Science Advances

AAAS

Supplementary Materials for

Dense sampling of ethnic groups within African countries reveals fine-scale genetic structure and extensive historical admixture.

Nancy Bird[1]*, Louise Ormond[1], Paschal Awah[2], Elizabeth F. Caldwell[3], Bruce Connell[4], Mohamed Elamin[5], Faisal M. Fadlelmola[6], Forka Leypey Matthew Fomine[7], Saioa López[8], Scott MacEachern[9], Yves Moñino[10], Sam Morris[11], Pieta Näsänen-Gilmore[12,13], Nana Nketsia[14], Krishna Veeramah[15], Michael E. Weale[16], David Zeitlyn[17], Mark G. Thomas[1], Neil Bradman[18], Garrett Hellenthal[1]

*****Corresponding author. Email: nancy.bird.18@ucl.ac.uk

This PDF file includes:

Supplementary Text 1 to 4
Figs. S1 to S21

Other Supplementary Materials for this manuscript include the following:

Data S1 to S8
Fig. S6

Supplementary Text 1

Brief description of newly reported samples from each country

Ghana

The project includes newly reported data from 211 individuals sampled in Ghana and representing 21 ethnic groups, 16 of which have a sample size of more than three individuals. Ethnic groups represented by data from more than 15 sampled individuals include: Dagaati, Frafra, Kusaasi, Bulsa and Wali. All included ethnic groups speak a Niger-Congo language. They are divided into two main linguistic branches: North Volta-Congo speakers and Kwa Volta-Congo speakers from the north and south of the country respectively (*33*). Within this, all North Volta-Congo speakers speak a Gur language, with the majority of these being Northern Central Gur languages. The exceptions are Sisaala and Kasem which are both Southern Central Gur. Of the Kwa Volta-Congo speakers, the majority of individuals speak a Nyo language, while one ethnic group, the Ewe speak a Gbe language.

Nigeria

Newly reported data was generated from 311 Nigerian individuals, from 6 different ethnic groups. These include: Anang, Efik, Ekoi, Ibibio, Igbo, and Oron. All ethnic groups have data from more than 15 individuals, with the largest group, Ibibio, having data from 161 individuals. All groups are from the South South and Southeast regions of Nigeria, with the majority sampled in the Cross River State. All ethnic groups speak a Niger-Congo language. Within this, most ethnic groups speak a Delta Cross language, while the Ekoi speak a Bantoid language and the Igbo speak an Igboid language.

Republic of the Congo

Newly reported genetic data was generated from 122 individuals from 17 ethnic groups sampled in the Republic of the Congo. Of these, 11 groups have a sample size of more than three individuals, but only one, the Yombe, has more than 15 individuals. All ethnic groups speak a Central-Western Bantu language. Of these, all are West-Coastal Bantu speakers except the Mboshi, who speak North Zaire River Bantu. Within the West-Coastal Bantu speakers, ethnic groups are divided into Njebe-Mbete-Teke language speakers (Koukouya, Lali, Mbere and Teke) and Kongo-Yaka-Suku-Hungan speakers (all other groups). The majority of samples have birthplaces in the southern half of the country.

Cameroon

Genetic data from Cameroon were obtained from individuals sampled as part of a North-South transect. The data contains newly reported genotype data from 484 Cameroonians from 77 different ethnic groups. 30 of these groups have data from more than three individuals, with nine represented by more than 15 individuals: Kotoko, Mambila, Noni, Bamun, Tikar, Yamba, Nso, Wimbum and Kwandja. Sampled groups include speakers of three main African language families; Afro-Asiatic, Nilo-Saharan, and Niger-Congo. The former two all inhabit the north of the country. There are two main branches of Nilo-Saharan speakers. Firstly, the Kanuri speak a Saharan language, and secondly the Sara and Ngambai are Central Sudanic speakers. Of the Afro-Asiatic speakers, one ethnic group, the Arabe, speak Arabic. All other Afro-Asiatic speakers have languages that are members of the Chadic branch. Most notably, these include Guidar, Kotoko, Mandara, Massa and Mousgoum.

The Niger-Congo speakers primarily inhabit the southern half of Cameroon. An exception is Fulani, who speak a North-Central Atlantic language (primarily found in Senegal). Another exception are Ubangian speakers, such as Doowaayo, Moundang, and Toupouri, who also live in the northern half of Cameroon. The remaining Niger-Congo speakers are then divided into two main classifications; Northern Bantoid and Southern Bantoid. Northern Bantoid speakers live primarily in the Adamawa region near the border with Nigeria, and include the ethnic groups Kwandja, Mambila and Tikar. Southern Bantoid speakers are again divided into two main groups: Wide Grassfields and Narrow Bantu. Wide Grassfields speakers primarily inhabit the Northwest and West regions of Cameroon and include the Bamileke, Bamun, Nso', Wimbum and Yamba. Lastly, the Narrow Bantu speakers include Bassa, Ewondo, Mbo and Ngoumba. The Narrow Bantu speakers in our dataset inhabit the Southwest, South and Littoral regions. Narrow Bantu is the language family spoken by roughly 30% of the African population as a result of the expansion of Bantu-speaking peoples. It originated in the Nigeria/Cameroon border region (*76*). However, there is much debate over the classification of Wide Grassfields and Narrow Bantu languages (*52*).

