. & Kuhlwilm, M. Harnessing deep learning for population genetic inference. *Nat Rev Genet* **25**, 61–78 (2024).
94. Mooney, J. A., Agranat-Tamir, L., Pritchard, J. K. & Rosenberg, N. A. On the number of genealogical ancestors tracing to the source groups of an admixed population. *GENETICS* **224**, iyad079 (2023).
95. Agranat-Tamir, L., Mooney, J. A. & Rosenberg, N. A. Counting the genetic ancestors from source populations in members of an admixed population. *GENETICS* **226**, iyae011 (2024).
96. Murtagh, F. Multilayer perceptrons for classification and regression. *Neurocomputing* **2**, 183–197 (1991).
97. Chen, J., Zhao, F., Sun, Y. & Yin, Y. Improved XGBoost model based on genetic algorithm. *IJCAT* **62**, 240 (2020).
98. Yelmen, B. & Jay, F. An Overview of Deep Generative Models in Functional and Evolutionary Genomics. *Annu. Rev. Biomed. Data Sci.* **6**, 173–189 (2023).
99. Skoglund, P. *et al.* Reconstructing Prehistoric African Population Structure. *Cell* **171**, 59-71.e21 (2017).
100. Vicente, M. & Schlebusch, C. M. African population history: an ancient DNA perspective. *Current Opinion in Genetics & Development* **62**, 8–15 (2020).
101. Schlebusch, C. Issues raised by use of ethnic-group names in genome study. *Nature* **464**, 487–487 (2010).
102. Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucl Acids Res* **16**, 1215–1215 (1988).
103. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
104. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491–498 (2011).
105. Van Der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *CP in Bioinformatics* **43**, (2013).
106. Breton, G., Johansson, A. C. V., Sjödin, P., Schlebusch, C. M. & Jakobsson, M. Comparison of sequencing data processing pipelines and application to underrepresented African human populations. *BMC Bioinformatics* **22**, 488 (2021).
107. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
108. Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.* **21**, 830–839 (2011).
109. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
110. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).
111. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
112. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
113. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
114. Hublin, J.-J. *et al.* New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature* **546**, 289–292 (2017).
115. Richter, D. *et al.* The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature* **546**, 293–296 (2017).
116. Excoffier, L. & Foll, M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332–1334 (2011).

117. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet* **9**, e1003905 (2013).
118. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2015).

Acknowledgments

We are grateful to all subjects who participated in this research. The computations were performed at the Swedish National Infrastructure for Computing (SNIC-UPPMAX). Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. We thank the platform Paléogénomique et Génétique Moléculaire (P2GM) of the MNHN at the Musée de l'Homme in Paris for helping to prepare, in part, the Central African samples for sequencing. We thank the Working Group of Indigenous Minorities in Southern Africa (WIMSA) and the South African San Council for their support and facilitating fieldwork. The project was approved by the South African San Council.

Funding

GB was funded in part by a grant from the Wilhelm och Martina Lundgrens Vetenskapsfond (2023-SA-4327). CS is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 759933) and the Knut and Alice Wallenberg foundation. MJ is funded by the Knut and Alice Wallenberg foundation, and the Vetenskapsrådet/Swedish Research Council (no: 642-2013-8019 and 2018-05537). PV was funded in part by the UMR7206 Eco-anthropology and the French Agence Nationale de la Recherche ANR-METHIS (15-CE32-0009-1).

Authors' contributions

GB – Designed the study, conducted data curation and quality controls, conducted statistical descriptions, conducted simulations and ABC inferences, interpreted the results, wrote the original draft.

PS – Designed the study, conducted statistical descriptions, interpreted the results, participated in writing and editing.

PLZ – Conducted data curation and quality controls, conducted statistical descriptions, participated in editing.

RL –Conducted simulations and ABC inferences, participated in writing and editing.

AF – Conducted fieldwork and provided samples, participated in editing.

AES – Conducted fieldwork and provided samples, participated in editing.

BSH – Conducted fieldwork and provided samples, participated in editing.

LBB – Conducted fieldwork and provided samples, participated in editing.

GHP – Conducted fieldwork and provided samples, participated in editing.

HS – Conducted fieldwork and provided samples, participated in editing.

EH – Acquired funding, interpreted the results, participated in editing.

CMS – Supervised the project, designed the study, acquired funding, conducted fieldwork and provided samples, interpreted the results, participated in writing and editing.

MJ – Supervised the project, designed the study, acquired funding, interpreted the results, participated in writing and editing.

PV – Supervised the project, designed the study, acquired funding, conducted fieldwork and provided samples, conducted statistical descriptions, conducted simulations and ABC inferences, interpreted the results, wrote the original draft.

Competing interests

The authors declare no competing nor conflict of interests.

GitHub repositories for material availability

https://github.com/Gwennid/africa-wgs-descriptive
https://github.com/Gwennid/africa-wgs-abc

Data availability statement

All raw whole-genome sequence FASTQ data original to this study are made available on the European Genome-Phenome Archive (https://ega-archive.org/), provided that requests conform to ethical requirements and informed consents provided by donors, as detailed in the corresponding Data Access Policies associated with each data repository. Data access requests are evaluated by the associated Data Access Committees. **EGA ACCESSION NUMBERS PENDING ACCEPTANCE**.

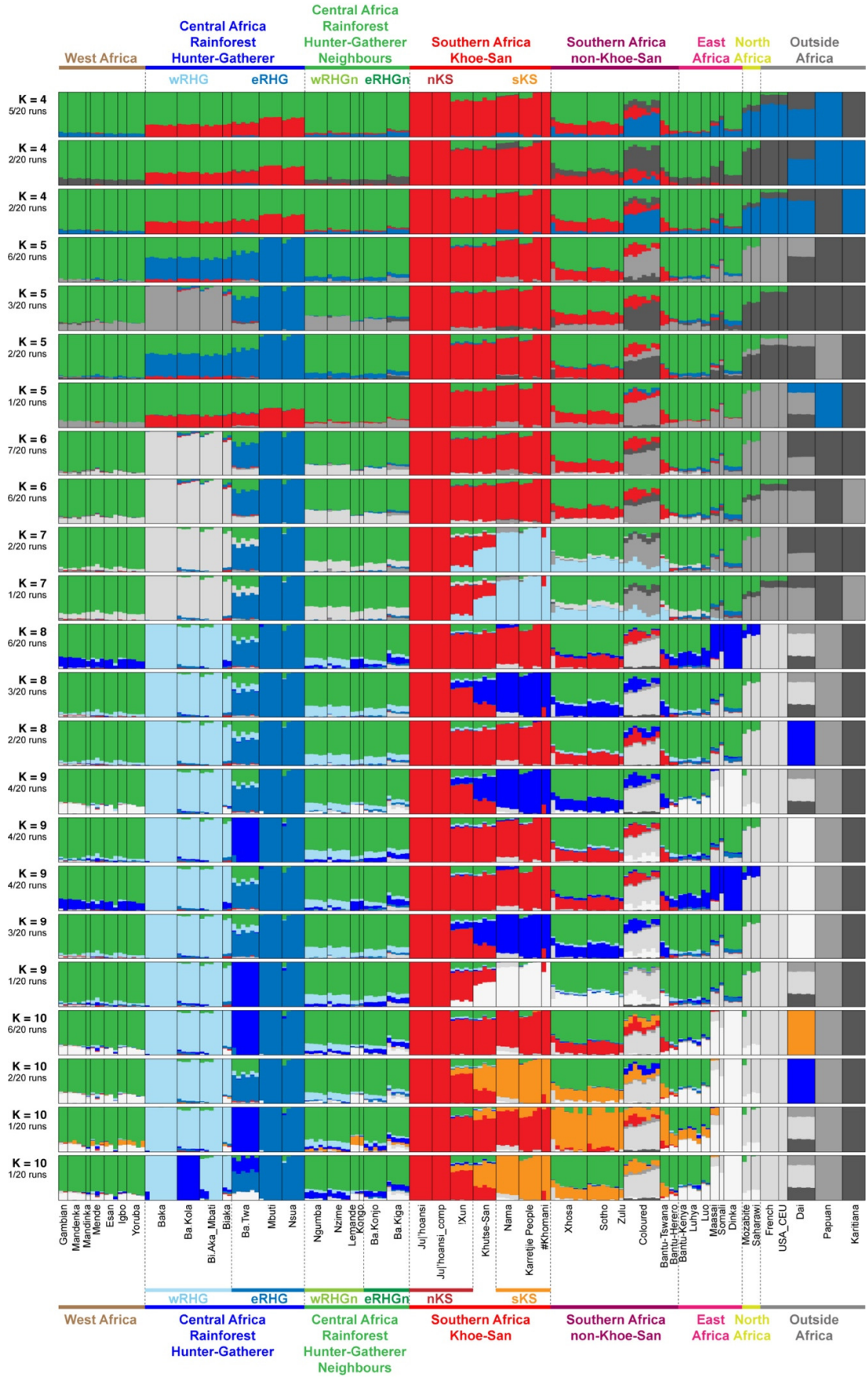
Ethical statement

DNA samples from individuals were collected with the subjects' informed consent following the Declaration of Helsinki guidelines and approved by the following Ethics committees: Institut de Recherche pour le Développement (IRD, France); Cameroon Ministry of Scientific Research and Technology; Central African Republic Ministère de l'Enseignement Supérieur et de la Recherche Scientifique; France Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (Comité de Protection des Personnes DC-2016-2740); French Commission Nationale Informatique et Liberté CNIL (declaration n°1972648); Uganda National Council for Science and Technology; Institutional Review Board of the University of Chicago Biological Sciences Division and the Pritzker School of Medicine University of Chicago Hospitals (protocol number 16986A); Committee for the Protection of Human Subjects of the Trustees of Dartmouth College and the Dartmouth-Hitchcock Medical Center (protocol number 22410); Human Research Ethics Committee (Medical) at the University of the Witwatersrand, Johannesburg (Protocol Numbers: M980553, M050902, M090576, M10270, M180654); Working Group of Indigenous Minorities in Southern Africa (WIMSA); South African San Council (SASC); and the Swedish Ethical Review Authority (Dnr 2019-05174).

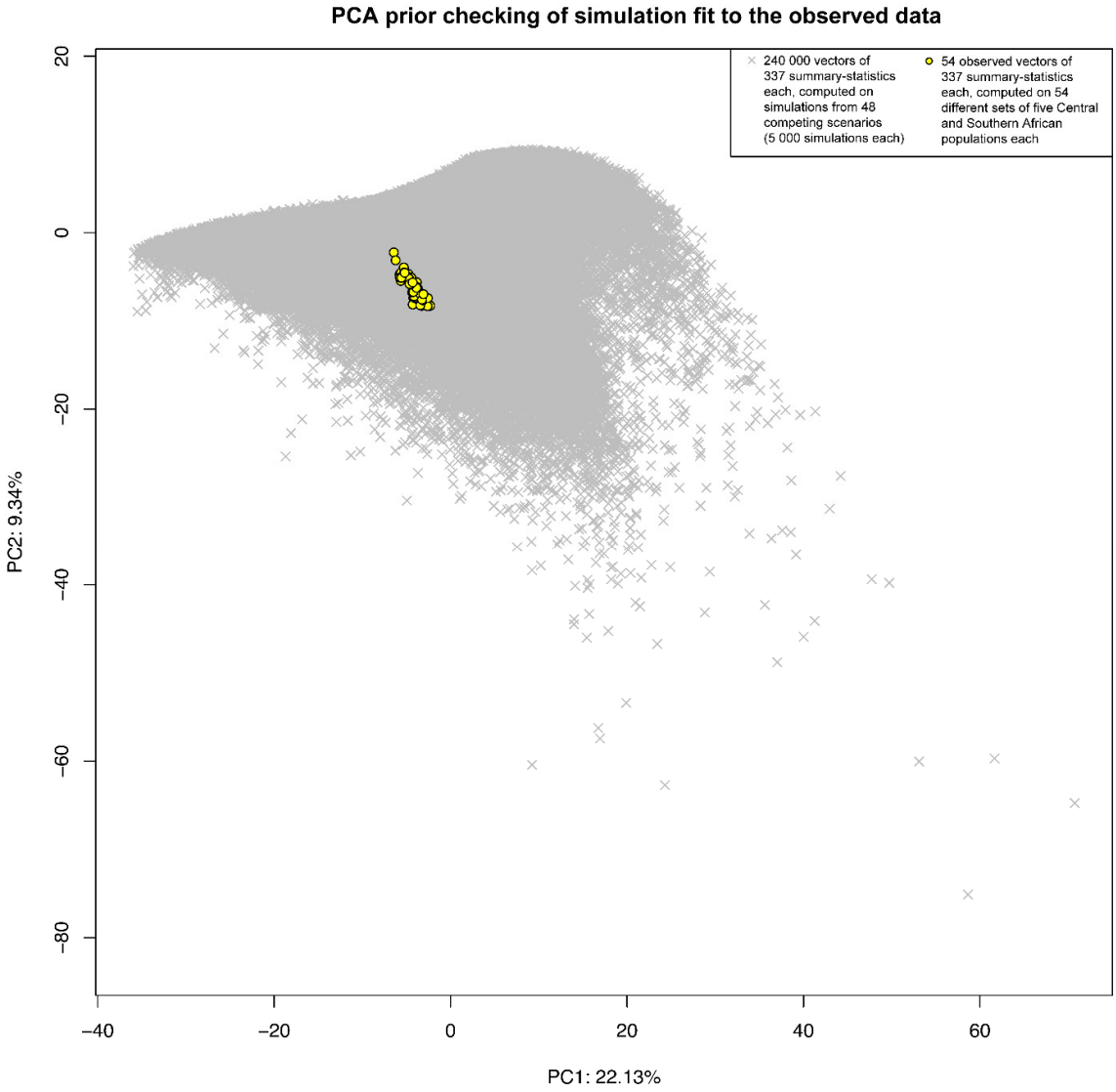
Supplementary Material

20 Supplementary Figures and 14 Supplementary Tables

<u>FigureS1</u>: Alternative ADMIXTURE modal results for K=4 to 10.
Modal results each containing the highest number of runs among the 20 independent runs performed for each value of K are provided in main **Figure4**. All alternative modal results are provided here. See **Material and Methods** for details.

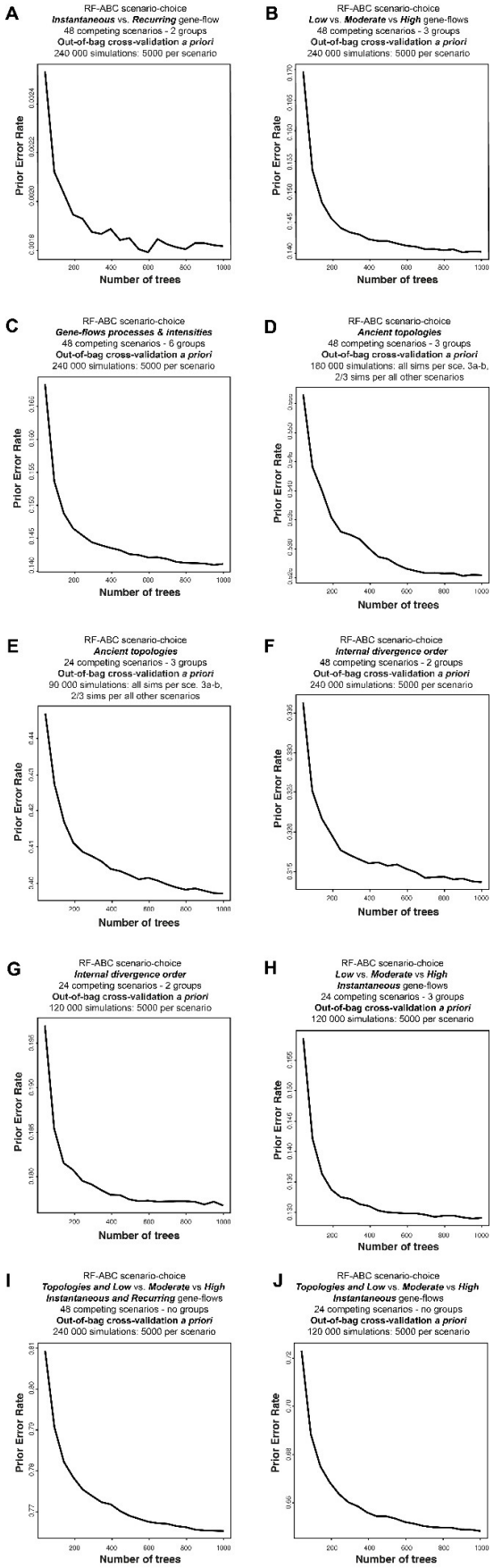


<u>FigureS2</u>: PCA for ABC prior-checking of simulations fit to the observed data.
We calculated 337 summary-statistics for each one of the 240 000 simulated data sets, 5000 under each one of the 48 competing scenarios (**Figure5** and **Material and Methods**), which we projected on the first two axes of a principal component analysis. Each vector (corresponding thus to a single simulation) is represented by a gray cross. We separately computed the same 337 summary-statistics, separately for each one of the 54 sets of five observed Central and Southern African populations included in our analyses. Each observed vector is then projected, in turn, on the PCA obtained from simulations only. Each such observed vector is represented as a single yellow dot. We can see that all the 54 observed sets of populations fall well within the space of simulated dataset, thus allowing us a priori to conduct machine-learning ABC scenario choice and posterior parameter estimation procedures.