Sudan

The project includes data from 233 individuals sampled in Sudan representing 31 ethnic groups. 17 of these groups have data from more than 3 individuals, with 4 having data from 15 or more individuals: Ja'aliya, Halfawieen, Beni-Amer and Rubatab. Sudanese individuals were primarily sampled as part of a Nile transect, where samples were collected at different cities and towns along the river (n=144, samples from Abu Hamad, Atbara, Dongola, Ad Douiem, Kosti, Madani, Shendi and Wadi Halfa). Other samples (mostly from Beni-Amer) were collected at the coast in Port Sudan. Lastly, samples were collected from individuals in the South Kordofan region, near the border with South Sudan. The majority of the sampled Sudanese individuals speak Arabic. Within Arabic-speakers, there are two separate ethnicities related to the pre-Arabic language spoken: Beja, and Nubian. Beja ethnic groups include Beni-Amer, Hadendowa and Halenga. Nubian ethnic groups include Danagla and Halfawieen. Ethnic groups represented in this study from the south Kordofan region speak either Nilo-Saharan or Kordofanian languages. Within Nilo-Saharan speakers, the Lagori speak a Dajuic language and the Kadugli, Keiga and Korongo speak a Kadugli-Krongo language. Within Kordofanian languages, the Acheron speak a Talodi language and the Moro speak a Heibanic language. We have grouped the Kordofanian languages into one phyla for this paper, although there is much debate about whether they constitute a single branch and some evidence indicates it may be a group of language isolates (*36*).

Note: Languages were classified using Glottolog for all countries except Sudan, where Ethnologue was used. This is because we inferred genetic differentiation between languages from the south Kordofan region classified as Nilo-Saharan and Kordofanian in Ethnologue. This level of classification is not included in Glottolog, as it is more disputed.

Supplementary Text 2

<u>Brief description of the historical polities mentioned in the text</u>

<u>Grassfields of Cameroon</u>

The Grassfields are situated in broadly the North-West and West regions of Cameroon. The region is characterised by mountainous terrain and Sudanian savannah vegetation. It is home to many different ethnolinguistic groups living in close proximity (*35*). Historically, these groups have adopted a number of different political systems, and have spheres of influence of varying sizes. Some groups, such as the Bamun, 'Nso, and Bafut had large polities in pre-colonial times, and were involved in both trade and conflict with other nearby groups (*62*).

<u>Makuria and the fall of Dongola</u>

Makuria was a Christian Nubian Kingdom established in the 4th century CE. It was located along the Nile, and at its height extended from the area around modern Khartoum to north of Luxor in what is now Egypt. Its capital was the city of Dongola (*67*). Arabs expanded into Egypt in the 7th century but formed a peace treaty with Makuria, preventing further expansion along the Nile (*91*). However, the kingdom began to decline from the 13th century onwards. The history of the Christian Makurian state from the 14th century onward is complex, with repeated Mamluk and Bedouin invasions and civil strife, and it had disappeared by the early 16th century.

<u>Kingdom of Aksum</u>

Also known as the Axumite Empire, this kingdom was regionally dominant from 100CE to about 800CE. It originated in northern Ethiopia, although as it expanded it incorporated parts of modern Eritrea and eastern Sudan. At the height of the empire in the 500CEs it also incorporated part of the southern Arabian peninsula. It was involved in trade networks linking the Roman Empire to the Middle East and India, and had a strong navy. The empire was linguistically and ethnically complex, although Ge'ez, an Afro-Asiatic Semitic language, was the official language of imperial rule and the Church (*68*). The period up to the peak of the empire was characterized by population growth (*92*).

<u>Kanem-Bornu Empire</u>

The Kanem-Bornu Empire existed in two phases. From the eighth century CE to approximately 1380CE, the political center of gravity of the state existed northeast of Lake Chad and the state was dominated by Nilo-Saharan-speaking Kanembu people. After 1380CE, the empire gradually moved its centre to Bornu in Nigeria, southwest of Lake Chad, traditionally held to be a result of attacks from Bulala people. This movement led to interactions with, and assimilation of, local Chadic-speaking people under the collective name of Sao, who lived in northern Cameroon and Nigeria. The merging of Nilo-Saharan-speaking people from the empire, and local Sao and related groups, gave rise to the Kanuri ethnic group. The empire covered parts of Niger, Cameroon, Central Africa Republic, Chad and Nigeria at different times and was an important trading centre. It was a natural connecting point for trans-Saharan and Sudanic trade routes (*38, 71*).

<u>Fulani Jihad and Sokoto Caliphate</u>

The Fulani Jihad occurred between 1804-1808 in Nigeria and continued throughout much of the C19th, eventually reaching present-day Cameroon. It was initially a conflict between armies led by the Fulani scholar Usman Dan Fodiyo and the Hausa kingdoms of northern Nigeria, and led to the creation of the multi-ethnic Sokoto Caliphate in 1804. Sokoto dominated much of west-central Africa in the 19th century, stretching from the Niger River in the west to the Adamawa Plateau in the southeast and engaging in warfare and slave-raiding with neighboring populations. The Sokoto Caliphate ended when the British and Germans conquered its territory between 1891 and 1903 (*73*).