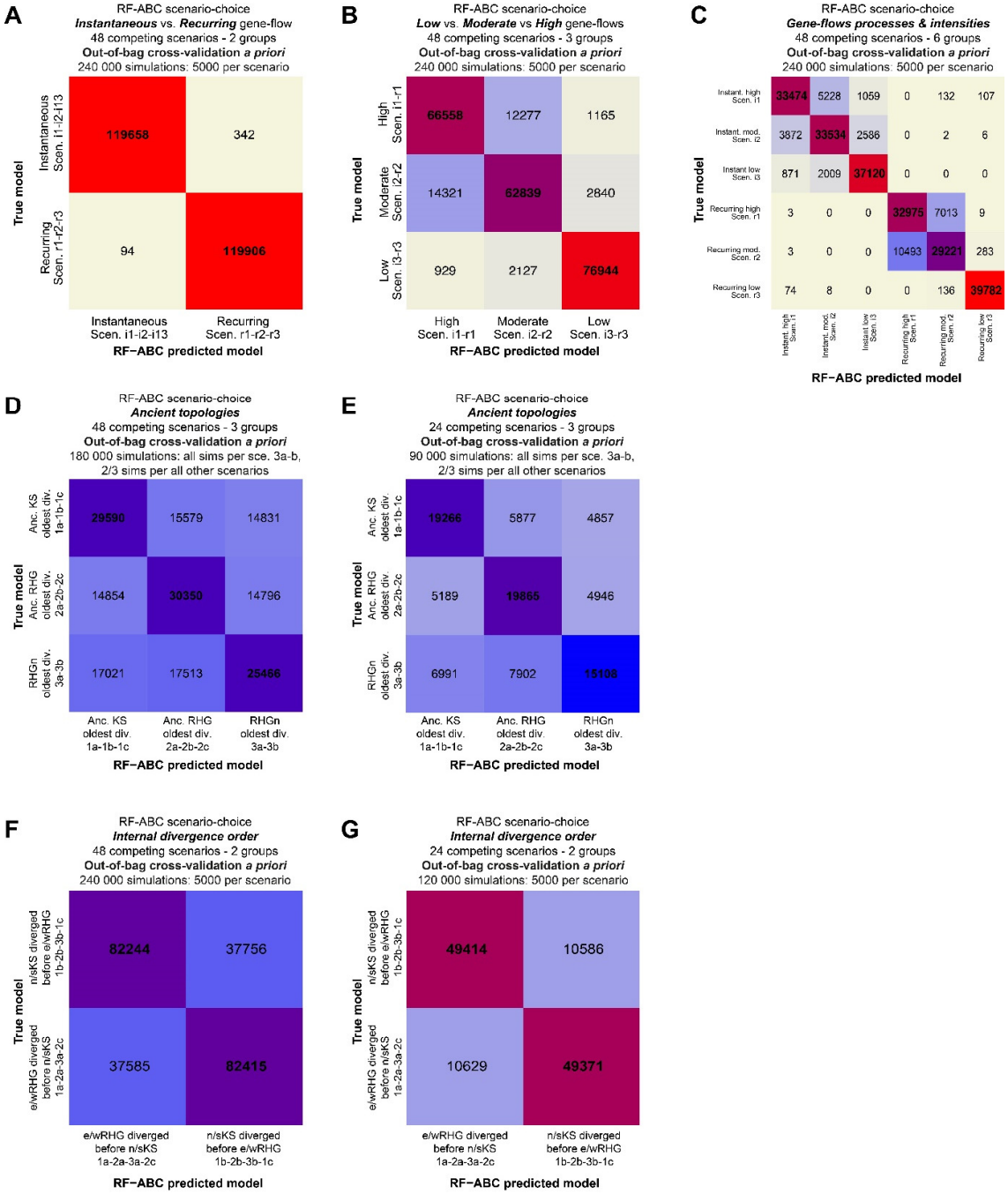


<u>FigureS3</u>: Cross-validation prior error of Random-Forest ABC scenario-choice procedures as a function of the number of trees in the forest.

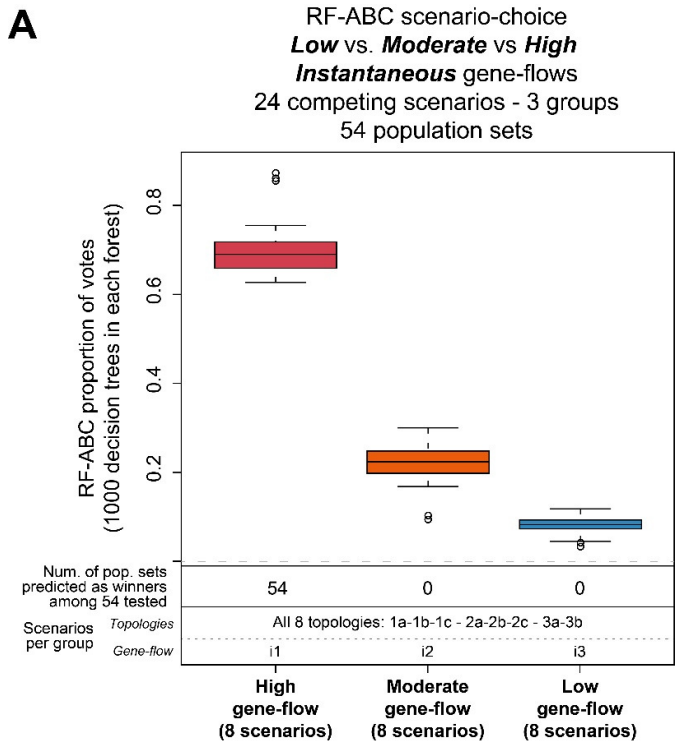
For panels A to G, out-of-bag cross-validation prior error results for each RF-ABC scenario choice analysis are shown in **FigureS4 panels A to G** respectively**.** For panel H, the corresponding out-of-bag cross-validation prior error results are shown in **FigureS5 panel B.** For panel I, the corresponding out-of-bag cross-validation prior error results are shown in **FigureS6 panel B.** For panel J, the corresponding out-of-bag cross-validation prior error results are shown in **FigureS7 panel B.**

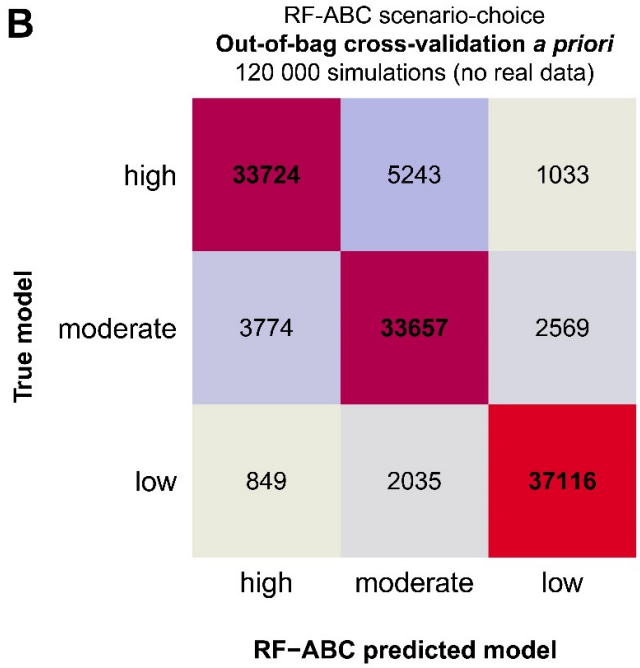


<u>FigureS4</u>: Out-of-bag cross-validation errors for Random-Forest ABC scenario-choices among 48 competing scenarios. All RF-ABC cross-validation analyses are conducted *a priori*, without using any observed data, by considering, in-turn, each simulation as observed data while all remaining simulations are used to train the Random Forest. Each panel provides the *a priori* cross-validation errors for the corresponding RF-ABC scenario-choice analysis detailed in each panel in **Figure6**. While these cross-validation errors do not predict the specific outcome of RF-ABC prediction for the observed data, they provide levels of scenarios and groups of scenarios nestedness *a priori* for the entire space of parameters used for simulations. (**A**) shows that intermediate and recurring gene-flow processes are, all topologies and gene-flow intensities "being equal", clearly differentiable by RF-ABC and little nested. (**B**) and (**C**) show that different classes of gene-flow intensities are also relatively clearly identifiable, albeit more nested, as expected by design since the "high gene-flow" class comprises parameter values of the moderate and the low intensity classes of gene-flows (**Material and Methods**, **TableS8**). (**D**) and (**E**) show that ancient tree topologies are relatively well distinguishable *a priori* all gene-flow processes "being equal", despite some amounts of errors due to the expected nestedness in the space of parameter values where ancient divergence times are resembling and thus undistinguishable. (**F**) and (**G**) show that the chronological order between the divergence-time of Northern and Southern KS lineages, and that of Western and Eastern RHG lineages, are also relatively well distinguishable a priori with RF-ABC scenario-choice, albeit with an increased cross-validation error expected due to high levels of nestedness in the spaces of parameter-values where both separate events may occur at similar time.



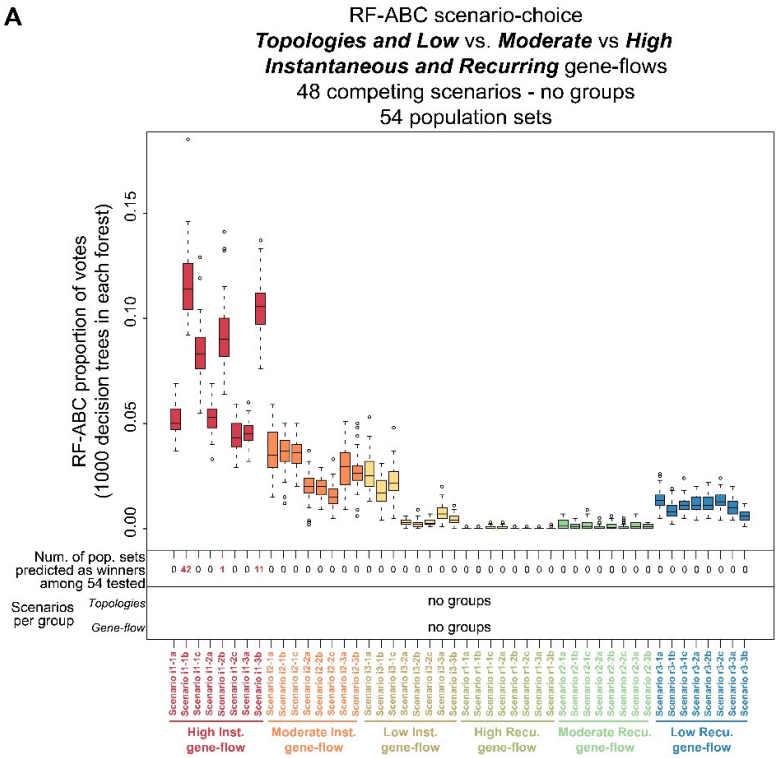
2035 **<u>FigureS5</u>: RF-ABC scenario choice among 24 competing scenarios with instantaneous asymmetric gene-flow processes,**
2036 **for three groups of gene-flow intensities, all topologies being equal.**
2037 (**A**) corresponds to the scenario-choice results for the same test as in **Figure6C**, restricted to the 24 competing scenarios
2038 considering instantaneous gene-flow processes only, all scenario topologies "being equal". Description of the x and y axis
2039 legends are provided in **Figure6**. Posterior probabilities of choosing the correct scenarios for each 54 sets of population
2040 combinations ranged from 0.63935 to 0.87000, compared to a prior probability of 1/3 in this particular test. (**B**) corresponds
2041 to the RF-ABC cross-validation prior error for the scenario-choice procedure conducted in (**A**). Errors were obtained without
2042 considering observed data and using, instead, all simulations in turn as pseudo-observed data and the remaining simulations
2043 to train the RF.
2044

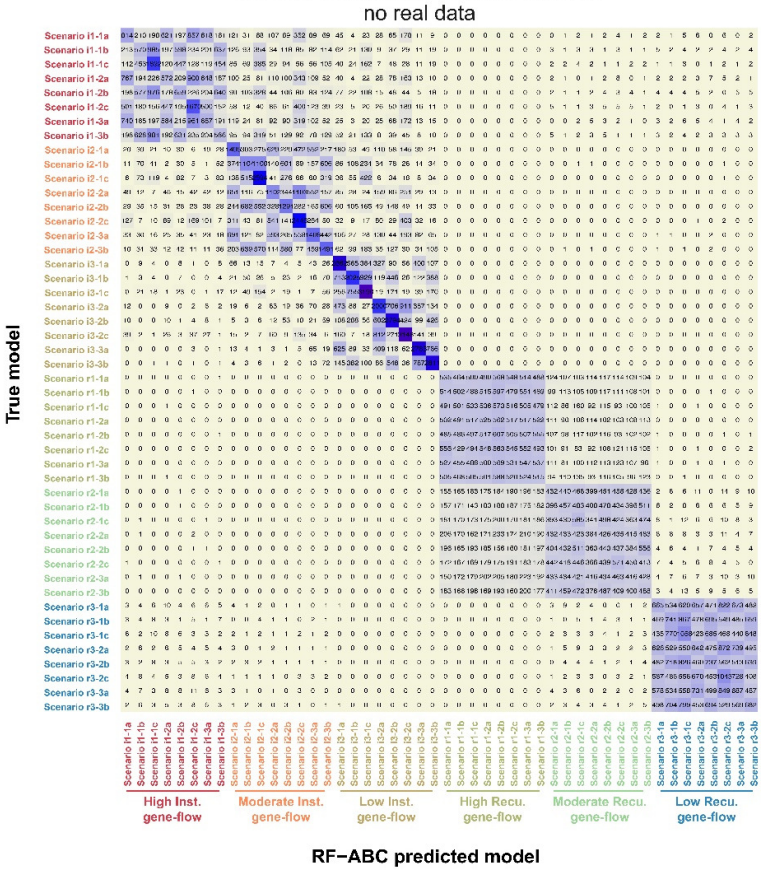




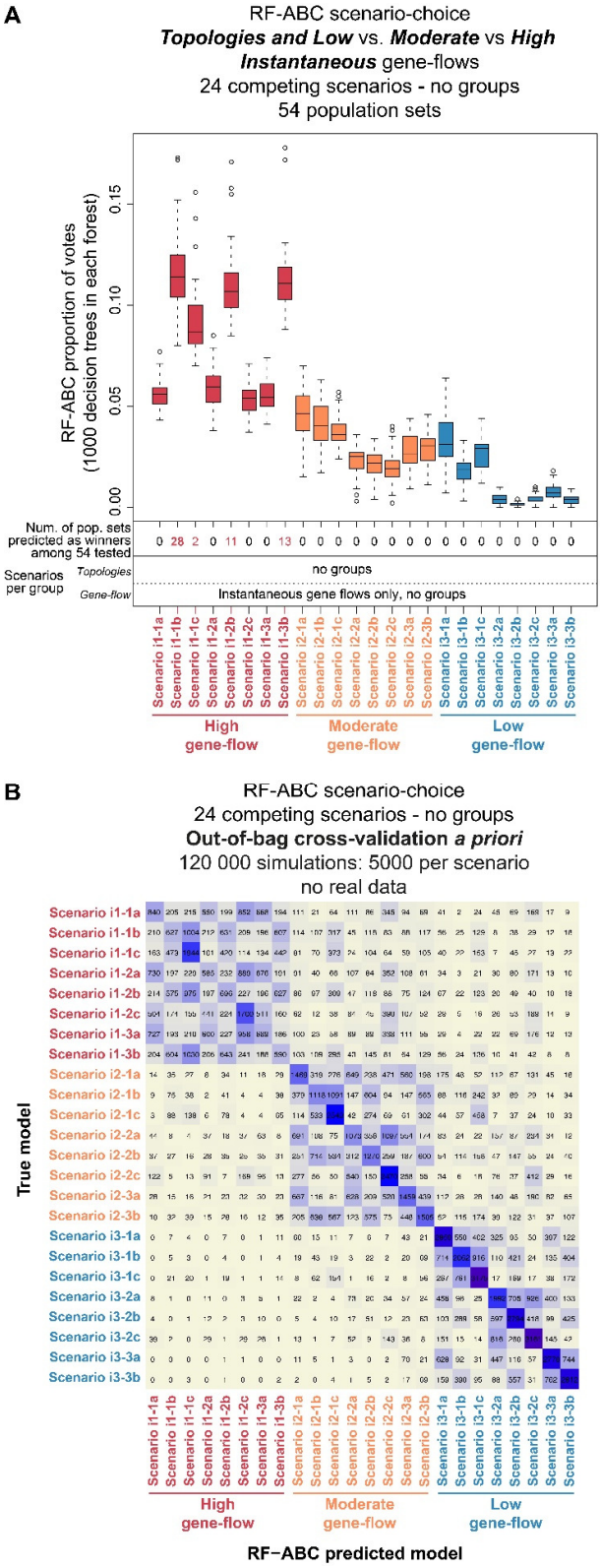
<u>FigureS6</u>: RF-ABC scenario choice among 48 competing scenarios without groups.
(**A**) corresponds to the RF-ABC scenario-choice results obtained when considering all 48 scenarios (5,000 simulations per scenario), as separate competitors without grouping them, for each 54 combinations of five sampled-populations separately. Description of the x and y axis legends are provided in **Figure6**. Posterior probabilities of choosing the correct scenarios for each 54 sets of population combinations ranged from 0.20553 to 0.28588, compared to a prior probability of 1/48 in this particular test. (**B**) corresponds to the RF-ABC cross-validation prior error for the scenario-choice procedure conducted in (**A**). Errors were obtained without considering observed data and using, instead, all simulations in turn as pseudo-observed data and the remaining simulations to train the RF. Note that, in (**B**), whichever the intensity of the gene-flow, when considering only recurring gene-flow processes, topologies are harder to distinguish from one another a priori. Conversely, such difficulty to distinguish a priori among topologies remains overall limited for instantaneous gene-flow processes, and errors increase only when considering the possibility of intense instantaneous gene-flows. However, as mentioned throughout the article, the capacity to distinguish a priori among scenarios in the entire space of summary-statistics values does not predict the power to predict a best scenario in the specific space occupied by observed data [45].

FigureS6: RF-ABC scenario choice among 48 competing scenarios without groups.

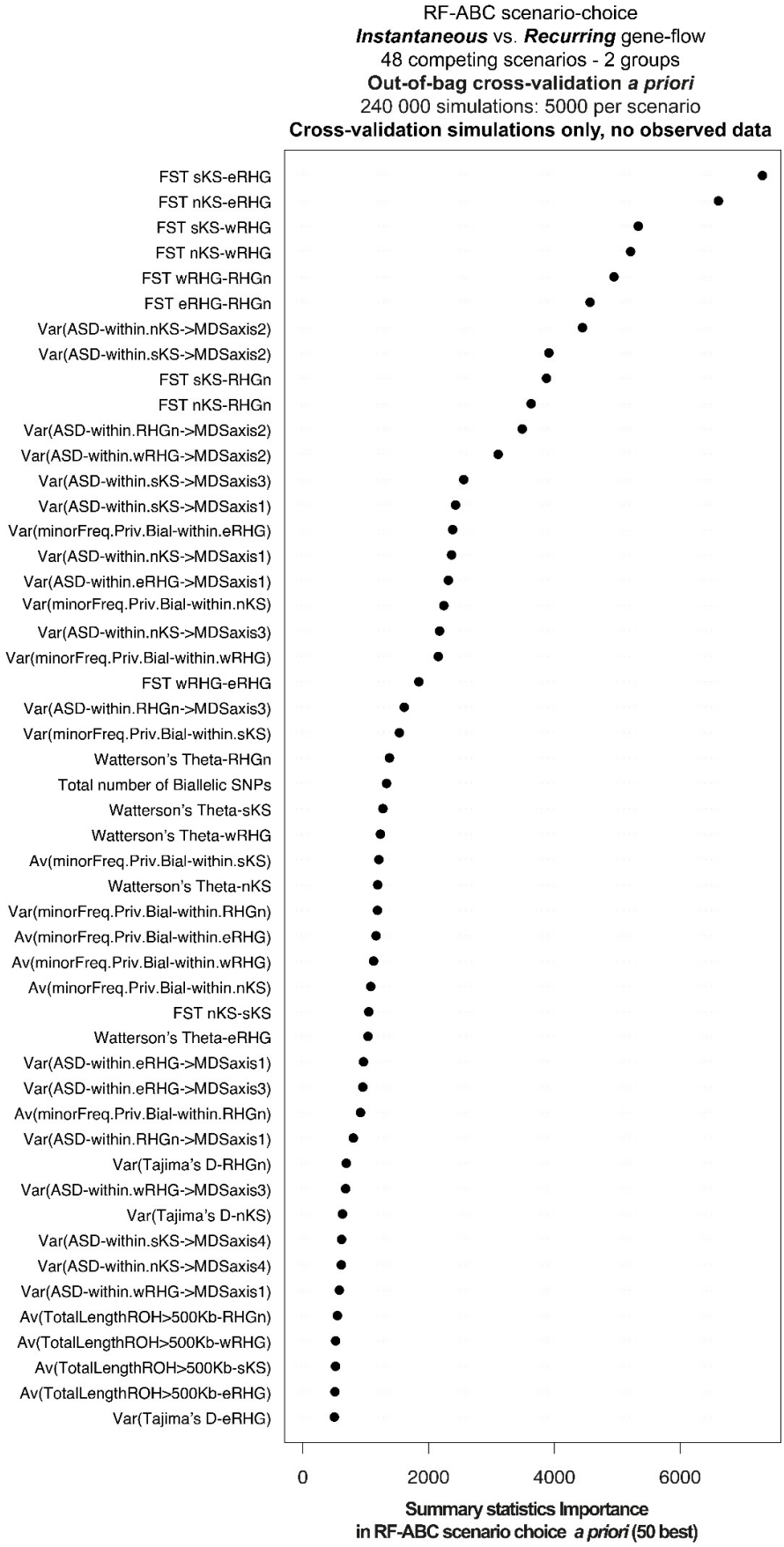




<u>FigureS7</u>: RF-ABC scenario choice among 24 competing scenarios without groups.
(**A**) corresponds to the RF-ABC scenario-choice results obtained when considering all 24 instantaneous gene-flow scenarios (5,000 simulations per scenario), as separate competitors without grouping them, for each 54 combinations of five sampled-populations separately. Description of the x and y axis legends are provided in **Figure6**. Posterior probabilities of choosing the correct scenarios for each 54 sets of population combinations ranged from 0.22380 to 0.30273, compared to a prior probability of 1/24 in this particular test. (**B**) corresponds to the RF-ABC cross-validation prior error for the scenario-choice procedure conducted in (**A**). Errors were obtained without considering observed data and using, instead, all simulations in turn as pseudo-observed data and the remaining simulations to train the RF. Note that prediction and cross-validation error results resemble those obtained for these 24 specific scenarios obtained when considering instead all the 48 scenarios in competition (**FigureS6**).

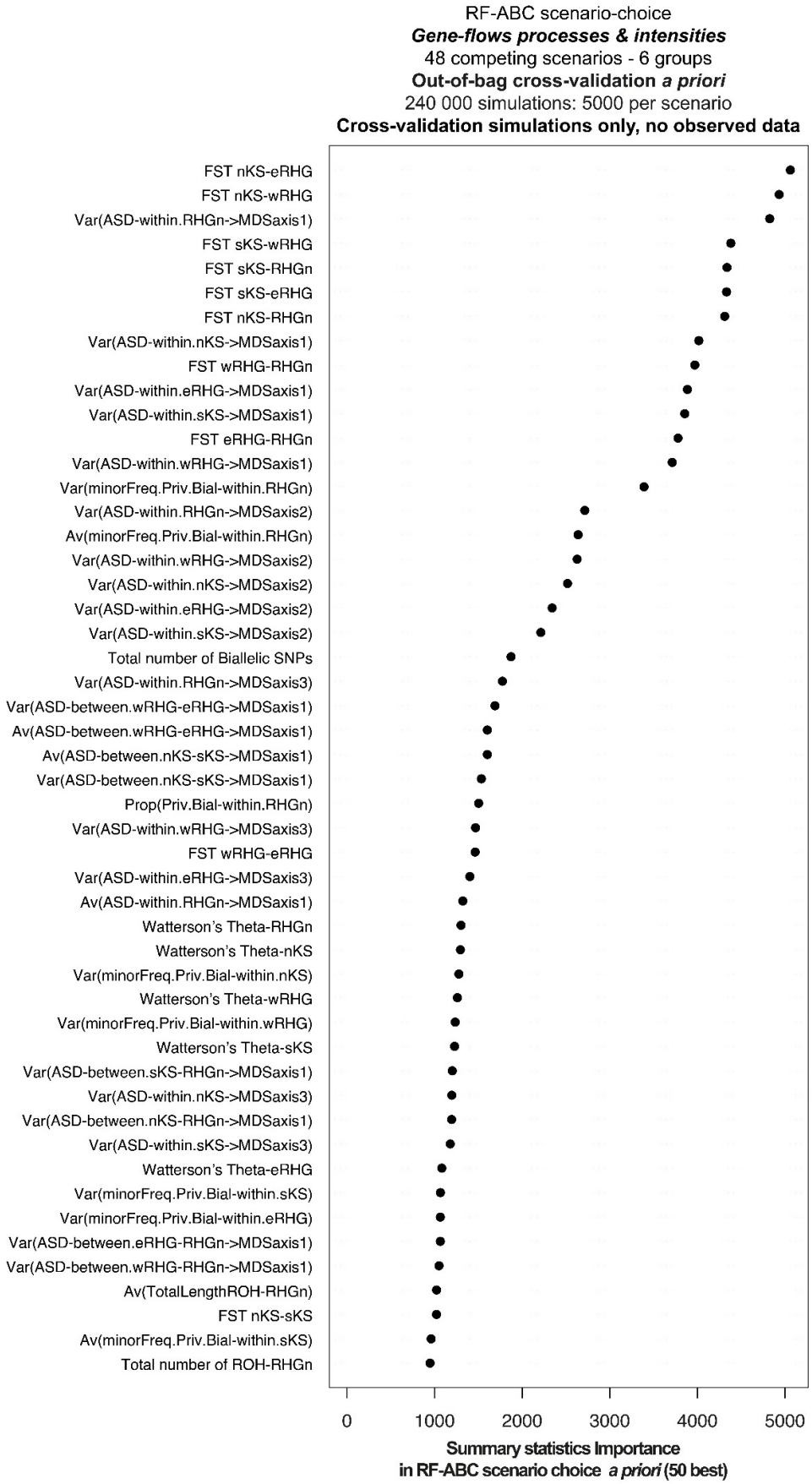


2074 **<u>FigureS8</u>: 50 best summary statistics relative contribution to the Random Forest ABC scenario-choice *a priori* based**
2075 **on simulations only without observed data – Instantaneous vs Recurring gene-flow processes all gene-flow intensities**
2076 **and all topologies "being equal".**
2077 Out-of-bag cross-validation prior error results for this RF-ABC scenario choice analysis are shown in **FigureS4 panel A.**
2078
2079
2080

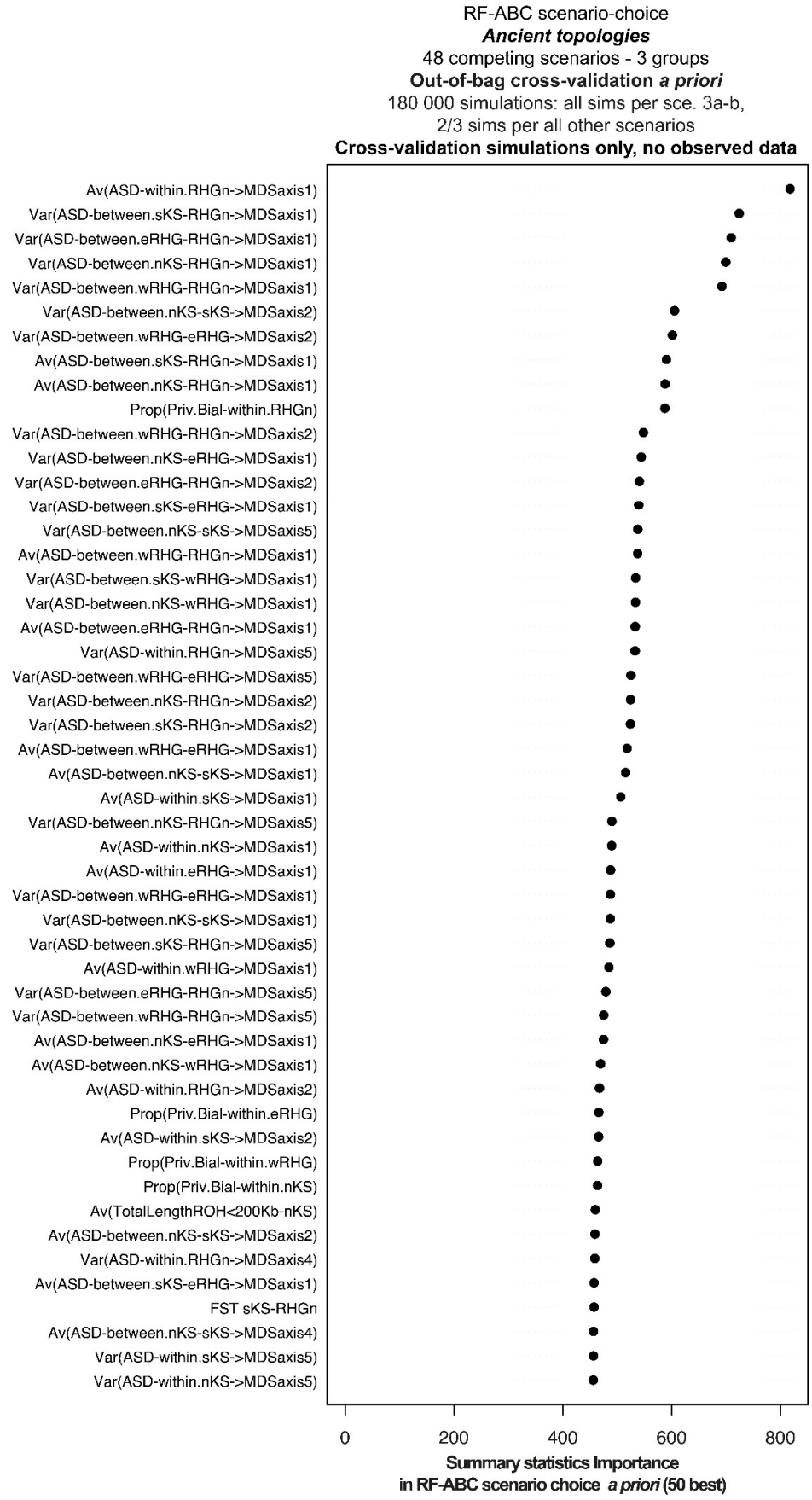


2081 **<u>FigureS9</u>: 50 best summary statistics relative contribution to the Random Forest ABC scenario-choice *a priori* based**
2082 **on simulations only without observed data – Low vs Moderate vs High gene-flow intensities, all gene-flow processes**
2083 **(instantaneous and recurring) and all topologies "being equal".**
2084 Out-of-bag cross-validation prior error results for this RF-ABC scenario choice analysis are shown in **FigureS4 panel B.**
2085
2086
2087
2088
2089

2090 **<u>FigureS10</u>: 50 best summary statistics relative contribution to the Random Forest ABC scenario-choice *a priori* based**
2091 **on simulations only without observed data – Instantaneous vs Recurring & Low vs Moderate vs High gene-flow**
2092 **intensities, all topologies "being equal".**
2093 Out-of-bag cross-validation prior error results for this RF-ABC scenario choice analysis are shown in **FigureS4 panel C.**
2094
2095
2096

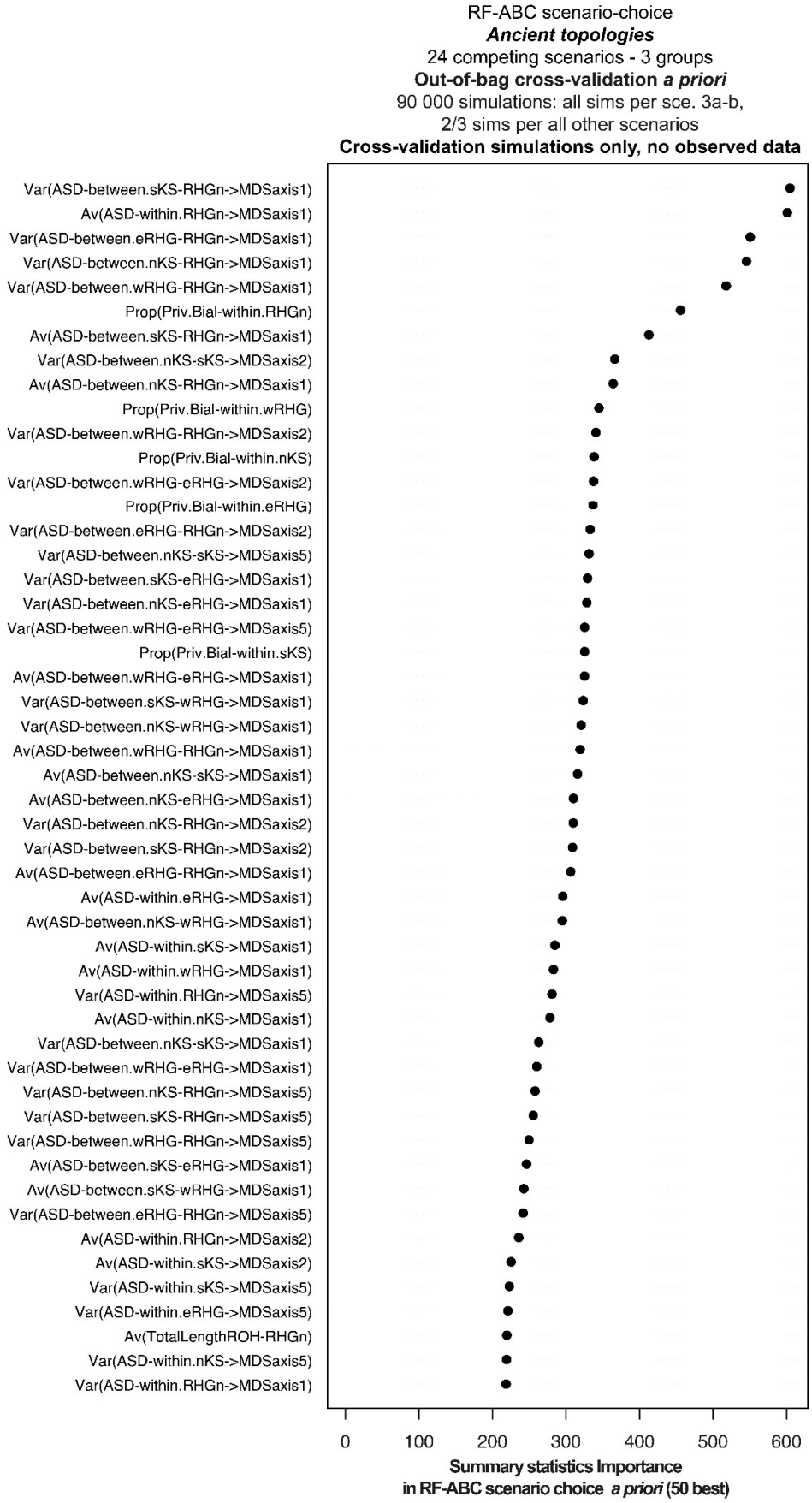


2097 **<u>FigureS11</u>: 50 best summary statistics relative contribution to the Random Forest ABC scenario-choice *a priori* based**
2098 **on simulations only without observed data – Ancient topologies, all gene-flow processes (instantaneous and recurring)**
2099 **and intensities (Low, moderate, and high), and all recent topologies "being equal".**
2100 Out-of-bag cross-validation prior error results for this RF-ABC scenario choice analysis are shown in **FigureS4 panel D.**
2101
2102
2103
2104

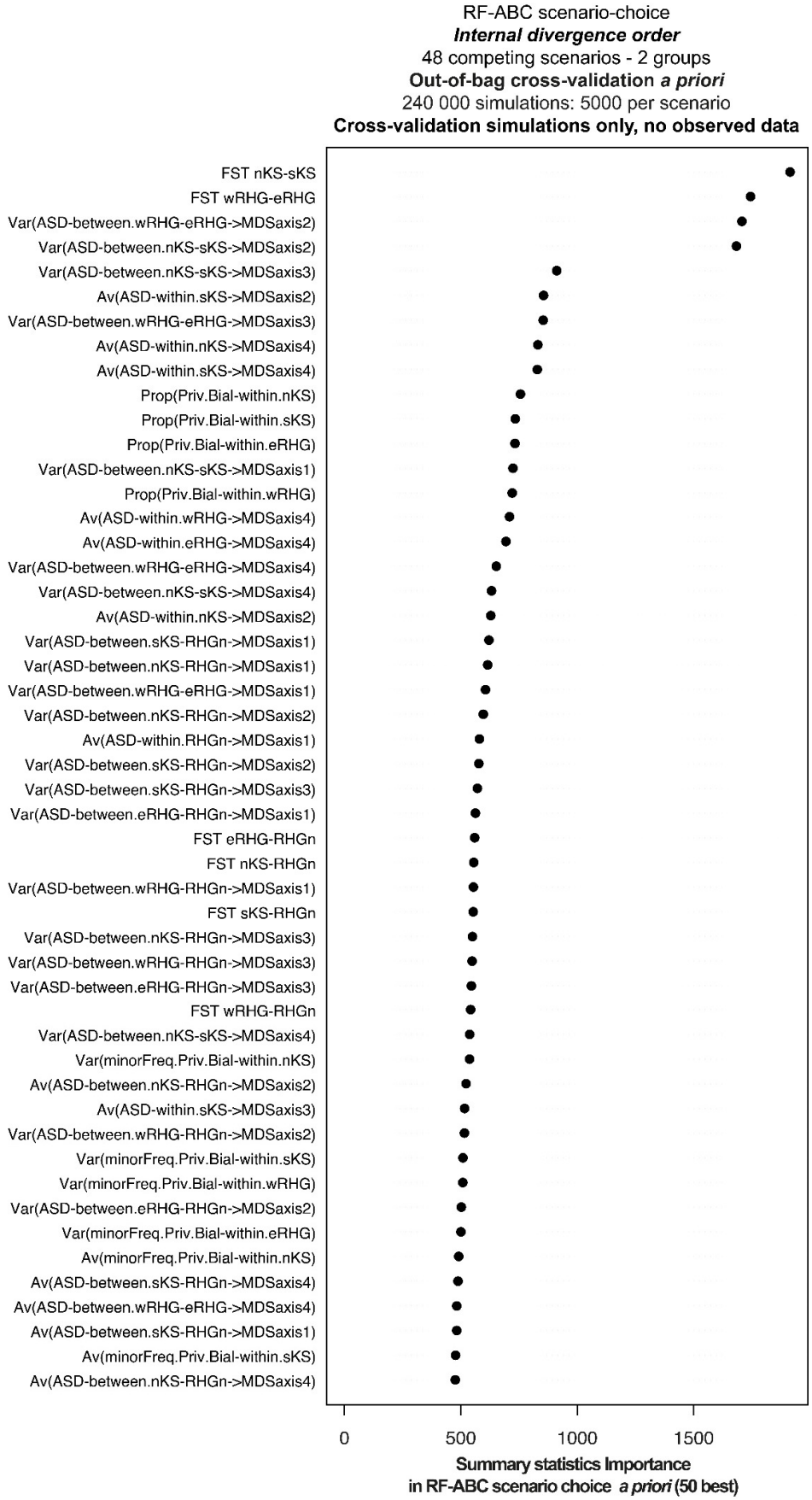


<u>FigureS12</u>: 50 best summary statistics relative contribution to the Random Forest ABC scenario-choice *a priori* based on simulations only without observed data – Ancient topologies for instantaneous gene-flow processes only, all intensities (Low, moderate, and high), and all recent topologies "being equal".