Supplementary Text 3

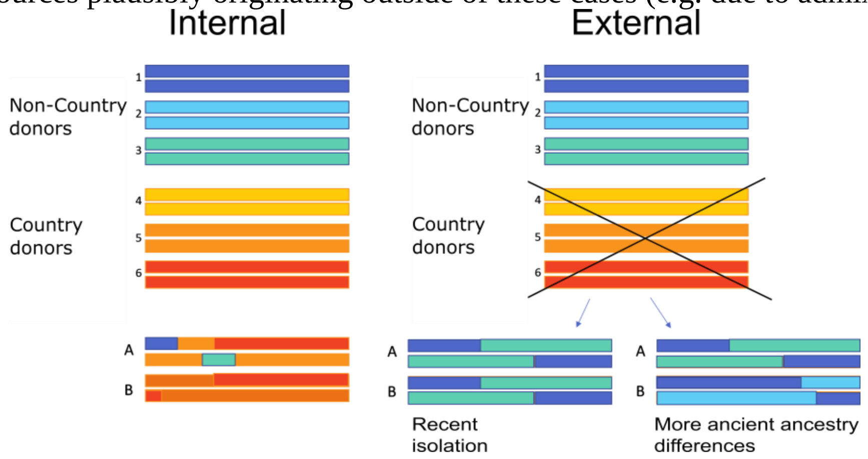
ChromoPainter analyses

ChromoPainter uses haplotype sharing patterns to represent (or ‘paint’) each of a focal individual’s two haploid genomes as a mosaic of ‘chunks’ of DNA copied from a set of other so-called ‘donor’ haploid genomes in the population. A chunk is a set of contiguous SNPs copied from the same donor haploid, with the donor of each chunk changing as one moves along the focal individual’s chromosome to reflect historical recombination events (*93*). At each genomic location, the donor copied is inferred to be the one that shares a most recent ancestor (i.e., more recent than any other donor haploid) with the focal individual. Chromosome painting can detect fine-scale genetic structure missed by other commonly used methods based on independent SNPs and is less susceptible to biases as a result of SNP array ascertainment (*8, 29, 47*). ChromoPainter requires two scaling parameters, the switch parameter, Ne (‘-n’), which controls the rate of switching from donor to donor within a focal haploid, and the mutation parameter, θ (‘-M’), which controls how often a focal haploid is allowed to copy from a donor that carries a different allele from that of the focal haploid.

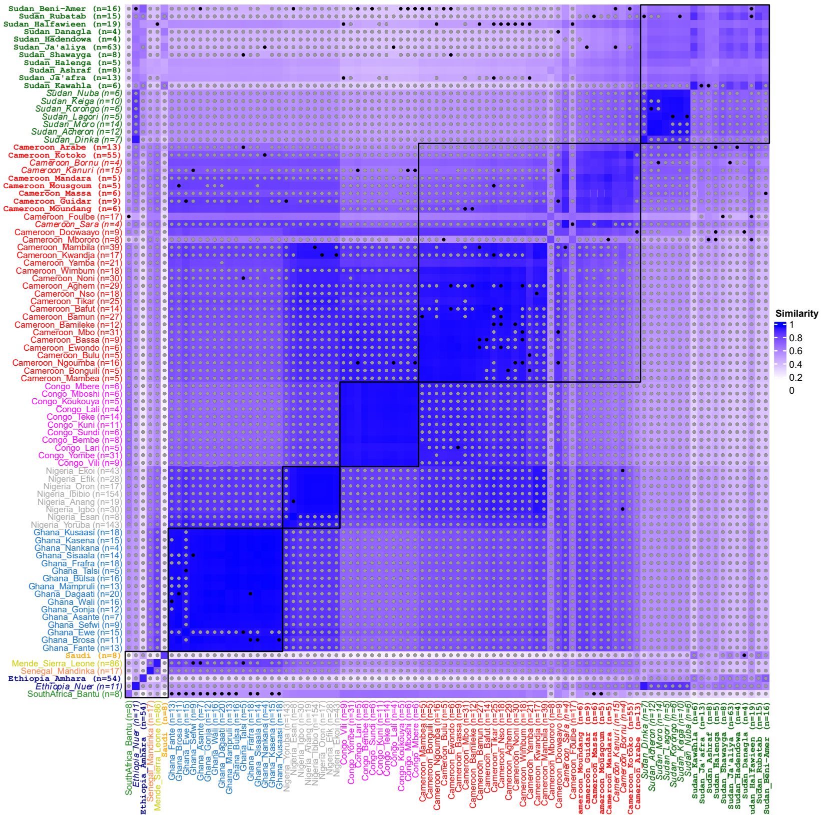
Analogous to van Dorp et al. (*9*), for this study we performed two separate chromosome paintings, each comparing individuals to a different set of donors:

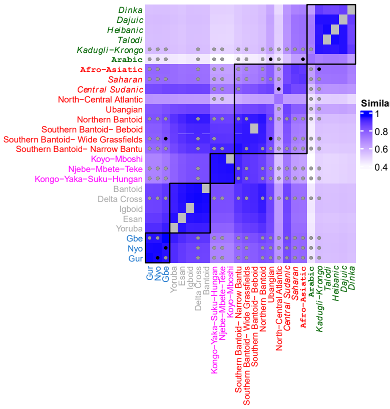
1. **Internal:** all individuals in the dataset are painted against (i.e., compared to) every other individual (Ne= 172.735, Mu= 0.000619838).
2. **External:** individuals are painted without allowing donors from Cameroon, Republic of the Congo, Ghana, Nigeria and Sudan (3673 donor individuals, 281 populations represented) (Ne= 213.426, Mu= 0.000676708).

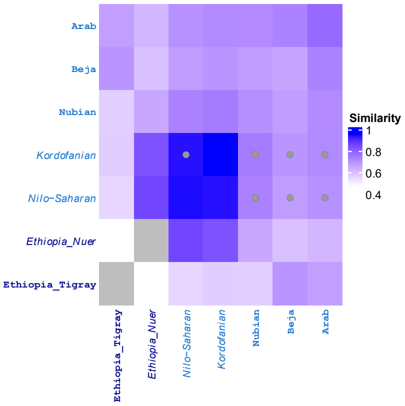
Under the internal painting, isolated groups are likely to copy long segments from other individuals within the same ethnic group or geographic region, which may make their painting appear very different to individuals from other nearby ethnic groups. The external painting acts to mitigate this effect by preventing individuals copying from their own or nearby groups (*9*). Consistent with this, when calculating a measure of genetic similarity based on the chromosome paintings (1-TVD, Methods), the mean value for individuals from Cameroon, Republic of the Congo, Ghana, Nigeria, and Sudan is 0.57 for the internal painting and 0.77 for the external painting, indicating higher genetic similarity in the latter. Thus, the external painting can reveal older relationships among populations not affected by recent isolation. Furthermore, by excluding neighbouring populations as donors, the external painting can also highlight ancestry derived from sources plausibly originating outside of these cases (e.g. due to admixture).



In the ‘internal’ analysis (Figure above left), ChromoPainter uses individuals 1-6 as donors with which to paint individuals A and B. Each haploid (row) of A and B is constructed as a mosaic of the donors’ haploid genomes. In the ‘external’ analysis (above right), the recipients can only be painted with donors from outside of the recipient’s country or region. Individual A and B will on average look more similar to each other in the external analysis as the effects of recent isolation are mitigated. Alternatively, if A and B have different sources of ancestry related to the non-Country donors (e.g. due to separate admixture events), they may still look different under the external analysis.







Above are heatmaps analogous to the ones shown in Fig. S8-9 except using the ‘external’ painting described above. The mean similarity between ethnic groups is higher in these heatmaps, and patterns of population structure differ. For example, in Fig S8 many of the ethnic groups in Ghana are genetically differentiable, but the majority of this structure disappears in the above heatmap, suggesting it is driven by endogamy/isolation rather than differences in ancestry. This is also seen in southern Cameroon, where genetic similarity between ethnic groups increases in the external analysis, indicating similar patterns of ancestry in this region. However, in some cases genetic structure is still detected when using the external painting, for example the structure between Arabic and Kordofanian-speaking Sudanese and between ethnic groups from different countries. In these cases, it suggests the genetic differences are driven by more external differences in ancestry rather than simple isolation/endogamy.

Another interesting result from the external analyses is in the heatmap of genetic differences between Sudanese language families. When comparing Kordofanian and Nilo-Saharan speakers, they have a high genetic similarity but are still differentiable, with the Nilo-Saharan speakers having a greater similarity to the Nilo-Saharan-speaking Ethiopian Nuer. This suggests that these two Sudanese groups, in addition to some degree of recent isolation, may have some older differences related to speaking languages from different phyla.

The external painting was used as the input for the SOURCEFIND results in Fig. 5 and Fig. S15 and the fastGLOBETROTTER analyses (Methods). This is because it allows detection of admixture events between sources more external to the focal group, which are likely to be older, rather than focusing on recent mixture between more proximate sources (*9*).

Supplementary Text 4

Code for msprime simulations in Fig. S21

(A)

```
mu=1.25e-8 # mutation rate per bp
rho=1e-8 # recombination rate per bp
nbp = 1e8
demography = msprime.Demography()
demography.add_population(name="A", initial_size=1000000, growth_rate=0.2763102)
demography.add_population_parameters_change(time=25, population="A", growth_rate=0,
initial_size=1000)
demography.add_population_parameters_change(time=45, population="A", initial_size=1000,
growth_rate=-0.9868222)
demography.add_population_parameters_change(time=52, population="A",
initial_size=1000000, growth_rate=0)
demography.add_population_parameters_change(time=68, population="A",
initial_size=1000000, growth_rate=0.163001)
demography.add_population_parameters_change(time=92, population="A", initial_size=5000,
growth_rate=0)
demography.sort_events()
ts = msprime.sim_ancestry({"A": 100}, demography=demography, recombination_rate=1e-8,
sequence_length=nbp, random_seed=seed1)
ts_mutated=msprime.sim_mutations(ts, rate=mu, random_seed=seed2)
```

(B)

```
mu=1.25e-8 # mutation rate per bp
rho=1e-8 # recombination rate per bp
nbp = 1e8
demography = msprime.Demography()
demography.add_population(name="A", initial_size=1000000, growth_rate=0.1564809)
demography.add_population_parameters_change(time=25, population="A", growth_rate=0,
initial_size=20000)
demography.add_population_parameters_change(time=45, population="A", initial_size=20000,
growth_rate=-0.5588604)
demography.add_population_parameters_change(time=52, population="A",
initial_size=1000000, growth_rate=0)
demography.add_population_parameters_change(time=68, population="A",
initial_size=1000000, growth_rate=0.163001)
demography.add_population_parameters_change(time=92, population="A", initial_size=5000,
growth_rate=0)
demography.sort_events()
ts = msprime.sim_ancestry({"A": 100}, demography=demography, recombination_rate=1e-8,
sequence_length=nbp, random_seed=seed1)
ts_mutated=msprime.sim_mutations(ts, rate=mu, random_seed=seed2)
```

(C)

```
mu=1.25e-8 # mutation rate per bp
rho=1e-8 # recombination rate per bp
nbp = 1e8
demography = msprime.Demography()
demography.add_population(name="A", initial_size=100000000, growth_rate=0.3406877)
demography.add_population_parameters_change(time=25, population="A", growth_rate=0,
initial_size=20000)
demography.add_population_parameters_change(time=45, population="A", initial_size=20000,
growth_rate=-0.5588604)
demography.add_population_parameters_change(time=52, population="A",
initial_size=1000000, growth_rate=0)
demography.add_population_parameters_change(time=68, population="A",
initial_size=1000000, growth_rate=0.163001)
demography.add_population_parameters_change(time=92, population="A", initial_size=5000,
growth_rate=0)
demography.sort_events()
ts = msprime.sim_ancestry({"A": 100}, demography=demography, recombination_rate=1e-8,
sequence_length=nbp, random_seed=seed1)
ts_mutated=msprime.sim_mutations(ts, rate=mu, random_seed=seed2)
```