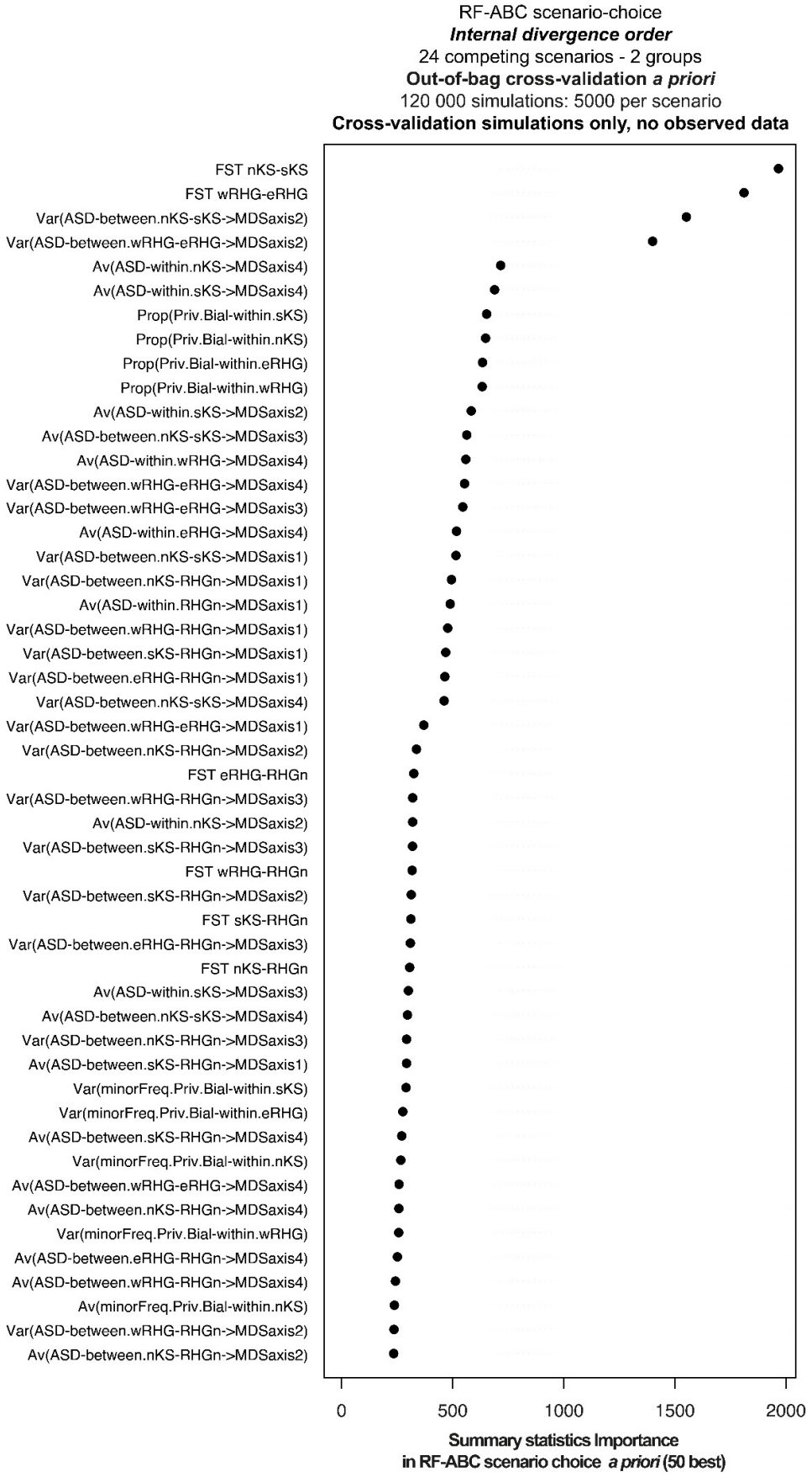
Out-of-bag cross-validation prior error results for this RF-ABC scenario choice analysis are shown in **FigureS4 panel E.**



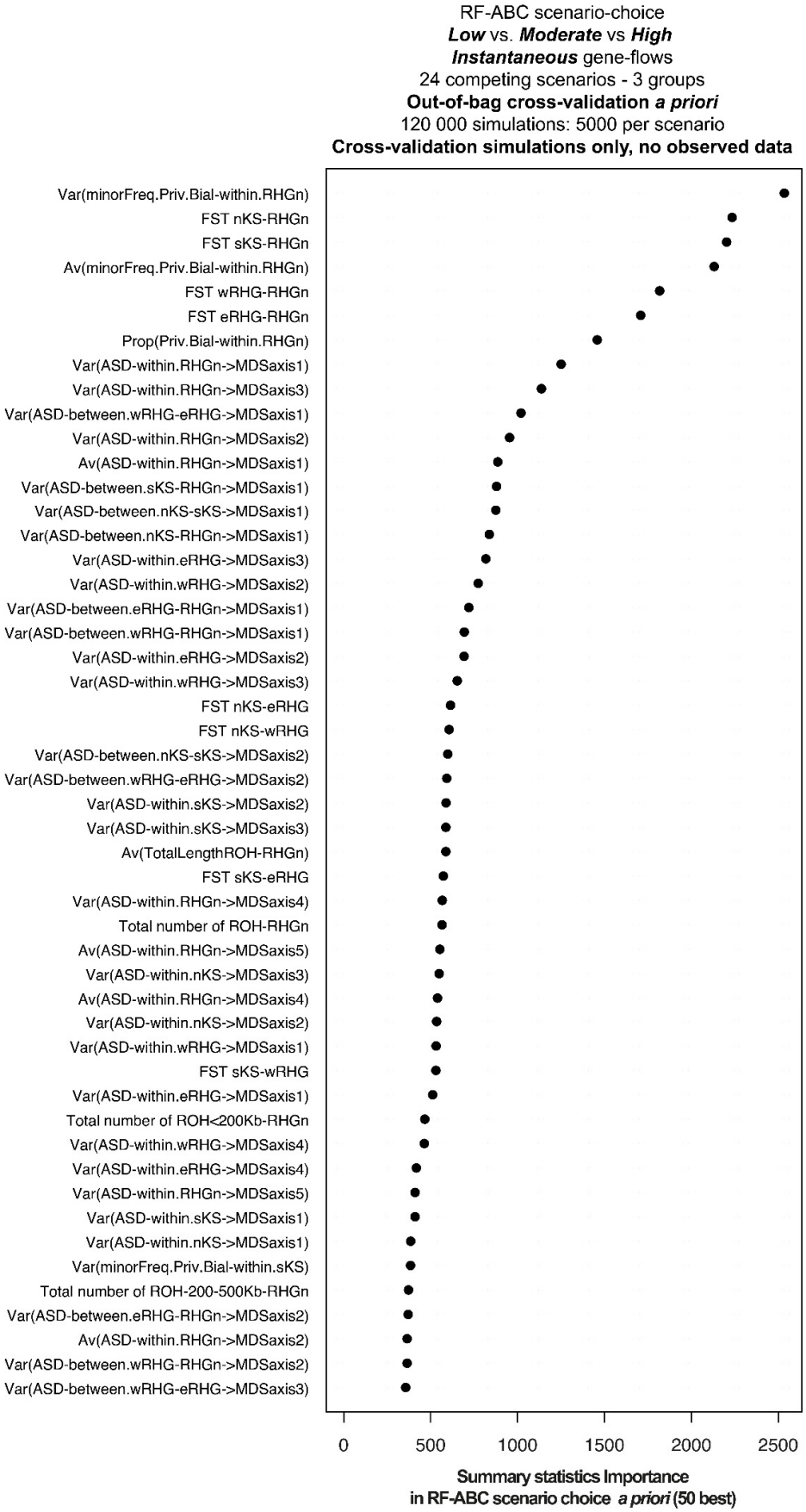
2113 **FigureS13: 50 best summary statistics relative contribution to the Random Forest ABC scenario-choice *a priori* based**
2114 **on simulations only without observed data – Internal divergence order, all gene-flow processes (instantaneous and**
2115 **recurring) and intensities (Low, moderate, and high), and all ancient topologies "being equal".**
2116 Out-of-bag cross-validation prior error results for this RF-ABC scenario choice analysis are shown in **FigureS4 panel F.**
2117
2118
2119
2120



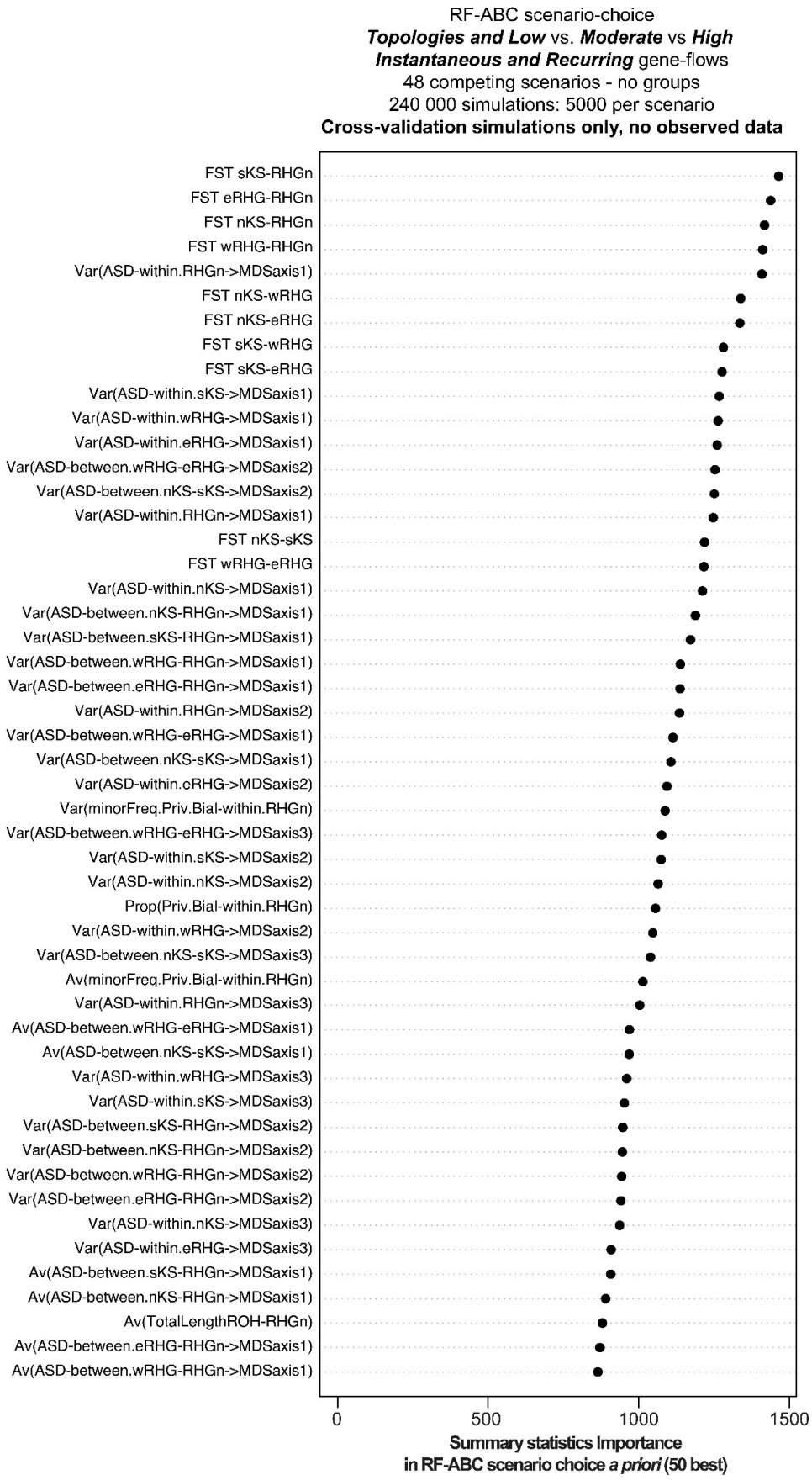
2121 **<u>FigureS14</u>: 50 best summary statistics relative contribution to the Random Forest ABC scenario-choice *a priori* based**
2122 **on simulations only without observed data – Internal divergence order for instantaneous gene-flow processes only, all**
2123 **intensities (Low, moderate, and high), and all ancient topologies "being equal".**
2124 Out-of-bag cross-validation prior error results for this RF-ABC scenario choice analysis are shown in **FigureS4 panel G.**
2125
2126
2127
2128



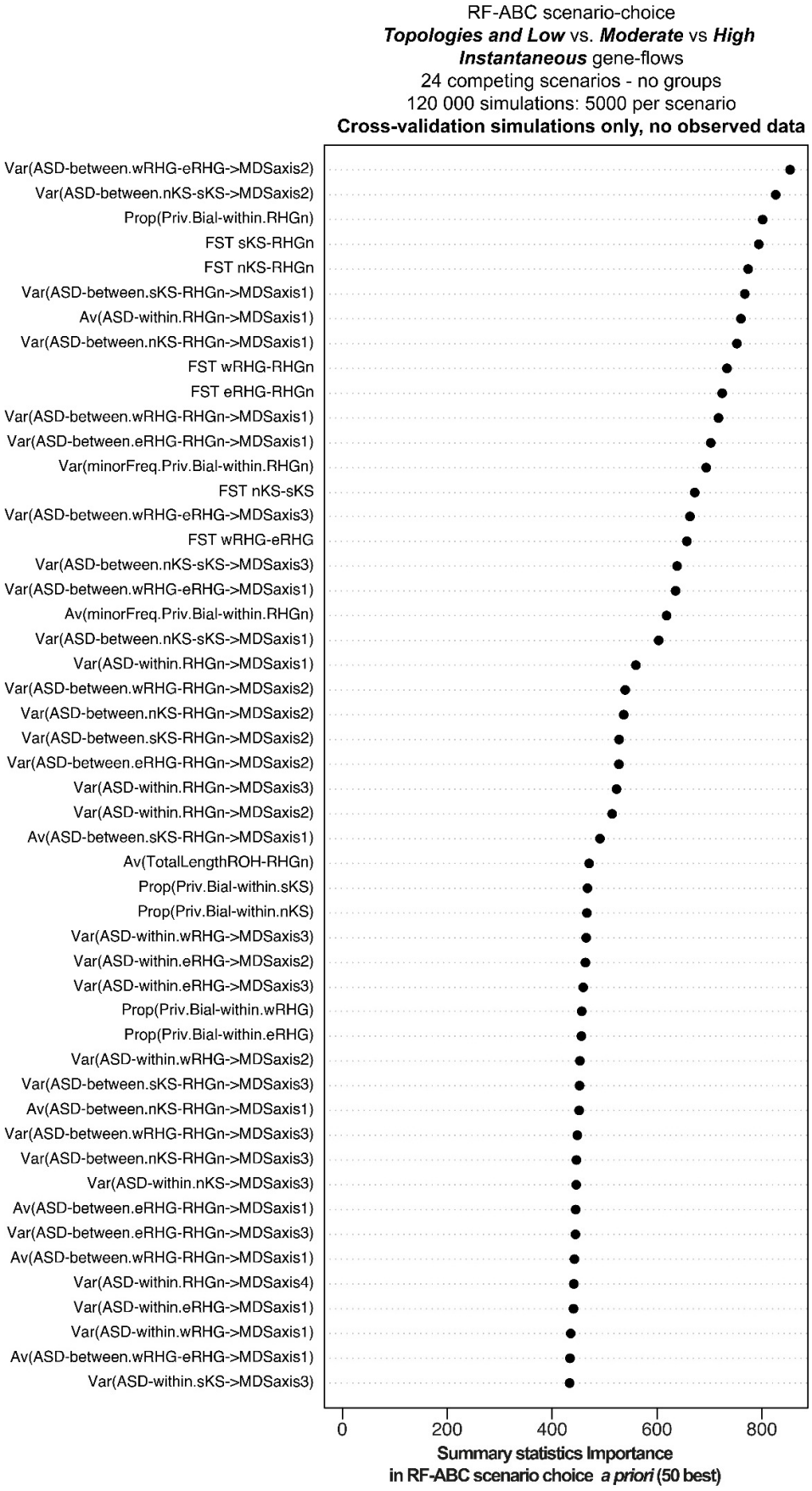
2129 **<u>FigureS15</u>: 50 best summary statistics relative contribution to the Random Forest ABC scenario-choice *a priori* based**
2130 **on simulations only without observed data – Low vs Moderate vs High gene-flow intensities for instantaneous gene-**
2131 **flow processes only, all topologies "being equal".**
2132 Out-of-bag cross-validation prior error results for this RF-ABC scenario choice analysis are shown in **FigureS5 panel B.**
2133
2134
2135
2136



2137 **<u>FigureS16</u>: 50 best summary statistics relative contribution to the Random Forest ABC scenario-choice *a priori* based**
2138 **on simulations only without observed data – All models in competition.**
2139 Out-of-bag cross-validation prior error results for this RF-ABC scenario choice analysis are shown in **FigureS6 panel B.**
2140
2141
2142
2143

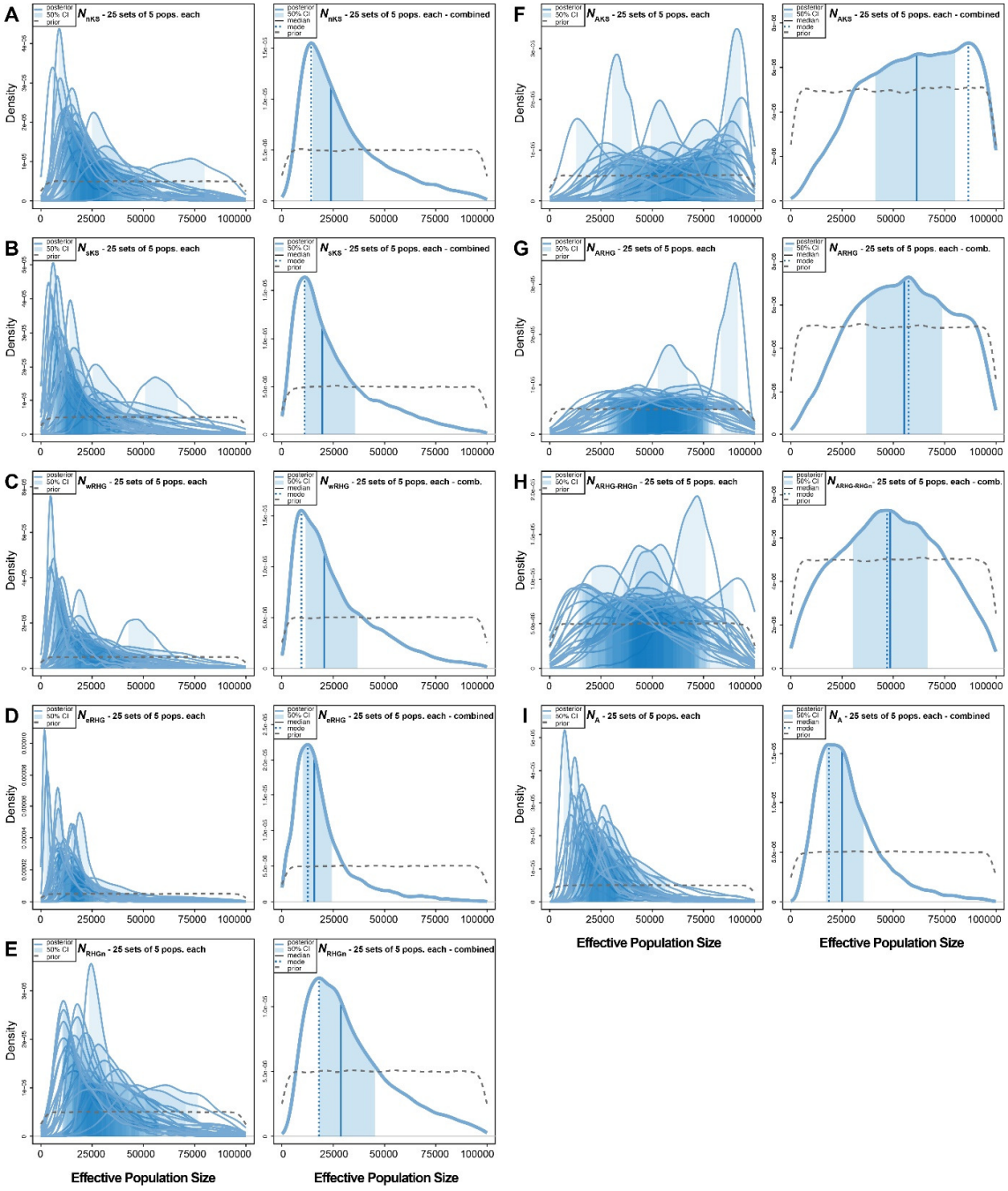


2144 **<u>FigureS17</u>: 50 best summary statistics relative contribution to the Random Forest ABC scenario-choice *a priori* based**
2145 **on simulations only without observed data – All models in competition for instantaneous gene-flow processes only.**
2146 Out-of-bag cross-validation prior error results for this RF-ABC scenario choice analysis are shown in **FigureS7 panel B.**
2147
2148
2149
2150