(D)

```
mu=1.25e-8 # mutation rate per bp
rho=1e-8 # recombination rate per bp
nbp = 1e8
demography = msprime.Demography()
demography.add_population(name="A", initial_size=1000000, growth_rate=0.1564809)
demography.add_population_parameters_change(time=25, population="A", growth_rate=0,
initial_size=20000)
demography.add_population_parameters_change(time=45, population="A", initial_size=20000,
growth_rate=-0.5588604)
demography.add_population_parameters_change(time=52, population="A",
initial_size=1000000, growth_rate=0)
demography.add_population_parameters_change(time=68, population="A",
initial_size=1000000, growth_rate=0.1248222)
demography.add_population_parameters_change(time=92, population="A", initial_size=50000,
growth_rate=0)
demography.sort_events()
ts = msprime.sim_ancestry({"A": 100}, demography=demography, recombination_rate=1e-8,
sequence_length=nbp, random_seed=seed1)
ts_mutated=msprime.sim_mutations(ts, rate=mu, random_seed=seed2)
```

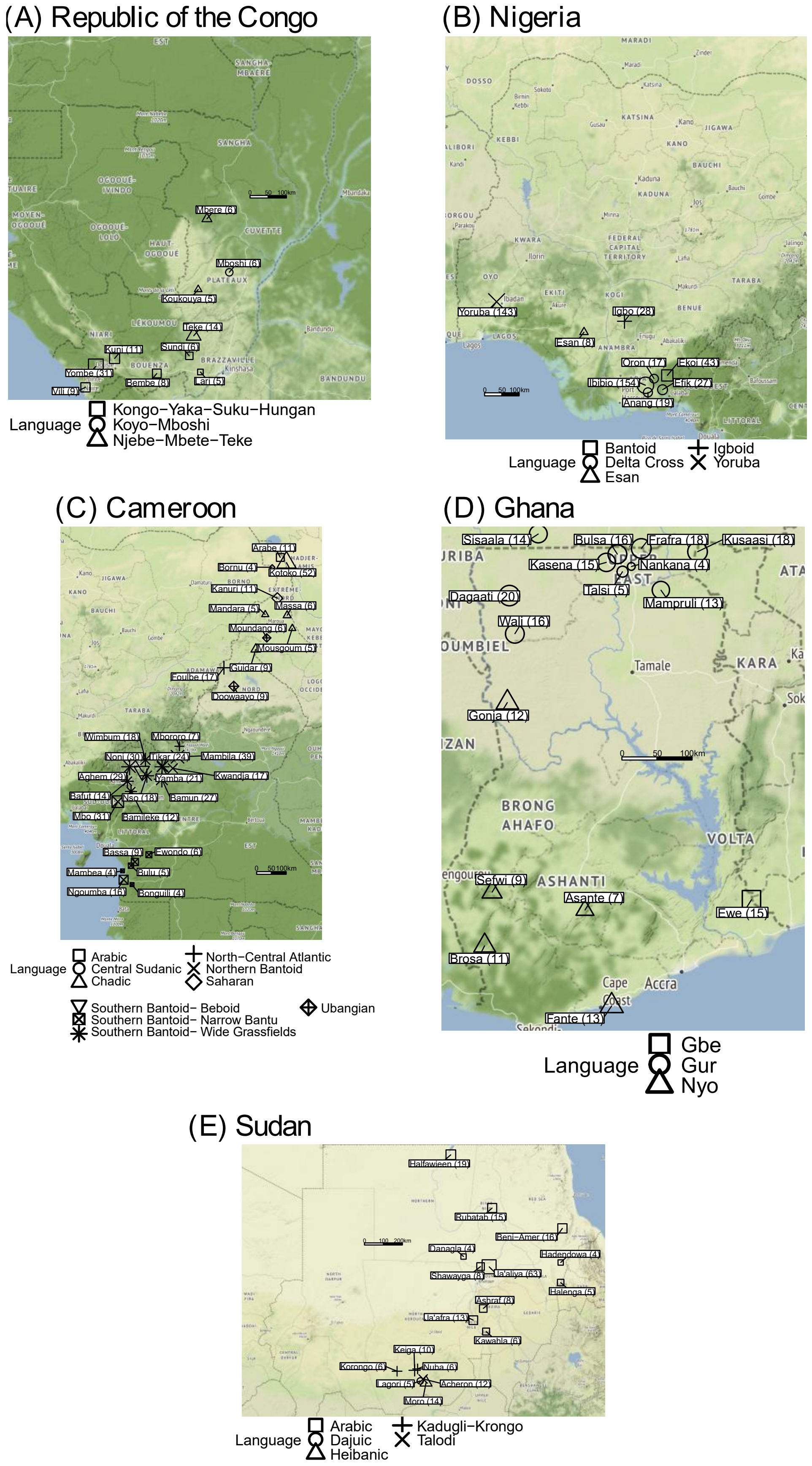

(A) Republic of the Congo
Mbere (6)
Mboshi (6)
Koukouya (5)
Teke (14)
Sundi (6)
Kuni (11)
Yombe (31)
Bembe (8)
Lari (5)
Vili (9)
0 50 100km
Language
Kongo−Yaka−Suku−Hungan
Koyo−Mboshi
Njebe−Mbete−Teke
(B) Nigeria
Yoruba (143)
Igbo (28)
Esan (8)
Oron (17)
Ekoi (43)
Ibibio (154)
Efik (27)
Anang (19)
0 50 100km
Language
Bantoid
Delta Cross
Esan
Igboid
Yoruba
(C) Cameroon
Arabe (11)
Bornu (4)
Kotoko (52)
Kanuri (11)