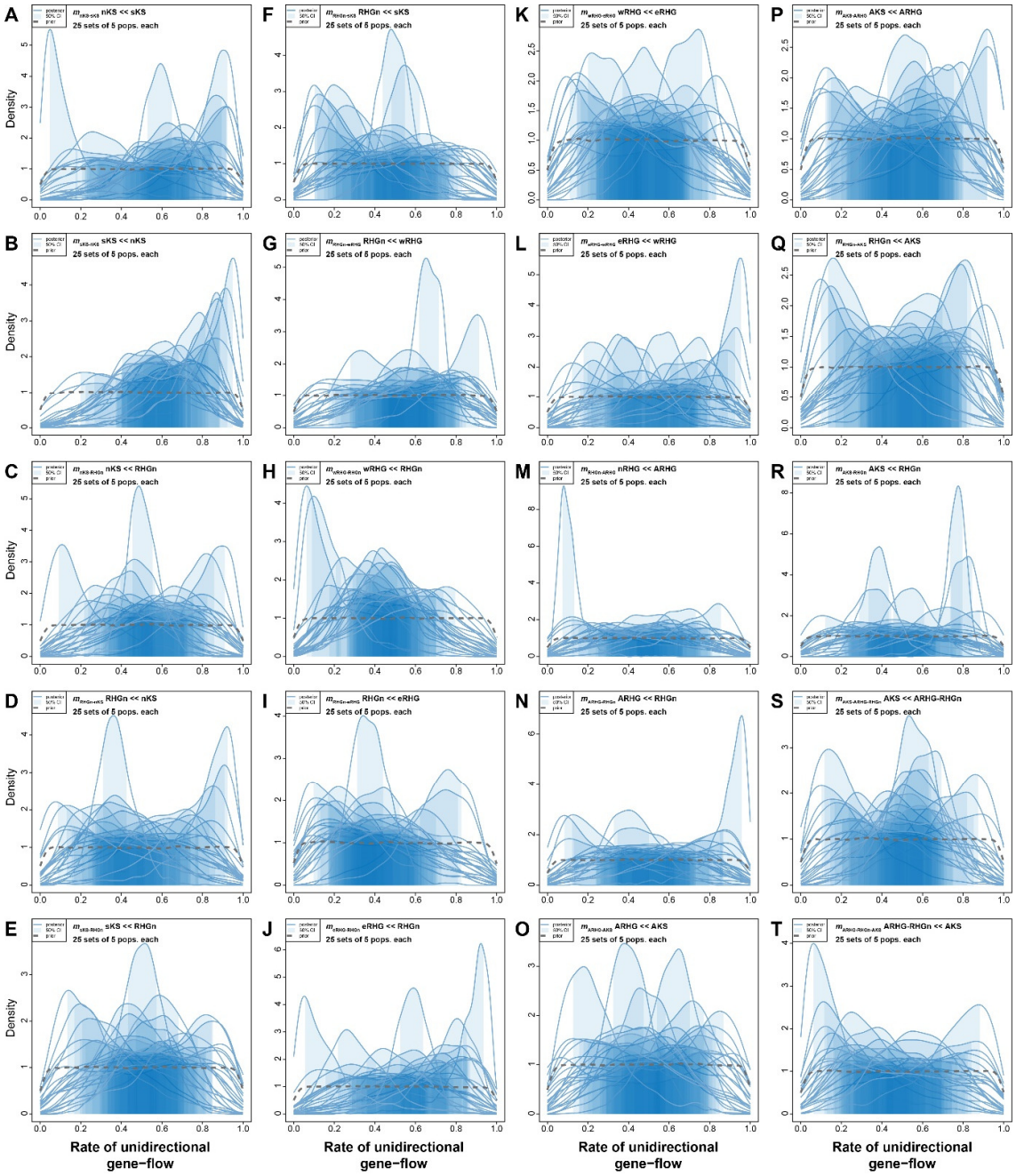


<u>FigureS18</u>: ABC posterior distribution of effective population size parameters.

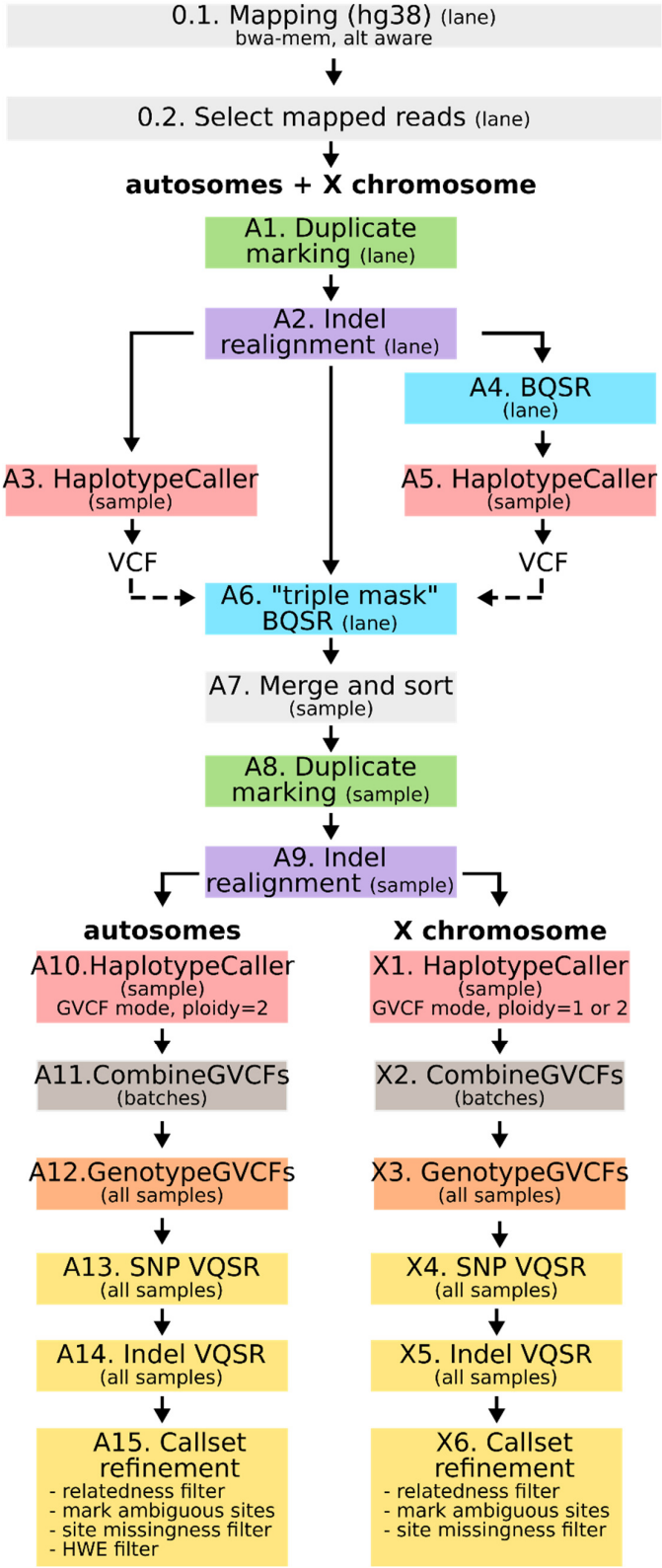
Neural Network Approximate Bayesian Computation [43,44], posterior parameter joint estimations of Effective population sizes *Ne* (in generations before present) for 25 sets of five Central and Southern African populations for which the best scenario identified by RF-ABC was Scenario i1-1b (**Figure5**, **TableS11**). NN ABC posterior parameter estimation procedures were conducted using 100,000 simulations under Scenario i1-1b, each simulation corresponding to a single vector of parameter values drawn randomly from prior distributions provided in **TableS8**. We considered 42 neurons in the hidden layer of the NN and a tolerance level of 0.01, corresponding to the 1,000 simulations providing summary-statistics closest to the observed ones, for each 25 separate analyses. NN posterior estimates are based on the logit transformation of parameter values using an Epanechnikov kernel between the corresponding parameter's prior bounds (see **Material and Methods** and **TableS8**). Posterior parameter densities are represented with solid blue lines. 50% Credibility Intervals are represented as the light blue area under the density. The median and mode values are represented as a solid and dotted blue vertical line, respectively. Parameter prior distributions are represented as dotted grey lines. For all panels, the left plots represent the NN-ABC posterior parameter distributions for each 25 sets of five Central and Southern African populations under Scenario i1-1b, separately (**TableS11** and **TableS12**). See **Figure5** and **TableS8** for all parameters' descriptions in each panel. Results are also provided in **TableS14**.



<u>FigureS19</u>: ABC posterior distribution of potentially asymmetric instantaneous gene-flow intensity parameters. Neural Network Approximate Bayesian Computation [43,44], posterior parameter joint estimations of gene-flow intensity parameters *m* (in generations before present) for 25 sets of five Central and Southern African populations for which the best scenario identified by RF-ABC was Scenario i1-1b (**Figure5**, **TableS11**). NN ABC posterior parameter estimation procedures were conducted using 100,000 simulations under Scenario i1-1b, each simulation corresponding to a single vector of parameter values drawn randomly from prior distributions provided in **TableS8**. We considered 42 neurons in the hidden layer of the NN and a tolerance level of 0.01, corresponding to the 1,000 simulations providing summary-statistics closest to the observed ones, for each 25 separate analyses. NN posterior estimates are based on the logit transformation of parameter values using an Epanechnikov kernel between the corresponding parameter's prior bounds (see **Material and Methods** and **TableS8**). Posterior parameter densities are represented with solid blue lines. 50% Credibility Intervals are represented as the light blue area under the density. Parameter prior distributions are represented as dotted grey lines. For all panels, the plots represent the NN-ABC posterior parameter distributions for each 25 sets of five Central and Southern African populations under Scenario i1-1b, separately (**TableS11** and **TableS12**). See **Figure5** and **TableS8** for all parameters' descriptions in each panel. Note that, overall, parameters are relatively little departing from their priors for numerous sets of population combinations among the 25. Also, posterior parameter distributions that are substantially departing from their priors are often highly differing from one another for each parameter. Therefore, we considered these parameters as unsatisfactorily estimated in our analyses and discuss this limitation in the **Discussion** section of the main text.



2185 **FigureS20: Schematic overview of the sequencing data processing pipeline.**
2186 All template codes with accompanying detailed explanations for all steps are provided in **Material and Methods** and in the
2187 corresponding GitHub repository (https://github.com/Gwennid/africa-wgs-descriptive).
2188



<u>TableS1</u>: Complementary population information for genetic-geographic comparisons, and for linguistic and life-style categories for ANOVA analyses.

Geographical coordinates correspond to locations in **Table1** and **Figure1**. "Traditional food-producing strategies" were obtained via ethno-anthropological investigations on the field for all original population-data, and from previously published material for comparison reference-datasets from the SGDP, HGDP, and KGP (see **Table1** and **Main Text** for references). "*nd*" stands for no explicit data concerning the genetic dataset specifically. "AGR" stands for agriculturalist and/or agropastoralist populations. "HG" stands for hunting, gathering, fishing, and horticulturalist populations. "Herder" stands for more strictly pastoral populations.

Population Name	N	Dataset	GeoCoord:X	GeoCoord:Y	Linguistic Family	Sub-Linguistic Family	Traditional Food-Producing Strategy
Baka	7	this study	13.86667	3.11667	Niger-Congo	Niger-Congo,Adamawa-Ubangian	HG
Nzime	5	this study	14.10000	2.98333	Niger-Congo	Niger-Congo, Narrow-Bantu	AGR
Ba.Kola	5	this study	10.48010	3.12090	Niger-Congo	Niger-Congo, Narrow-Bantu	HG
Ngumba	5	this study	10.72630	3.23490	Niger-Congo	Niger-Congo, Narrow-Bantu	AGR
Bi.Aka_Mbati	5	this study	17.83462	3.76069	Niger-Congo	Niger-Congo, Narrow-Bantu	HG
Ba.Kiga	5	this study	29.61434	-0.98659	Niger-Congo	Niger-Congo, Narrow-Bantu	AGR
Ba.Twa	6	this study	29.61434	-0.98659	Niger-Congo	Niger-Congo, Narrow-Bantu	HG
Nsua	5	this study	30.25151	0.95051	Nilo-Saharan	Nilo-Saharan, Eastern-Sudanic	HG
Ba.Konjo	5	this study	30.25151	0.91938	Niger-Congo	Niger-Congo, Narrow-Bantu	AGR
Karretjie People	5	this study	25.10101	-30.71264	Khoisan	Khoisan	Herder
Khutse-San	5	this study	24.66980	-23.65090	Khoisan	Khoisan	HG
Ju|'hoansi	5	this study	20.49500	-19.59740	Khoisan	Khoisan	HG
Nama	5	this study	17.07275	-22.55856	Khoisan	Khoisan	Herder
!Xun	5	this study	17.66602	-14.62894	Khoisan	Khoisan	Herder
Bantu_Herero	2	SGDP	19.00000	-22.00000	Niger-Congo	Niger-Congo, Narrow-Bantu	Herder
Bantu_Kenya	2	SGDP	37.00000	-3.00000	Niger-Congo	Niger-Congo, Narrow-Bantu	AGR
Bantu_Tswana	2	SGDP	24.00000	-28.00000	Niger-Congo	Niger-Congo, Narrow-Bantu	AGR
Biaka	2	SGDP	17.00000	4.00000	Niger-Congo	Niger-Congo, Narrow-Bantu	HG
Coloured	8	SAHGP	18.41740	-33.92899	Indo-European	Indo-European, Afrikaans	*nd*
Dinka	4	SGDP, HGDP	27.40000	8.78333	Nilo-Saharan	Nilo-Saharan, Eastern-Sudanic	AGR
Esan	3	SGDP, KGP	6.00000	6.50000	Niger-Congo	Niger-Congo, Volta-Niger	AGR
Gambian	2	SGDP	-16.70000	13.40000	Niger-Congo	Niger-Congo, Mande	AGR
Igbo	2	SGDP	7.00000	6.00000	Niger-Congo	Niger-Congo, Volta-Niger	AGR
Ju|'hoansi_comp	4	SGDP, HGDP	21.50000	-18.90000	Khoisan	Khoisan	HG
#Khomani	2	SGDP	20.80000	-27.00000	Khoisan	Khoisan	Herder
Kongo	1	SGDP	11.80000	2.50000	Niger-Congo	Niger-Congo, Narrow-Bantu	AGR
Lemande	2	SGDP	11.10000	4.50000	Niger-Congo	Niger-Congo, Narrow-Bantu	AGR
Luhya	3	SGDP, KGP	36.80000	1.30000	Niger-Congo	Niger-Congo, Narrow-Bantu	AGR
Luo	2	SGDP	34.30000	-0.10000	Nilo-Saharan	Nilo-Saharan, Eastern-Sudanic	AGR
Maasai	2	SGDP	35.20000	-1.50000	Nilo-Saharan	Nilo-Saharan, Eastern-Sudanic	Herder
Mandenka	4	SGDP, HGDP	-12.00000	12.00000	Niger-Congo	Niger-Congo, Mande	AGR
Mandinka	1	KGP	-16.70000	13.40000	Niger-Congo	Niger-Congo, Mande	AGR
Mbuti	5	SGDP, HGDP	29.00000	1.00000	Niger-Congo	Niger-Congo, Narrow-Bantu	HG
Mende	3	SGDP	-13.20000	8.20000	Niger-Congo	Niger-Congo, Mande	AGR
Mozabite	2	SGDP	3.00000	32.00000	Afro-Asiatic	Afro-Asiatic, Berber	AGR
Saharawi	2	SGDP	-8.90000	27.30000	Afro-Asiatic	Afro-Asiatic, Hassaniya	*nd*
Somali	1	SGDP	48.30000	5.60000	Afro-Asiatic	Afro-Asiatic, Cushitic	*nd*
Sotho	7	SAHGP	27.13588	-28.08954	Niger-Congo	Niger-Congo, Narrow-Bantu	*nd*
Xhosa	8	SAHGP	25.62075	-33.96171	Niger-Congo	Niger-Congo, Narrow-Bantu	*nd*
Yoruba	4	SGDP, HGDP	3.90000	7.40000	Niger-Congo	Niger-Congo, Volta-Niger	AGR
Zulu	1	SAHGP	28.04972	-26.20500	Niger-Congo	Niger-Congo, Narrow-Bantu	*nd*

2200 **TableS2: Whole genome total variant counts among 177 unrelated worldwide individuals.**
2201 Variant counts compared to the reference sequence for the human genome GRCh38 and previously reported variants in dbSNP
2202 156. "SD" stands for "standard deviation". See **Material and Methods** for details about the quality control, relatedness
2203 filtering, and variant count procedures.
2204

	Total variant count		Per individual variant count among 177 worldwide unrelated individuals			
	Original dataset: 73 Central and Southern African unrelated individuals	Entire dataset: 177 worldwide unrelated individuals	Total	SD	Min	Max
Number of biallelic SNPs	26,780,319	36,272,545	4,245,000	375,379	3,093,785	4,642,544
Number of biallelic SNPs not in dbSNP 156	854,114	1,055,245	8,052	4,563	1,152	27,406
Number of multiallelic SNPs	241,428	257,906	56,330	5,162	38,057	61,716
Number of simple indels	2,454,965	3,159,306	385,391	35,522	278,237	421,976
Number of novel simple indels not in dbSNP 156	114,362	189,124	1,651	514	771	627
Number of complex indels	969,025	977,542	414,037	71,029	289,993	566,951
Number of variants appearing in a single sample	8,404,499	12,638,316	71,403	15,094	22,444	104,211
Number of filtered SNPs	2,940,629	3,296,775	247,172	48,579	66,597	350,377

2205
2206

TableS3: Whole genome biallelic SNPs counts in 46 worldwide populations.