Mandara (5)
Massa (6)
Moundang (6)
Mousgoum (5)
Guidar (9)
Foulbe (17)
Doowaayo (9)
Wimbum (18)
Mbororo (7)
Noni (30)
Tikar (24)
Mambila (39)
Aghem (29)
Yamba (21)
Kwandja (17)
Bafut (14)
Nso (18)
Bamun (27)
Mbo (31)
Bamileke (12)
Bassa (9)
Ewondo (6)
Mambea (4)
Bulu (5)
Ngoumba (16)
Bonguili (4)
0 50 100km
Language
Arabic
Central Sudanic
Chadic
North−Central Atlantic
Northern Bantoid
Saharan
Southern Bantoid− Beboid
Southern Bantoid− Narrow Bantu
Southern Bantoid− Wide Grassfields
Ubangian
(D) Ghana
Sisaala (14)
Bulsa (16)
Frafra (18)
Kusaasi (18)
Kasena (15)
Nankana (4)
Talsi (5)
Dagaati (20)
Mampruli (13)
Wali (16)
Gonja (12)
0 50 100km
Sefwi (9)
Asante (7)
Ewe (15)
Brosa (11)
Fante (13)
Language
Gbe
Gur
Nyo
(E) Sudan
Halfawieen (19)
Rubatab (15)
Beni−Amer (16)
0 100 200km
Danagla (4)
Hadendowa (4)
Shawayga (8)
Ja'aliya (63)
Halenga (5)
Ashraf (8)
Ja'afra (13)
Kawahla (6)
Keiga (10)
Korongo (6)
Nuba (6)
Lagori (5)
Acheron (12)
Moro (14)
Language
Arabic
Dajuic
Heibanic
Kadugli−Krongo
Talodi

Fig. S1.

Map of samples from (A) Republic of the Congo, (B) Nigeria, (C) Cameroon, (D) Ghana and (E) Sudan. Each point indicates an ethnic group, placed according to the mean birthplaces of its individuals' maternal grandmothers and paternal grandfathers. Only ethnic groups that contain more than three individuals are shown. Point shape shows language family.

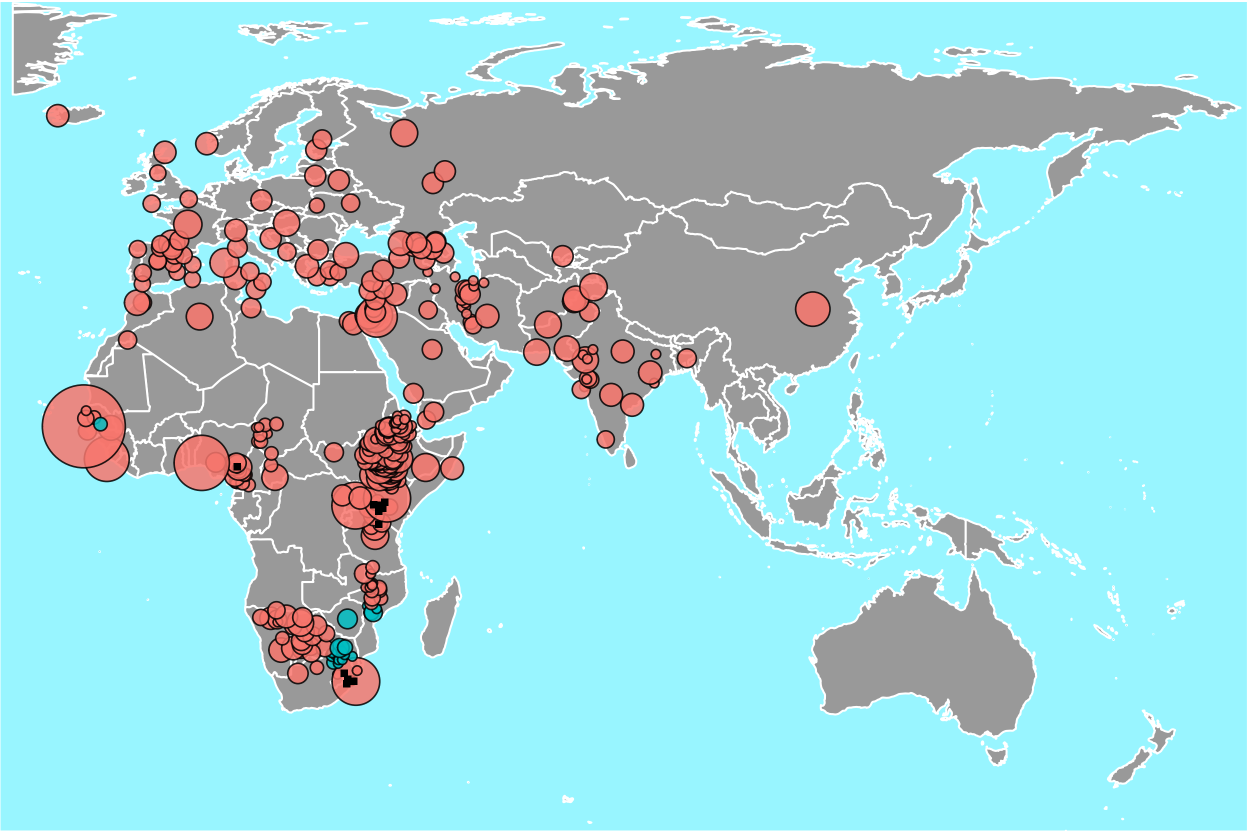


Fig. S2.