Mean number of bi-allelic SNPs across unrelated individuals from 46 worldwide populations, and associated standard deviations (when applicable), compared to the reference sequence for the human genome GRCh38 and compared to previously reported variants in dbSNP 156. See **Material and Methods** for details about the quality control, relatedness filtering, and variant count procedures. Populations are ordered in increasing mean number of SNPs per pop. The information original to this study is indicated in bold. "SD" stands for standard deviation. "*na*" stands for not applicable. Additional population information and geographical location of samples can be found in **Figure1**, **Table1**, and **TableS1**.

Population Name[1]	Sampling location	Dataset[2]	N[3]	Mean coverage X (SD)	Mean number of biallelic SNPs (SD)	Mean number of biallelic SNPs not in dbSNP 156 (SD)
Karitiana	Brazil	SGDP, HGDP	5	31.1 (7.8)	3,214,086 (69,241)	1,030 (547)
Papuan	Papua New Guinea	SGDP, HGDP	6	35.8 (7.8)	3,414,518 (74,893)	841 (290)
French	France	SGDP, HGDP	4	33.8 (8.7)	3,455,358 (77,954)	652 (225)
Dai	China	SGDP, KGP	6	37.7 (10.5)	3,480,744 (67,644)	695 (407)
CEU	United States of America	KGP	2	67.3 (0.2)	3,551,792 (7,819)	1,539 (57)
Saharawi	Western Sahara	SGDP	2	43.7 (2.2)	3,726,018 (27,793)	590 (21)
Mozabite	Algeria	SGDP	2	34.9 (1.6)	3,736,308 (69,393)	563 (11)
Somali	Kenya	SGDP	1	37.7 (*na*)	3,940,793 (*na*)	540 (*na*)
Coloured	South Africa	SAHGP	8	44.9 (2.2)	4,000,396 (97,447)	10,169 (2,319)
Dinka	Sudan	SGDP, HGDP	4	31.3 (5.8)	4,142,138 (97,001)	3,146 (4,811)
Maasai	Kenya	SGDP	2	38.6 (5.5)	4,169,732 (5,281)	654 (27)
Mandinka	Gambia	KGP	1	27.2 (*na*)	4,193,651 (*na*)	1,602 (*na*)
Mandenka	Senegal	SGDP, HGDP	4	30 (7.6)	4,217,398 (107,847)	914 (464)
Yoruba	Nigeria	SGDP, HGDP	4	31.1 (5)	4,233,500 (11,5371)	755 (195)
Ba.Kiga	**Mukono (Uganda)**	**this study**	**5**	**39.4 (4.6)**	**4,274,466 (25,201)**	**5,391 (191)**
Gambian	Gambia	SGDP	2	35.5 (0.8)	4,277,874 (12,976)	816 (107)
Esan	Nigeria	SGDP, KGP	3	39.7 (9.5)	4,284,041 (18,359)	704 (67)
Bantu_Kenya	Kenya	SGDP	2	38.7 (6.8)	4,288,533 (10,954)	611 (88)
Luo	Kenya	SGDP	2	34 (0.2)	4,293,632 (1,003)	722 (4)
Luhya	Kenya	SGDP, KGP	3	40.6 (12.1)	4,294,198 (28,096)	704 (37)
Igbo	Nigeria	SGDP	2	36.3 (8)	4,294,870 (1,867)	1,166 (9)
Lemande	Cameroon	SGDP	2	36.3 (0.5)	4,306,880 (5,284)	1,530 (37)
Bantu_Herero	Namibia	SGDP	2	38.4 (2.2)	4,318,604 (33,525)	690 (161)
Ba.Konjo	**Mulimassenge (Uganda)**	**this study**	**5**	**43 (6.4)**	**4,322,177 (6,336)**	**5,349 (773)**
Nzime	**Messea (Cameroon)**	**this study**	**5**	**35.5 (2.6)**	**4,322,734 (18,671)**	**5,305 (208)**
Ngumba	**Dispersed between Lolodorf and Kribi (Cameroon)**	**this study**	**5**	**40.2 (2.2)**	**4,332,324 (9,273)**	**4,810 (552)**
Kongo	Cameroon	SGDP	1	42.8 (*na*)	4,336,731 (*na*)	1233 (*na*)
Mende	Sierra Leone	SGDP, KGP	3	39.6 (8.3)	4,341,512 (13,486)	866 (180)
Zulu	South Africa	SAHGP	1	42.8 (*na*)	4,376,691 (*na*)	8,146 (*na*)
Xhosa	South Africa	SAHGP	8	42.4 (1.1)	4,385,032 (56,902)	10,085 (1,300)
Ba.Twa	**Kebiremu, Byumba, Kitariro, Mgungu, Nteko (Uganda)**	**this study**	**6**	**53.8 (17.7)**	**4,426,223 (32,552)**	**20,251 (1,022)**
Bantu_Tswana	South Africa	SGDP	2	36.7 (3.1)	4,430,870 (55,537)	720 (37)
Mbuti	Democratic Republic of Congo	SGDP, HGDP	5	32.1 (7.4)	4,434,425 (110,658)	1,227 (488)
Sotho	South Africa	SAHGP	7	42.9 (6.4)	4,446,796 (23,172)	12,040 (1,150)
Ba.Kola	**Dispersed between Lolodorf and Kribi (Cameroon)**	**this study**	**5**	**39.2 (1.9)**	**4,475,922 (50,261)**	**6,995 (1,948)**
Nsua	**Bundimassoli (Uganda)**	**this study**	**5**	**37.6 (4.9)**	**4,486,217 (22,774)**	**16,443 (984)**
Baka	**Bosquet (Cameroon)**	**this study**	**7**	**40.7 (3.5)**	**4,503,977 (23,946)**	**3,927 (144)**
Bi.Aka_Mbati	**Bombeketi section of Bagandou (Central African Republic)**	**this study**	**5**	**43.4 (13.6)**	**4,514,318 (17,376)**	**4,587 (198)**
Biaka	Central African Republic	SGDP	2	39.1 (0.3)	4,521,140 (8,155)	726 (52)
Nama	**Windhoek (Namibia)**	**this study**	**5**	**48.9 (4.3)**	**4,530,168 (18,807)**	**27,902 (462)**
Ju|'hoansi_comp	Namibia	SGDP, HGDP	4	33.9 (7)	4,554,593 (103,099)	1,382 (273)
#Khomani	South Africa	SGDP	2	44.7 (2.8)	4,559,846 (79,851)	1,364 (48)
Khutse_San	**Kutse Game reserve (Botswana)**	**this study**	**5**	**46.3 (6.7)**	**4,592,852 (27,794)**	**27,989 (1,353)**
Karretjie	**Colesberg (South Africa)**	**this study**	**5**	**46.6 (5.4)**	**4,601,724 (26,951)**	**27,720 (1,968)**
!Xun	**Omega camp (Namibia) and Schmidtsdrift (Sout Africa)[4]**	**this study**	**5**	**45.6 (1.9)**	**4,614,835 (10,047)**	**28,923 (1,371)**
Ju|'hoansi	**Tsumkwe (Namibia)**	**this study**	**5**	**49 (4.6)**	**4,628,957 (8,139)**	**26,665 (6,491))**

2217 **<u>TableS4</u>: Correlations between genetic pairwise differentiation and geography.**
2218 Standardized average Allele Sharing Dissimilarities across pairs of individuals from different populations are compared, for
2219 all population pairs, with the logarithm of geographic distances using Spearman and Pearson correlations, and tested using
2220 1000 Mantel permutations. See **Figure1**, **Table1, and TableS1** for population geographic location and information.
2221

Geographic Region	Population Samples	Nb pops.	Nb indivs.	Spearman ρ	Mantel Permutation p	Pearson r	Mantel Permutation p
Africa	All African population samples in **Figure1**, **Table1, and TableS1**	41	154	0.03573	0.32268	0.05562	0.16384
Sub-Sahara Africa	All African population samples in **Figure1**, **Table1, and TableS1,** except Northern African Sarahwi and Mozabites	38	150	0.07321	0.16883	0.09707	0.05294
West Western Africa	Gambian, Mandenka, Mandinka, Mende	4	10	0.50022	0.09291	0.38444	0.33467
Wider Central Africa	Igbo, Esan, Yoruba, Ba.Kola (wRHG), Ngumba (wRHGn), Baka (wRHG), Nzime(wRHGn), Kongo, Lemande, Bi.Aka_Mbati (wRHG), Bi.Aka (wRHG), Ba.Twa (eRHG), Ba.Kiga (eRHGn), Nsua (eRHG), Ba.Konjo (eRHGn), Mbuti (eRHG), Dinka	17	71	0.15587	0.07393	0.06298	0.21378
Central Africa - Congo Basin	Ba.Kola (wRHG), Ngumba (wRHGn), Baka (wRHG), Nzime(wRHGn), Kongo, Lemande, Bi.Aka_Mbati (wRHG), Bi.Aka (wRHG), Ba.Twa (eRHG), Ba.Kiga (eRHGn), Nsua (eRHG), Ba.Konjo (eRHGn), Mbuti (eRHG)	13	58	0.19952	0.03996	0.13451	0.05594
Eastern Africa	Bantu_Kenya, Luhya, Luo, Maasai, Somali	5	10	0.13939	0.25574	-0.04705	0.36863
Southern Africa	!Xun (nKS), Ju|'hoansi (nKS), Ju|'hoansi_comp (nKS), Bantu_Herero, Nama (sKS), Khutse San (Central KS), #Khomani, Zulu, Sotho, Bantu_Tswana, Karretjie_People (sKS), Coloured, Xhosa	13	59	0.07123	0.20979	0.10440	0.15485

2222
2223

2224 **TableS5: Permutation ANOVA of average population pairwise ASD explained by population linguistic families at**
2225 **different geographical scale in Africa.**
2226 See **Figure1**, **Table1** and **TableS1** for population geographical location and linguistic family categorization criteria.
2227 Significant ANOVA results are indicated in bold and non-significant results in italics. "*na*" indicates that the ANOVA test is
2228 not applicable due to insufficient numbers of levels in the linguistic family category.
2229

				ANOVA partial r^2: % Var(Av. ASD among pairs of populations) explained		
Geographic Region	**Population Samples**	**Nb pops**	**Nb Linguistic families**	**among linguistic families**	**within linguistic families**	**1000 ANOVA permutation p**
Africa	All African population samples in **Figure1**, **Table1**, and **TableS1**	41	5	**15.064%**	**84.936%**	**<0.001**
Sub-Sahara Africa	All African population samples in **Figure1**, **Table1**, and **TableS1**, except Northern African Sarahwi and Mozabites	39	5	**14.811%**	**85.189%**	**<0.001**
West Western Africa	Gambian, Mandenka, Mandinka, Mende	4	1	*Na*	*Na*	*na*
Wider Central Africa	Igbo, Esan, Yoruba, Ba.Kola (wRHG), Ngumba (wRHGn), Baka (wRHG), Nzime(wRHGn), Kongo, Lemande, Bi.Aka_Mbati (wRHG), Bi.Aka (wRHG), Ba.Twa (eRHG), Ba.Kiga (eRHGn), Nsua (eRHG), Ba.Konjo (eRHGn), Mbuti (eRHG), Dinka	17	2	*6.389%*	*93.611%*	*0.1698*
Central Africa - Congo Basin	Ba.Kola (wRHG), Ngumba (wRHGn), Baka (wRHG), Nzime(wRHGn), Kongo, Lemande, Bi.Aka_Mbati (wRHG), Bi.Aka (wRHG), Ba.Twa (eRHG), Ba.Kiga (eRHGn), Nsua (eRHG), Ba.Konjo (eRHGn), Mbuti (eRHG)	13	2	*9.172%*	*90.828%*	*0.1489*
Eastern Africa	Bantu_Kenya, Luhya, Luo, Maasai, Somali	5	3	*50.311%*	*49.689%*	*0.297*
Southern Africa	!Xun (nKS), Ju|'hoansi (nKS), Ju|'hoansi_comp (nKS), Bantu_Herero, Nama (sKS), Khutse San (Central KS), #Khomani, Zulu, Sotho, Bantu_Tswana, Karretjie_People (sKS), Coloured, Xhosa	13	3	**21.544%**	**78.456%**	**<0.001**

2230
2231

2232 **TableS6: Permutation ANOVA of average population pairwise ASD explained by population sub-linguistic families at**
2233 **different geographical scale in Africa.**
2234 See **Figure1**, **Table1** and **TableS1** for population geographical location and sub-linguistic family categorization criteria.
2235 Significant ANOVA results are indicated in bold and non-significant results in italics. "*na*" indicates that the ANOVA test is
2236 not applicable due to insufficient numbers of levels in the linguistic family category.
2237

Geographic Region	Population Samples	Nb pops	Nb Sub-linguistic families	ANOVA partial r^2: % Var(Av. ASD among pairs of populations) explained		
				among Sub-linguistic families	within Sub-linguistic families	1000 ANOVA permutation p
Africa	All African population samples in **Figure1**, **Table1**, and **TableS1**	41	10	**26.109%**	**73.891%**	**<0.001**
Sub-Sahara Africa	All African population samples in **Figure1**, **Table1**, and **TableS1**, except Northern African Sarahwi and Mozabites	39	8	**22.759%**	**77.241%**	**<0.001**
West Western Africa	Gambian, Mandenka, Mandinka, Mende	4	1	*na*	*na*	*na*
Wider Central Africa	Igbo, Esan, Yoruba, Ba.Kola (wRHG), Ngumba (wRHGn), Baka (wRHG), Nzime(wRHGn), Kongo, Lemande, Bi.Aka_Mbati (wRHG), Bi.Aka (wRHG), Ba.Twa (eRHG), Ba.Kiga (eRHGn), Nsua (eRHG), Ba.Konjo (eRHGn), Mbuti (eRHG), Dinka	17	4	*19.441%*	*80.559%*	*0.1728*
Central Africa - Congo Basin	Ba.Kola (wRHG), Ngumba (wRHGn), Baka (wRHG), Nzime(wRHGn), Kongo, Lemande, Bi.Aka_Mbati (wRHG), Bi.Aka (wRHG), Ba.Twa (eRHG), Ba.Kiga (eRHGn), Nsua (eRHG), Ba.Konjo (eRHGn), Mbuti (eRHG)	13	3	*17.832%*	*82.168%*	*0.1538*
Eastern Africa	Bantu_Kenya, Luhya, Luo, Maasai, Somali	5	3	*50.311%*	*49.689%*	*0.297*
Southern Africa	!Xun (nKS), Ju|'hoansi (nKS), Ju|'hoansi_comp (nKS), Bantu_Herero, Nama (sKS), Khutse San (Central KS), #Khomani, Zulu, Sotho, Bantu_Tswana, Karretjie_People (sKS), Coloured, Xhosa	13	3	**21.544%**	**78.456%**	**<0.001**

2238
2239

2240 **<u>TableS7</u>: Permutation ANOVA of average population pairwise ASD explained by population traditional food-**
2241 **production strategies at different geographical scale in Africa.**
2242 See **Figure1**, **Table1** and **TableS1** for population geographical location and traditional food-production strategy categorization
2243 criteria. Significant ANOVA results are indicated in bold and non-significant results in italics. "*na*" indicates that the ANOVA
2244 test is not applicable due to insufficient numbers of levels in the linguistic family category.
2245

				ANOVA partial r^2: % Var(Av. ASD among pairs of populations) explained		
Geographic Region	**Population Samples**	**Nb pops**	**Nb Traditional food-production strategies**	**among strategies**	**within strategies**	**1000 ANOVA permutation p**
Africa	All African population samples in Figure1 and Table1, except Saharawi, Coloured, Somali, Sotho, Xhosa, and Zulu populations (TableS1)	35	3	**9.79%**	**90.21%**	**<0.001**
Sub-Sahara Africa	All African population samples in Figure1 and Table1, except Northern African Saharawi and Mozabite, and Southern African Coloured, Somali, Sotho, Xhosa, and Zulu populations (TableS1)	34	3	**10.01%**	**89.99%**	**<0.001**
West Western Africa	Gambian, Mandenka, Mandinka, Mende	4	1	*na*	*Na*	*na*
Wider Central Africa	Igbo, Esan, Yoruba, Ba.Kola (wRHG), Ngumba (wRHGn), Baka (wRHG), Nzime(wRHGn), Kongo, Lemande, Bi.Aka_Mbati (wRHG), Bi.Aka (wRHG), Ba.Twa (eRHG), Ba.Kiga (eRHGn), Nsua (eRHG), Ba.Konjo (eRHGn), Mbuti (eRHG), Dinka	17	2	**9.538%**	**90.462%**	**<0.001**
Central Africa - Congo Basin	Ba.Kola (wRHG), Ngumba (wRHGn), Baka (wRHG), Nzime(wRHGn), Kongo, Lemande, Bi.Aka_Mbati (wRHG), Bi.Aka (wRHG), Ba.Twa (eRHG), Ba.Kiga (eRHGn), Nsua (eRHG), Ba.Konjo (eRHGn), Mbuti (eRHG)	13	2	**11.153%**	**88.847%**	**0.003**
Eastern Africa	Bantu_Kenya, Luhya, Luo, Maasai	4	2	*34.198%*	*65.802%*	*0.25*
Southern Africa	!Xun (nKS), Ju|'hoansi (nKS), Ju|'hoansi_comp (nKS), Bantu_Herero, Nama (sKS), Khutse San (Central KS), #Khomani, Bantu_Tswana, Karretjie_People (sKS)	9	3	*25.979%*	*74.201%*	*0.1728*

2246
2247

2248 **<u>TableS8</u>: 48 competing scenarios parameters' description and prior distributions.**
2249 Codes for scenarios topologies and gene-flow processes and all scenario parameters are detailed in **Figure5** and in **Material**
2250 **and Methods**.
2251

TableS8: 48 competing scenarios parameters' description and prior distributions.

Parameter	Description	Prior	Scenario topology	Scenario gene-flow	Condition
t_{RHG}	Time of divergence between wRHG and eRHG	Uniform[1,15000]	1a-2a-2c-3a 1b-1c-2b-3b	i1-i2-i3 and r1-r2-r3	$t_{RHG}>t_{KS}$ $t_{KS}>t_{RHG}$
t_{KS}	Time of divergence between nKS and sKS	Uniform[1,15000]	1a-2a-2c-3a 1b-1c-2b-3b	i1-i2-i3 and r1-r2-r3	$t_{RHG}>t_{KS}$ $t_{KS}>t_{RHG}$
$t_{RHG\text{-}RHGn}$	Time of divergence between the RHGn lineage and the lineage ancestral to all RHG lineages	Uniform[1,15000]	1a-1b-1c	i1-i2-i3 and r1-r2-r3	$t_{RHG\text{-}RHGn}>t_{RHG}$
$t_{KS\text{-}RHGn}$	Time of divergence between the RHGn lineage and the lineage ancestral to all KS lineages	Uniform[1,15000]	2a-2b-2c	i1-i2-i3 and r1-r2-r3	$t_{KS\text{-}RHGn}>t_{KS}$
$t_{KS\text{-}RHG}$	Time of divergence between the KS and RHG ancestral lineages	Uniform[1,15000]	3a-3b	i1-i2-i3 and r1-r2-r3	$t_{KS\text{-}RHG}>t_{KS}$ and $t_{KS\text{-}RHG}>t_{RHG}$
$t_{KS\text{-}RHG\text{-}RHGn}$	Time of the divergence event ancestral to all lineages	Uniform[1,15000]	all 8 topol.	i1-i2-i3 and r1-r2-r3	$t_{KS\text{-}RHG\text{-}RHGn}>$ all other times
$tad_{nKS\text{-}sKS}$	Time for both the unidirectional instantaneous gene-flow events between nKS and sKS	Uniform[1,15000]	all 8 topol.	i1-i2-i3	$t_{KS}>tad_{nKS\text{-}sKS}$
m(nKS-sKS) or m(sKS-nKS)	Instantaneous gene-flow intensity from sKs into nKS, or from nKS into sKS independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	all 8 topol.	i1 i2 i3	x
m(nKS-sKS) or m(sKS-nKS)	Recurring gene-flow intensity from sKS into nKS, or from nKS into sKs independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	all 8 topol.	r1 r2 r3	x
$tad_{nKS\text{-}RHGn}$	Time for both the unidirectional instantaneous gene-flow events between nKS and RHGn	Uniform[1,15000]	all 8 topol.	i1-i2-i3	$t_{KS}>tad_{nKS\text{-}RHGn}$ and $t_{RHG\text{-}RHGn}>tad_{nKS\text{-}RHGn}$
m(nKS-RHGn) or m(RHGn-nKS)	Instantaneous gene-flow intensity from RHGn into nKS, or from nKS into RHGn independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	all 8 topol.	i1 i2 i3	x
m(nKS-RHGn) or m(RHGn-nKS)	Recurring gene-flow intensity from RHGn into nKS, or from nKS into RHGn independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	all 8 topol.	r1 r2 r3	x
$tad_{sKS\text{-}RHGn}$	Time for both the unidirectional instantaneous gene-flow events between sKS and RHGn lineages	Uniform[1,15000]	all 8 topol.	i1-i2-i3	$t_{KS}>tad_{sKS\text{-}RHGn}$ and $t_{RHG\text{-}RHGn}>tad_{sKS\text{-}RHGn}$
m(sKS-RHGn) or m(RHGn-sKS)	Instantaneous gene-flow intensity from RHGn into sKS, or from sKS into RHGn independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	all 8 topol.	i1 i2 i3	x
m(sKS-RHGn) or m(RHGn-sKS)	Recurring gene-flow intensity from RHGn into sKS, or from sKS into RHGn independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	all 8 topol.	r1 r2 r3	x
$tad_{wRHG\text{-}eRHG}$	Time for both the unidirectional instantaneous gene-flow events between wRHG and eRHG	Uniform[1,15000]	all 8 topol.	i1-i2-i3	$t_{RHG}>tad_{wRHG\text{-}eRHG}$
m(wRHG-eRHG) or m(eRHG-wRHG)	Instantaneous gene-flow intensity from eRHG into wRHG, or from wRHG into eRHG independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	all 8 topol.	i1 i2 i3	x
m(wRHG-eRHG) or m(eRHG-wRHG)	Recurring gene-flow intensity from eRHG into wRHG, or from wRHG into eRHG independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	all 8 topol.	r1 r2 r3	x
$tad_{wRHG\text{-}RHGn}$	Time for both the unidirectional instantaneous gene-flow events between wRHG and RHGn	Uniform[1,15000]	all 8 topol.	i1-i2-i3	$t_{RHG}>tad_{wRHG\text{-}RHGn}$ and $t_{KS\text{-}RHGn}>tad_{nwRHG\text{-}RHGn}$
m(wRHG-RHGn) or m(RHGn-wRHG)	Instantaneous gene-flow intensity from RHGn into wRHG, or from wRHG into RHGn independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	all 8 topol.	i1 i2 i3	x
m(wRHG-RHGn) or m(RHGn-wRHG)	Recurring gene-flow intensity from RHGn into wRHG, or from wRHG into RHGn independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	all 8 topol.	r1 r2 r3	x
$tad_{eRHG\text{-}RHGn}$	Time for both the unidirectional instantaneous gene-flow events between eRHG and RHGn	Uniform[1,15000]	all 8 topol.	i1-i2-i3	$t_{RHG}>tad_{eRHG\text{-}RHGn}$ and $t_{KS\text{-}RHGn}>tad_{eRHG\text{-}RHGn}$
m(eRHG-RHGn) or m(RHGn-eRHG)	Instantaneous gene-flow intensity from RHGn into eRHG, or from eRHG into RHGn independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	all 8 topol.	i1 i2 i3	x
m(eRHG-RHGn) or m(RHGn-eRHG)	Recurring gene-flow intensity from RHGn into eRHG, or from eRHG into RHGn independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	all 8 topol.	r1 r2 r3	x
$tad_{ARHG\text{-}RHGn}$	Time for both the unidirectional instantaneous gene-flow events between the RHGn and the lineage ancestral to wRHG and eRHG lineages	Uniform[1,15000]	all 8 topol. except 2c	i1-i2-i3	$tad_{RHGn\text{-}ARHG}>t_{RHG}$
m(RHGn-ARHG) or m(ARHG-RHGn)	Instantaneous gene-flow intensity from the ancestral lineage ARHG into RHGn, or from RHGn into ARHG independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	all 8 topol. except 2c	i1 i2 i3	x
m(RHGn-ARHG) or m(ARHG-RHGn)	Recurring gene-flow intensity from the ancestral lineage ARHG into RHGn, or from RHGn into ARHG independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	all 8 topol. except 2c	r1 r2 r3	x
$tad_{AKS\text{-}RHGn}$	Time for both the unidirectional instantaneous gene-flow events between the lineage ancestral to nKS and sKS, and the RHGn lineage	Uniform[1,15000]	all 8 topol. except 1c	i1-i2-i3	$tad_{AKS\text{-}RHGn}>t_{KS}$
m(RHGn-AKS) or m(AKS-RHGn)	Instantaneous gene-flow intensity from the ancestral lineage AKS into RHGn, or from RHGn into AKS independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	all 8 topol. except 1c	i1 i2 i3	x
m(RHGn-AKS) or m(AKS-RHGn)	Recurring gene-flow intensity from the ancestral lineage AKS into RHGn, or from RHGn into AKS independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	all 8 topol. except 1c	r1 r2 r3	x
$tad_{AKS\text{-}ARHG}$	Time for both the unidirectional instantaneous gene-flow events between the lineage ancestral to nKS and sKS, and the lineage ancestral to wRHG and eRHG lineages	Uniform[1,15000]	all 8 topol. except 1c-2c	i1-i2-i3	$tad_{AKS\text{-}ARHG}>t_{KS}$ and $tad_{AKS\text{-}ARHG}>t_{RHG}$

m(AKS-ARHG) or m(ARHG-AKS)	Instantaneous gene-flow intensity from the ancestral lineage ARHG into the ancestral lineage AKS, or from AKS into ARHG independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	all 8 topol. except 1c-2c	i1 i2 i3	x
m(AKS-ARHG) or m(ARHG-AKS)	Recurring gene-flow intensity from the ancestral lineage ARHG into the ancestral lineage AKS, or from AKS into ARHG independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	all 8 topol. except 1c-2c	r1 r2 r3	x
$tad_{nKS\text{-}ARHG\text{-}RHGn}$	Time for both the unidirectional instantaneous gene-flow events between the nKS and the lineage ancestral to ARHG and RHGn lineages	Uniform[1,15000]	1c	i1-i2-i3	$t_{KS}>tad_{nKS\text{-}ARHG\text{-}RHGn}$ and $tad_{nKS\text{-}ARHG\text{-}RHGn}>t_{RHG\text{-}RHGn}$
m(nKS-ARHG-RHGn) or m(ARHG-RHGn-nKS)	Instantaneous gene-flow intensity from the lineage ARHG-RHGn ancestral to lineages RHGn and ARHG into nKS, or from nKS into ARHG-RHGn independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	1c	i1 i2 i3	x
m(nKS-ARHG-RHGn) or m(ARHG-RHGn-nKS)	Recurring gene-flow intensity from the lineage ARHG-RHGn ancestral to lineages RHGn and ARHG into nKS, or from nKS into ARHG-RHGn independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	1c	r1 r2 r3	x
$tad_{sKS\text{-}ARHG\text{-}RHGn}$	Time for both the unidirectional instantaneous gene-flow events between the sKS and the lineage ancestral to ARHG and RHGn lineages	Uniform[1,15000]	1c	i1-i2-i3	$t_{KS}>tad_{sKS\text{-}ARHG\text{-}RHGn}$ and $tad_{sKS\text{-}ARHG\text{-}RHGn}>t_{RHG\text{-}RHGn}$
m(sKS-ARHG-RHGn) or m(ARHG-RHGn-sKS)	Instantaneous gene-flow intensity from the lineage ARHG-RHGn ancestral to lineages RHGn and ARHG into sKS, or from sKS into ARHG-RHGn independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	1c	i1 i2 i3	x
m(sKS-ARHG-RHGn) or m(ARHG-RHGn-sKS)	Recurring gene-flow intensity from the lineage ARHG-RHGn ancestral to lineages RHGn and ARHG into sKS, or from sKS into ARHG-RHGn independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	1c	r1 r2 r3	x
$tad_{AKS\text{-}RHGn\text{-}wRHG}$	Time for both the unidirectional instantaneous gene-flow events between the wRHG and the lineage ancestral to AKS and RHGn lineages	Uniform[1,15000]	2c	i1-i2-i3	$t_{RHG}>tad_{AKS\text{-}RHGn\text{-}wRHG}$ and $tad_{AKS\text{-}RHGn\text{-}wRHG}>t_{KS\text{-}RHGn}$
m(wRHG-AKS-RHGn) or m(AKS-RHGn--wRHG)	Instantaneous gene-flow intensity from the lineage AKS-RHGn ancestral to lineages AKS and RHGn into wRHG, or from wRHG into AKS-RHGn independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	2c	i1 i2 i3	x
m(wRHG-AKS-RHGn) or m(AKS-RHGn--wRHG)	Recurring gene-flow intensity from the lineage AKS-RHGn ancestral to lineages AKS and RHGn into wRHG, or from wRHG into AKS-RHGn independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	2c	r1 r2 r3	x
$tad_{AKS\text{-}RHGn\text{-}eRHG}$	Time for both the unidirectional instantaneous gene-flow events between the eRHG and the lineage ancestral to AKS and RHGn lineages	Uniform[1,15000]	2c	i1-i2-i3	$t_{RHG}>tad_{AKS\text{-}RHGn\text{-}eRHG}$ and $tad_{AKS\text{-}RHGn\text{-}eRHG}>t_{KS\text{-}RHGn}$
m(eRHG-AKS-RHGn) or m(AKS-RHGn--eRHG)	Instantaneous gene-flow intensity from the lineage AKS-RHGn ancestral to lineages AKS and RHGn into eRHG, or from eRHG into AKS-RHGn independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	2c	i1 i2 i3	x
m(eRHG-AKS-RHGn) or m(AKS-RHGn--eRHG)	Recurring gene-flow intensity from the lineage AKS-RHGn ancestral to lineages AKS and RHGn into eRHG, or from eRHG into AKS-RHGn independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	2c	r1 r2 r3	x
$tad_{AKS\text{-}ARHG\text{-}RHGn}$	Time for both the unidirectional instantaneous gene-flow events between the lineage ancestral to KS and the lineage ancestral to RHG and RHGn lineages	Uniform[1,15000]	1a-1b-1c	i1-i2-i3	$tad_{AKS\text{-}ARHG\text{-}RHGn}>t_{RHG\text{-}RHGn}$ and $tad_{AKS\text{-}ARHG\text{-}RHGn}>t_{KS}$ and $t_{KS\text{-}RHG\text{-}RHGn}>tad_{AKS\text{-}ARHG\text{-}RHGn}$
m(AKS-ARHG-RHGn) or m(ARHG-RHGn-AKS)	Instantaneous gene-flow intensity from the lineage ARHG-RHGn ancestral to lineages ARHG and RHGn into the ancestral lineage AKS, or from AKS into ARHG-RHGn independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	1a-1b-1c	i1 i2 i3	x
m(AKS-ARHG-RHGn) or m(ARHG-RHGn-AKS)	Recurring gene-flow intensity from the lineage ARHG-RHGn ancestral to lineages ARHG and RHGn into the ancestral lineage AKS, or from AKS into ARHG-RHGn independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	1a-1b-1c	r1 r2 r3	x
$tad_{AKSRHGn\text{-}ARHG}$	Time for both the unidirectional instantaneous gene-flow events between the lineage ancestral to KS and RHGn, and the lineage ancestral to RHG lineages	Uniform[1,15000]	2a-2b-2c	i1-i2-i3	$tad_{AKSRHGn\text{-}ARHG}>t_{KS\text{-}RHGn}$ and $tad_{AKSRHGn\text{-}ARHG}>t_{RHG}$ and $t_{KS\text{-}RHG\text{-}RHGn}>tad_{AKSRHGn\text{-}ARHG}$
m(AKSRHGn-ARHG) or m(ARHG-AKSRHGn)	Instantaneous gene-flow intensity from the ancestral lineage ARHG into the lineage AKSRHGn ancestral to AKS and RHGn, or from AKSRHGn into ARHG independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	2a-2b-2c	i1 i2 i3	x
m(AKSRHGn-ARHG) or m(ARHG-AKSRHGn)	Recurring gene-flow intensity from the ancestral lineage ARHG into the lineage AKSRHGn ancestral to AKS and RHGn, or from AKSRHGn into ARHG independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	2a-2b-2c	r1 r2 r3	x
$tad_{AKSARHG\text{-}RHGn}$	Time for both the unidirectional instantaneous gene-flow events between the lineage ancestral to KS and RHG, and the RHGn lineage	Uniform[1,15000]	3a-3b	i1-i2-i3	$tad_{AKSARHG\text{-}RHGn}>t_{KS\text{-}RHG}$ and $t_{KS\text{-}RHG\text{-}RHGn}>tad_{AKSARHG\text{-}RHGn}$
m(AKSARHG-RHGn) or m(RHGn-AKSARHG)	Instantaneous gene-flow intensity from the lineage RHGn into the lineage AKSRHG ancestral to AKS and ARHG, or from AKSRHG into RHGn independantly	Uniform[0,1] Uniform[0,0.125] Uniform[0,0.01]	3a-3b	i1 i2 i3	x
m(AKSARHG-RHGn) or m(RHGn-AKSARHG)	Recurring gene-flow intensity from the lineage RHGn into the lineage AKSRHG ancestral to AKS and ARHG, or from AKSRHG into RHGn independantly	Uniform[0,0.05] Uniform[0,0.0125] Uniform[0,0.001]	3a-3b	r1 r2 r3	x
N_{nKS}	Effective population size for the nKS lineage	Uniform[10,100000]	all 8 topol.	i1-i2-i3 and r1-r2-r3	x
N_{sKS}	Effective population size for the sKS lineage	Uniform[10,100000]	all 8 topol.	i1-i2-i3 and r1-r2-r3	x
N_{AKS}	Effective population size for the lineage ancestral to the nKS and the sKS	Uniform[10,100000]	all 8 topol.	i1-i2-i3 and r1-r2-r3	x
N_{RHGn}	Effective population size for the RHGn lineage	Uniform[10,100000]	all 8 topol.	i1-i2-i3 and r1-r2-r3	x
N_{wRHG}	Effective population size for the wRHG lineage	Uniform[10,100000]	all 8 topol.	i1-i2-i3 and r1-r2-r3	x
N_{eRHG}	Effective population size for the eRHG lineage	Uniform[10,100000]	all 8 topol.	i1-i2-i3 and r1-r2-r3	x
N_{ARHG}	Effective population size for the lineage ancestral to the wRHG and the eRHG	Uniform[10,100000]	all 8 topol.	i1-i2-i3 and r1-r2-r3	x
$N_{ARHG\text{-}RHGn}$	Effective population size for the lineage ancestral to the RHGn and the ARHG	Uniform[10,100000]	1a-1b-1c	i1-i2-i3 and r1-r2-r3	x

$N_{AKS\text{-}RHGn}$	Effective population size for the lineage ancestral to the RHGn and the AKS		2a-2b-2c	i1-i2-i3 and r1-r2-r3	x
$N_{AKS\text{-}ARHG}$	Effective population size for the lineage ancestral to the AKS and the ARHG		3a-3b	i1-i2-i3 and r1-r2-r3	x
N_A	Effective population size for the lineage ancestral to all lineages		all 8 topol.	i1-i2-i3 and r1-r2-r3	x
Mutation rate	Fixed mutation rate	$1.25*10^{-8}$	all 8 topol.	i1-i2-i3 and r1-r2-r3	x
Recombination rate	Fixed recombination rate	$1*10^{-8}$	all 8 topol.	i1-i2-i3 and r1-r2-r3	x
Transition to transversion ratio	Fixed transition to transversion rate	0.33	all 8 topol.	i1-i2-i3 and r1-r2-r3	x

2254 **<u>TableS9</u>: Summary-statistics used in Random-Forest ABC scenario-choice and Neural-Network ABC posterior-**
2255 **parameter estimation**.
2256

Name of statistic (or group of statistics)	What was computed	Number of values	RF-ABC scenario-choice	NN-ABC posterior-parameter estimation
Total number of biallelic sites among the five populations	sum	1	yes	yes
Total number of multiallelic sites among the five populations	sum	1	yes	yes
ASD distance within population for each five population separately	average and variance	10	yes	no
First five dimensions of the projection of the ASD matrix within population for each five population separately	average and variance	50	yes	yes
Proportion of biallelic sites within each five population separately	sum in a population divided by total number of biallelic sites in the dataset (no missing data)	5	yes	yes
Proportion of homozygous sites within each five population separately	sum in a population divided by total number of biallelic sites in the dataset (no missing data)	5	yes	yes
Proportion of homozygous ancestral sites within each five population separately	sum in a population divided by total number of homozygous sites in the population (no missing data)	5	yes	no
Proportion of private biallelic sites within each five population separately	sum in a population divided by total number of biallelic sites in the dataset (no missing data)	5	yes	yes
Frequency of the minor allele of the private biallelic sites within each five population separately	average and variance	10	yes	no
Expected heterozygosity at a segregating site within each five population separately	average and variance	10	yes	no
Tajima's D within each five population separately	average and variance for chromosomes with at least one variant	10	yes	yes
Unfolded site frequency spectrum (SFS) within each five population separately	proportion for each of 9 classes	45	yes	no
Watterson's theta within each five population separately	sum of biallelic sites divided by the harmonic number of the sample size (n=10)	5	yes	no
Total number of runs of homozygosity (ROH) within each five population separately	sum	5	yes	no
Total number of ROH by length class (<200 kbp, 200-500 kbp, >500 kbp) within each five population separately	sum	15	yes	yes
ROH length for each length class within each five population separately	average	15	yes	no
Total ROH length within each five population separately	average and variance	10	yes	no
ASD distance between each pair of five populations	average and variance	20	yes	no
First five dimensions of the projection of the ASD matrix between each pair of five populations	average and variance	100	yes	yes
Pairwise FST between each pair of five population	Pairwise FST (Weir and Cockerham 1984)	10	yes	yes
	TOTAL number of statistics used in ABC inference		337	202

2257 ----------

TableS10: Correlations between standardized ASD matrices for each one of the 54 combinations of five population samples computed using 100 separate high-quality 1 Mb-long genomic windows, and corresponding individual-subsets of the standardized ASD matrix computed using the full genomic SNP set and used for Figure3.