Map of previously published populations merged with the dataset (pink, n=3866, Data S2) and newly reported data genotyped for this study sampled from Mozambique, South Africa and Zimbabwe (blue, n=54). This excludes newly reported data from Cameroon, Republic of the Congo, Ghana, Nigeria and Sudan (n=1333, Fig. 1, Data S1). Two of the individuals sampled in South Africa had a birthplace in Senegal. Ancient individuals are shown with a black square. Size of point corresponds with population size (range is 1-100, mean=12). Modern-day samples are from Byrska-Bishop et al. (*43*), Fan et al. (*6*), Gurdasani et al. (*4*), Lazaridis et al. (*41*), Lipson et al. (*12*), López et al. (*45*), López et al. (*8*), MalariaGEN (*44*), Skoglund et al. (*11*) and Zheng-Bradley et al. (*42*). The 20 ancient African individuals from Cameroon, Kenya, Tanzania and South Africa were published in Lipson et al. (*12*), Prendergast et al. (*13*) and Schlebusch et al. (*10*).

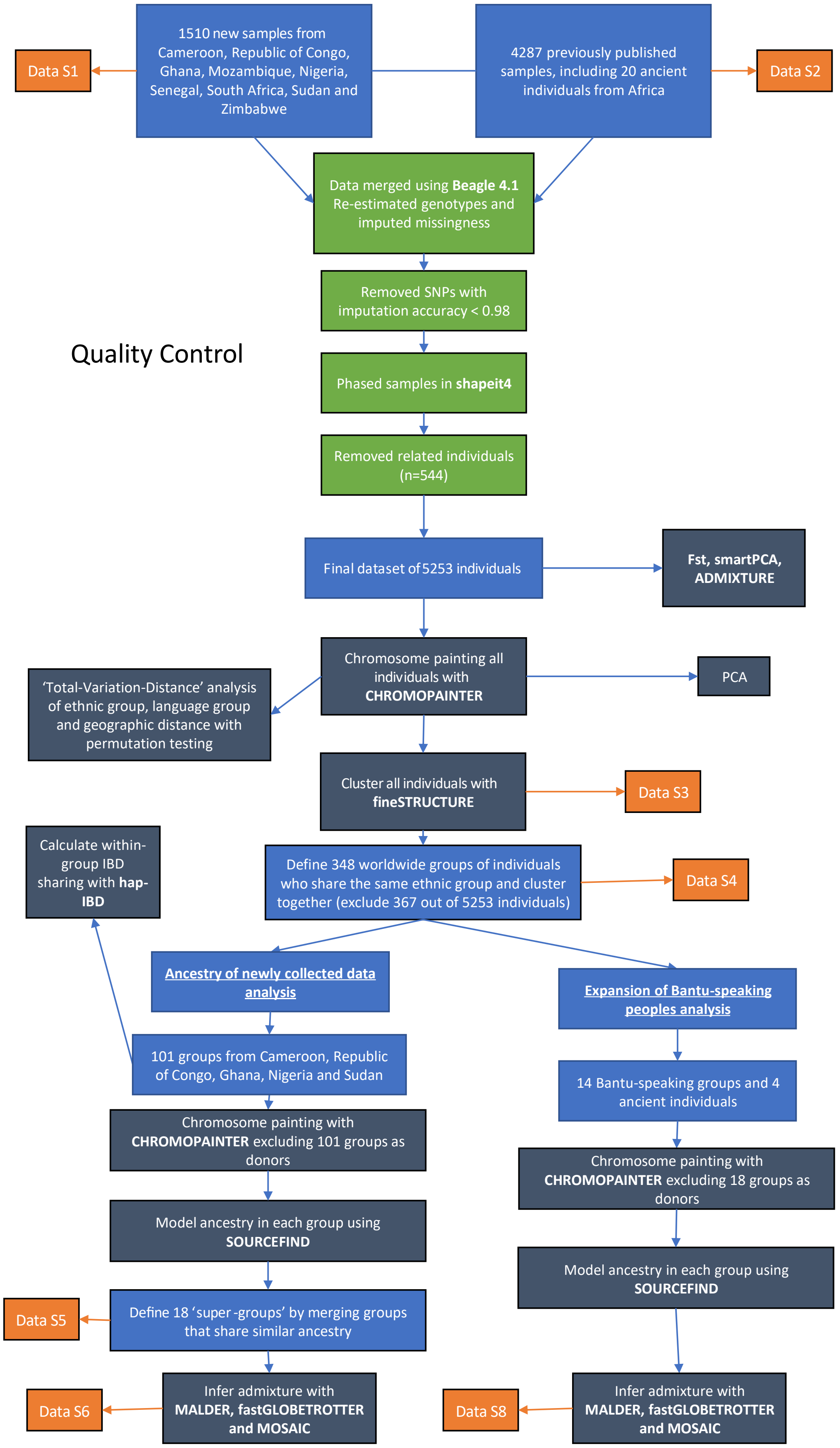

1510 new samples from Cameroon, Republic of Congo, Ghana, Mozambique, Nigeria, Senegal, South Africa, Sudan and Zimbabwe
Data S1
4287 previously published samples, including 20 ancient individuals from Africa
Data S2
Data merged using Beagle 4.1 Re-estimated genotypes and imputed missingness
Removed SNPs with imputation accuracy < 0.98
Quality Control
Phased samples in shapeit4
Removed related individuals (n=544)
Final dataset of 5253 individuals
Fst, smartPCA, ADMIXTURE
Chromosome painting all individuals with CHROMOPAINTER
PCA
'Total-Variation-Distance' analysis of ethnic group, language group and geographic distance with permutation testing
Cluster all individuals with fineSTRUCTURE
Data S3