Combinations of five populations used for ABC inferences								
nKS	**sKS**	**RHGn**	**wRHG**	**eRHG**	***Spearman_ρ***	**1000 Mantel prem. p**	***Pearson_r***	**1000 Mantel prem. p**
Ju|'hoansi	Karretjie_People	Ba.Konjo	Ba.Kola	Mbuti	0.893	<0.002	0.962	<0.002
Ju|'hoansi	Karretjie_People	Ba.Konjo	Ba.Kola	Nsua	0.896	<0.002	0.963	<0.002
Ju|'hoansi	Karretjie_People	Ba.Konjo	Baka	Mbuti	0.898	<0.002	0.964	<0.002
Ju|'hoansi	Karretjie_People	Ba.Konjo	Baka	Nsua	0.903	<0.002	0.967	<0.002
Ju|'hoansi	Karretjie_People	Ba.Konjo	Bi.Aka_Mbati	Mbuti	0.894	<0.002	0.963	<0.002
Ju|'hoansi	Karretjie_People	Ba.Konjo	Bi.Aka_Mbati	Nsua	0.895	<0.002	0.965	<0.002
Ju|'hoansi	Karretjie_People	Ba.Kiga	Ba.Kola	Ba.Twa	0.901	<0.002	0.961	<0.002
Ju|'hoansi	Karretjie_People	Ba.Kiga	Baka	Ba.Twa	0.908	<0.002	0.965	<0.002
Ju|'hoansi	Karretjie_People	Ba.Kiga	Bi.Aka_Mbati	Ba.Twa	0.887	<0.002	0.961	<0.002
Ju|'hoansi	Karretjie_People	Ngumba	Ba.Kola	Ba.Twa	0.885	<0.002	0.960	<0.002
Ju|'hoansi	Karretjie_People	Ngumba	Ba.Kola	Mbuti	0.907	<0.002	0.963	<0.002
Ju|'hoansi	Karretjie_People	Ngumba	Ba.Kola	Nsua	0.910	<0.002	0.964	<0.002
Ju|'hoansi	Karretjie_People	Nzime	Baka	Ba.Twa	0.914	<0.002	0.968	<0.002
Ju|'hoansi	Karretjie_People	Nzime	Baka	Mbuti	0.929	<0.002	0.971	<0.002
Ju|'hoansi	Karretjie_People	Nzime	Baka	Nsua	0.935	<0.002	0.973	<0.002
Ju|'hoansi	Karretjie_People	Nzime	Bi.Aka_Mbati	Ba.Twa	0.896	<0.002	0.965	<0.002
Ju|'hoansi	Karretjie_People	Nzime	Bi.Aka_Mbati	Mbuti	0.925	<0.002	0.970	<0.002
Ju|'hoansi	Karretjie_People	Nzime	Bi.Aka_Mbati	Nsua	0.927	<0.002	0.970	<0.002
Ju|'hoansi	Nama	Ba.Konjo	Ba.Kola	Nsua	0.932	<0.002	0.966	<0.002
Ju|'hoansi	Nama	Ba.Konjo	Baka	Nsua	0.948	<0.002	0.969	<0.002
Ju|'hoansi	Nama	Ba.Konjo	Bi.Aka_Mbati	Nsua	0.944	<0.002	0.967	<0.002
Ju|'hoansi	Nama	Ba.Kiga	Ba.Kola	Ba.Twa	0.948	<0.002	0.964	<0.002
Ju|'hoansi	Nama	Ba.Kiga	Baka	Ba.Twa	0.958	<0.002	0.966	<0.002
Ju|'hoansi	Nama	Ba.Kiga	Bi.Aka_Mbati	Ba.Twa	0.945	<0.002	0.962	<0.002
Ju|'hoansi	Nama	Ngumba	Ba.Kola	Ba.Twa	0.935	<0.002	0.962	<0.002
Ju|'hoansi	Nama	Ngumba	Ba.Kola	Nsua	0.945	<0.002	0.965	<0.002
Ju|'hoansi	Nama	Nzime	Baka	Ba.Twa	0.957	<0.002	0.970	<0.002
Ju|'hoansi	Nama	Nzime	Baka	Nsua	0.963	<0.002	0.974	<0.002
Ju|'hoansi	Nama	Nzime	Bi.Aka_Mbati	Ba.Twa	0.945	<0.002	0.965	<0.002
Ju|'hoansi	Nama	Nzime	Bi.Aka_Mbati	Nsua	0.945	<0.002	0.971	<0.002
!Xun	Karretjie_People	Ba.Konjo	Ba.Kola	Nsua	0.927	<0.002	0.952	<0.002
!Xun	Karretjie_People	Ba.Konjo	Baka	Nsua	0.936	<0.002	0.957	<0.002
!Xun	Karretjie_People	Ba.Konjo	Bi.Aka_Mbati	Nsua	0.934	<0.002	0.956	<0.002
!Xun	Karretjie_People	Ba.Kiga	Ba.Kola	Ba.Twa	0.923	<0.002	0.949	<0.002
!Xun	Karretjie_People	Ba.Kiga	Baka	Ba.Twa	0.936	<0.002	0.954	<0.002
!Xun	Karretjie_People	Ba.Kiga	Bi.Aka_Mbati	Ba.Twa	0.922	<0.002	0.950	<0.002
!Xun	Karretjie_People	Ngumba	Ba.Kola	Ba.Twa	0.913	<0.002	0.948	<0.002
!Xun	Karretjie_People	Ngumba	Ba.Kola	Nsua	0.930	<0.002	0.952	<0.002
!Xun	Karretjie_People	Nzime	Baka	Ba.Twa	0.945	<0.002	0.959	<0.002
!Xun	Karretjie_People	Nzime	Baka	Nsua	0.956	<0.002	0.963	<0.002
!Xun	Karretjie_People	Nzime	Bi.Aka_Mbati	Ba.Twa	0.931	<0.002	0.955	<0.002
!Xun	Karretjie_People	Nzime	Bi.Aka_Mbati	Nsua	0.935	<0.002	0.962	<0.002
!Xun	Nama	Ba.Konjo	Ba.Kola	Nsua	0.932	<0.002	0.954	<0.002
!Xun	Nama	Ba.Konjo	Baka	Nsua	0.945	<0.002	0.957	<0.002
!Xun	Nama	Ba.Konjo	Bi.Aka_Mbati	Nsua	0.937	<0.002	0.956	<0.002
!Xun	Nama	Ba.Kiga	Ba.Kola	Ba.Twa	0.933	<0.002	0.948	<0.002
!Xun	Nama	Ba.Kiga	Baka	Ba.Twa	0.942	<0.002	0.951	<0.002
!Xun	Nama	Ba.Kiga	Bi.Aka_Mbati	Ba.Twa	0.920	<0.002	0.947	<0.002
!Xun	Nama	Ngumba	Ba.Kola	Ba.Twa	0.910	<0.002	0.945	<0.002
!Xun	Nama	Ngumba	Ba.Kola	Nsua	0.930	<0.002	0.950	<0.002
!Xun	Nama	Nzime	Baka	Ba.Twa	0.940	<0.002	0.956	<0.002
!Xun	Nama	Nzime	Baka	Nsua	0.954	<0.002	0.962	<0.002
!Xun	Nama	Nzime	Bi.Aka_Mbati	Ba.Twa	0.915	<0.002	0.951	<0.002
!Xun	Nama	Nzime	Bi.Aka_Mbati	Nsua	0.919	<0.002	0.959	<0.002

2263 **<u>TableS11</u>: 25 combinations of five population samples out of 54 combinations tested for which Scenario i1-1b is best**
2264 **explaining observed data with Random-Forest ABC among the 48 competing scenarios.**
2265 Values in the table correspond to posterior densities plotted in the left panels of **Figure7**, **Figure8**, **FigureS18**, and **FigureS19**.
2266 Parameter definitions and priors are provided in **TableS8** and represented graphically in the corresponding scenario panel of
2267 **Figure5**. Values in italic are not satisfactorily departing from the priors.
2268

	nKS	**sKS**	**RHGn**	**wRHG**	**eRHG**
Combination 1	Ju|'hoansi	Karretjie People	Nzime	Baka	Nsua
Combination 2	Ju|'hoansi	Karretjie People	Ngumba	Ba.Kola	Nsua
Combination 3	Ju|'hoansi	Karretjie People	Ngumba	Ba.Kola	Ba.Twa
Combination 4	Ju|'hoansi	Karretjie People	Nzime	Bi.Aka_Mbati	Nsua
Combination 5	Ju|'hoansi	Karretjie People	Nzime	Bi.Aka_Mbati	Ba.Twa
Combination 6	Ju|'hoansi	Karretjie People	Ngumba	Ba.Kola	Mbuti
Combination 7	Ju|'hoansi	Karretjie People	Nzime	Bi.Aka_Mbati	Mbuti
Combination 8	Ju|'hoansi	Karretjie People	Ba.Konjo	Baka	Mbuti
Combination 9	Ju|'hoansi	Karretjie People	Ba.Konjo	Ba.Kola	Mbuti
Combination 10	Ju|'hoansi	Karretjie People	Ba.Konjo	Bi.Aka_Mbati	Mbuti
Combination 11	Ju|'hoansi	Nama	Ba.Konjo	Baka	Nsua
Combination 12	!Xun	Nama	Ba.Konjo	Baka	Nsua
Combination 13	Ju|'hoansi	Nama	Nzime	Baka	Nsua
Combination 14	Ju|'hoansi	Nama	Ngumba	Ba.Kola	Nsua
Combination 15	Ju|'hoansi	Nama	Nzime	Baka	Ba.Twa
Combination 16	Ju|'hoansi	Nama	Ngumba	Ba.Kola	Ba.Twa
Combination 17	Ju|'hoansi	Nama	Nzime	Bi.Aka_Mbati	Ba.Twa
Combination 18	Ju|'hoansi	Nama	Ba.Konjo	Ba.Kola	Nsua
Combination 19	Ju|'hoansi	Nama	Ba.Kiga	Ba.Kola	Ba.Twa
Combination 20	Ju|'hoansi	Nama	Ba.Kiga	Bi.Aka_Mbati	Ba.Twa
Combination 21	!Xun	Nama	Nzime	Bi.Aka_Mbati	Nsua
Combination 22	!Xun	Nama	Nzime	Bi.Aka_Mbati	Ba.Twa
Combination 23	!Xun	Karretjie People	Nzime	Bi.Aka_Mbati	Nsua
Combination 24	!Xun	Karretjie People	Nzime	Bi.Aka_Mbati	Ba.Twa
Combination 25	!Xun	Karretjie People	Ba.Konjo	Bi.Aka_Mbati	Nsua

2269
2270

2271 **<u>TableS12</u>: NN-ABC posterior parameter estimation of all parameters in Scenario i1-1b for results from each 25 sets of**
2272 **five populations each, separately.**
2273 Each sampled-population combination is provided as a single line. Table-headers in bold indicate the categorization of sampled
2274 populations into Northern Khoisan (nKS), Southern Khoisan (sKS), Rainforest Hunter-Gatherer neighbors (RHGn), Western
2275 Rainforest Hunter-Gatherer (wRHG), or Eastern Rainsforest Hunter-Gatherer (eRHG) groups, as per scenario topology and
2276 simulation design explicated in **Material and Methods** and in **Figure5**. Note that each sampled population considered in the
2277 RF-ABC scenario choice analyses are represented twice or more in the 25 combinations.
2278
2279 See "**TableS12.xlsx**" file as the table is to large to reasonably fit in a A4 page format.
2280
2281

<u>TableS13</u>: NN-ABC posterior parameter estimation cross-validation errors of all parameters in Scenario i1-1b for results from 25 sets of five populations each.

Cross-validation errors are based on 50 (5%) randomly chosen simulation out of the 1000 simulations closest to the observed data for each 25 combination of sampled populations, respectively, each considered in-turn as pseudo-observed data and for which NN-ABC joint posterior parameter estimations were performed with the same hyper-parameters. See **Material and Methods** for additional information. Relatively high errors (close-to or above 1) are highlighted in grey.

See "**TableS13.xlsx**" file as the table is to large to reasonably fit in a A4 page format.

TableS14: NN-ABC posterior parameter estimation of all parameters in Scenario i1-1b for results from 25 sets of five populations each, combined altogether.

Entire posterior distributions for each 25 analyses separately, for which summary statistics are presented in **TableS3**, are combined together before re-computing the Mode, Median, Mean, 50%CI and 90%CI presented in this table. Therefore, values in the table correspond to posterior densities plotted in the right panels of **Figure7**, **Figure8**, **FigureS18**, and **FigureS19**. Parameter definitions and priors are provided in **TableS8** and represented graphically in the corresponding scenario panel of **Figure5**. Values in this table are used for the synthetic schematic representation of the demographic and migration history of Central and Southern African populations presented in **Figure9**. Values in italic are not satisfactorily departing from the priors and are associated with high cross-validation posterior parameter estimation errors in the vicinity of the observed data (**TableS13**).

Parameter	Mode	Median	Mean	50% CI	90% CI
N_{nKS}	**14184**	**23881**	**29908**	**14721-39894**	**7825-73423**
N_{sKS}	**11042**	**19672**	**26050**	**11059-35737**	**4623-68767**
N_{wRHG}	**9461**	**20586**	**26551**	**11506-36934**	**5074-67317**
N_{eRHG}	**12592**	**15735**	**19968**	**10059-24368**	**3993-53065**
N_{RHGn}	**18022**	**28739**	**33965**	**18076-45474**	**9150-77157**
N_{AKS}	*86572*	*61359*	*59978*	*41254-80294*	*19523-94743*
N_{ARHG}	*57397*	*55256*	*54989*	*36771-73803*	*16922-91847*
$N_{ARHG\text{-}RHGn}$	*46968*	*48438*	*48585*	*30042-66886*	*9894-88177*
N_{A}	**18507**	**24995**	**28406**	**17135-35537**	**9399-60616**
t_{KS}	**2623**	**2960**	**3199**	**2287-3921**	**1481-5763**
t_{RHG}	**892**	**1111**	**1255**	**773-1574**	**422-2531**
$t_{RHG\text{-}RHGn}$	**6726**	**6729**	**6931**	**5393-8271**	**3876-10748**
$t_{KS\text{-}RHG\text{-}RHGn}$	**12117**	**12152**	**12001**	**10797-13422**	**8875-14538**
$tad_{nKS\text{-}sKS}$	**395**	**532**	**597**	**347-786**	**147-1232**
$tad_{nKS\text{-}RHGn}$	**1383**	**1659**	**1835**	**1140-2333**	**591-3710**
$tad_{sKS\text{-}RHGn}$	**1142**	**1378**	**1533**	**928-1970**	**418-3165**
$tad_{wRHG\text{-}RHGn}$	**302**	**348**	**418**	**227-550**	**97-944**
$tad_{eRHG\text{-}RHGn}$	**393**	**520**	**616**	**342-801**	**171-1358**
$tad_{wRHG\text{-}eRHG}$	**411**	**555**	**653**	**367-851**	**157-1429**
$tad_{RHGn\text{-}ARHG}$	**2756**	**3193**	**3523**	**2352-4347**	**1369-6828**
$tad_{AKS\text{-}RHGn}$	**4490**	**4588**	**4831**	**3731-5590**	**2828-7892**
$tad_{AKS\text{-}ARHG}$	**4408**	**4458**	**4730**	**3629-5498**	**2658-7963**
$tad_{AKS\text{-}ARHG\text{-}RHGn}$	**9679**	**9532**	**9524**	**7888-11066**	**6358-12919**
m(nKS-sKS)	*0.6727*	*0.6018*	*0.5685*	*0.3770-0.7830*	*0.0892-0.9385*
m(sKS-nKS)	*0.7330*	*0.6314*	*0.6072*	*0.4481-0.7911*	*0.1851-0.9346*
m(nKS-RHGn)	*0.4827*	*0.5336*	*0.5324*	*0.3595-0.7206*	*0.1277-0.9082*
m(RHGn-nKS)	*0.3668*	*0.4842*	*0.4966*	*0.2930-0.7092*	*0.0875-0.9153*
m(sKS-RHGn)	*0.4810*	*0.5093*	*0.5082*	*0.3190-0.6977*	*0.1062-0.9089*
m(RHGn-sKS)	*0.4920*	*0.4525*	*0.4587*	*0.2549-0.6432*	*0.0771-0.8880*
m(RHGn-wRHG)	*0.6461*	*0.5674*	*0.5482*	*0.3609-0.7438*	*0.1142-0.9258*
m(wRHG-RHGn)	*0.3925*	*0.4341*	*0.4447*	*0.2729-0.6058*	*0.0873-0.8413*
m(RHGn-eRHG)	*0.3305*	*0.4140*	*0.4365*	*0.2319-0.6291*	*0.0618-0.8750*
m(eRHG-RHGn)	*0.6072*	*0.5726*	*0.5507*	*0.3583-0.7577*	*0.1073-0.9243*
m(wRHG-eRHG)	*0.5034*	*0.4724*	*0.4766*	*0.2736-0.6758*	*0.0802-0.8891*
m(eRHG-wRHG)	*0.5453*	*0.5396*	*0.5340*	*0.3291-0.7465*	*0.1069-0.9388*
m(RHGn-ARHG)	*0.5802*	*0.4897*	*0.4851*	*0.2797-0.6870*	*0.0781-0.8904*
m(ARHG-RHGn)	*0.4081*	*0.5080*	*0.5110*	*0.2993-0.7299*	*0.0853-0.9345*
m(ARHG-AKS)	*0.4734*	*0.5082*	*0.5056*	*0.3033-0.7097*	*0.0850-0.9212*
m(AKS-ARHG)	*0.5824*	*0.5220*	*0.5121*	*0.3036-0.7210*	*0.0773-0.9255*
m(RHGn-AKS)	*0.7032*	*0.5235*	*0.5139*	*0.3053-0.7261*	*0.0887-0.9158*
m(AKS-RHGn)	*0.4066*	*0.5027*	*0.5066*	*0.3136-0.7197*	*0.0929-0.8935*
m(AKS-ARHG-RHGn)	*0.5357*	*0.5008*	*0.4916*	*0.3017-0.6756*	*0.0834-0.8904*
m(ARHG-RHGn-AKS)	*0.3359*	*0.4501*	*0.4680*	*0.2424-0.6856*	*0.0663-0.9148*

2303