Calculate within-group IBD sharing with hap-IBD
Define 348 worldwide groups of individuals who share the same ethnic group and cluster together (exclude 367 out of 5253 individuals)
Data S4
Ancestry of newly collected data analysis
Expansion of Bantu-speaking peoples analysis
101 groups from Cameroon, Republic of Congo, Ghana, Nigeria and Sudan
14 Bantu-speaking groups and 4 ancient individuals
Chromosome painting with CHROMOPAINTER excluding 101 groups as donors
Chromosome painting with CHROMOPAINTER excluding 18 groups as donors
Model ancestry in each group using SOURCEFIND
Model ancestry in each group using SOURCEFIND
Data S5
Define 18 'super-groups' by merging groups that share similar ancestry
Data S6
Infer admixture with MALDER, fastGLOBETROTTER and MOSAIC
Data S8
Infer admixture with MALDER, fastGLOBETROTTER and MOSAIC

Fig. S3.

Flowchart of the analysis process. Blue boxes indicate dataset/analysis descriptions, green boxes show quality control steps, grey boxes show analysis steps, and orange boxes indicate where data is reported.

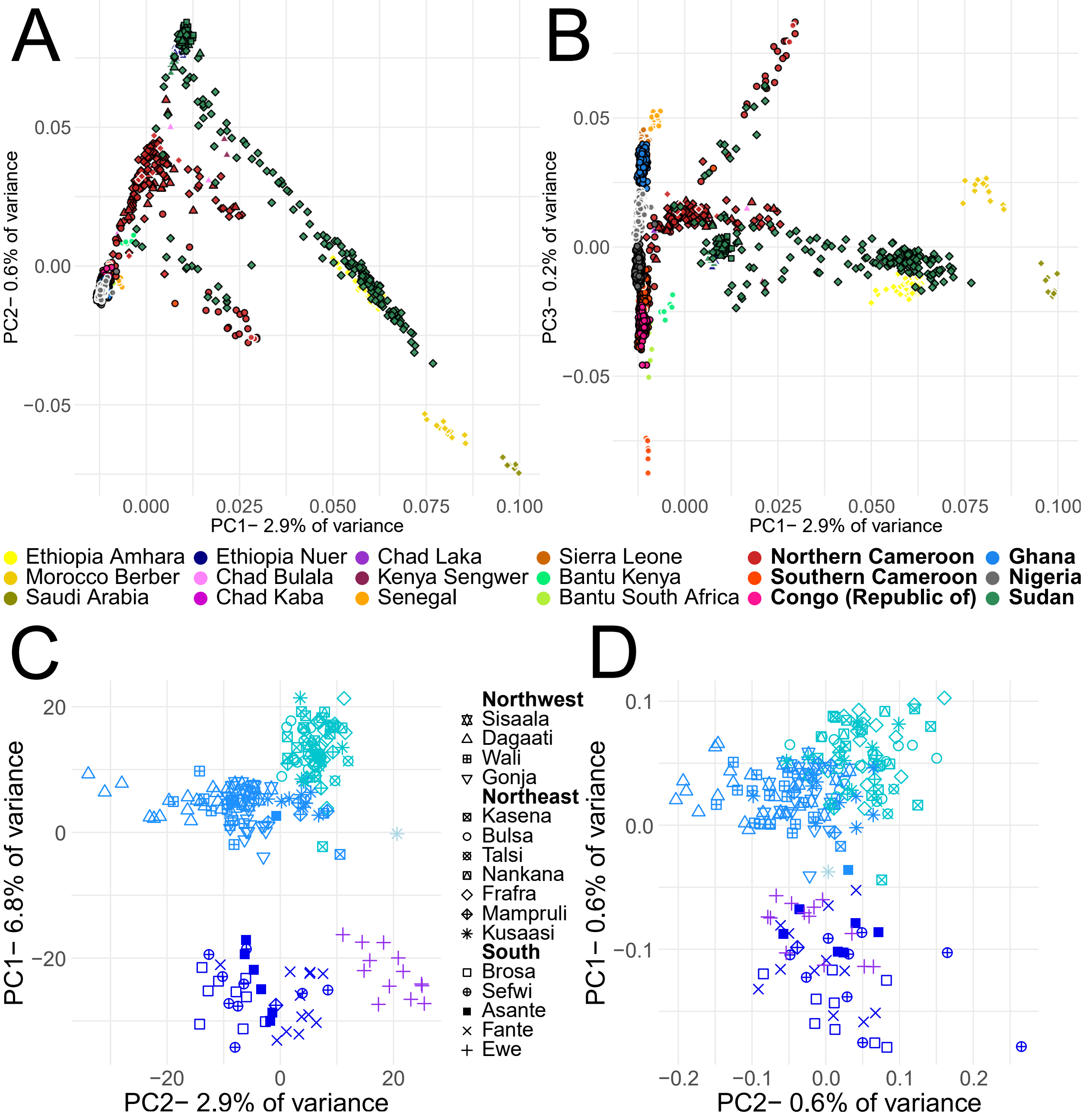


Fig. S4.

Comparing principal component analyses calculated either using smartPCA or from the ChromoPainter output (Methods). (A) and (B) principal component analysis (PCA) using the same samples as Fig. 2, but calculated using each SNP independently with smartPCA, rather than with haplotypes. (C) and (D) PCAs of all samples from Ghana, where shape corresponds to ethnic group and colour corresponds to fineSTRUCTURE cluster (see Fig. 3-4). In (C) PCs were generated using haplotypes, while (D) uses smartPCA. By using haplotypes, as in (C), there is increased power to differentiate between eastern and western clusters in Ghana.

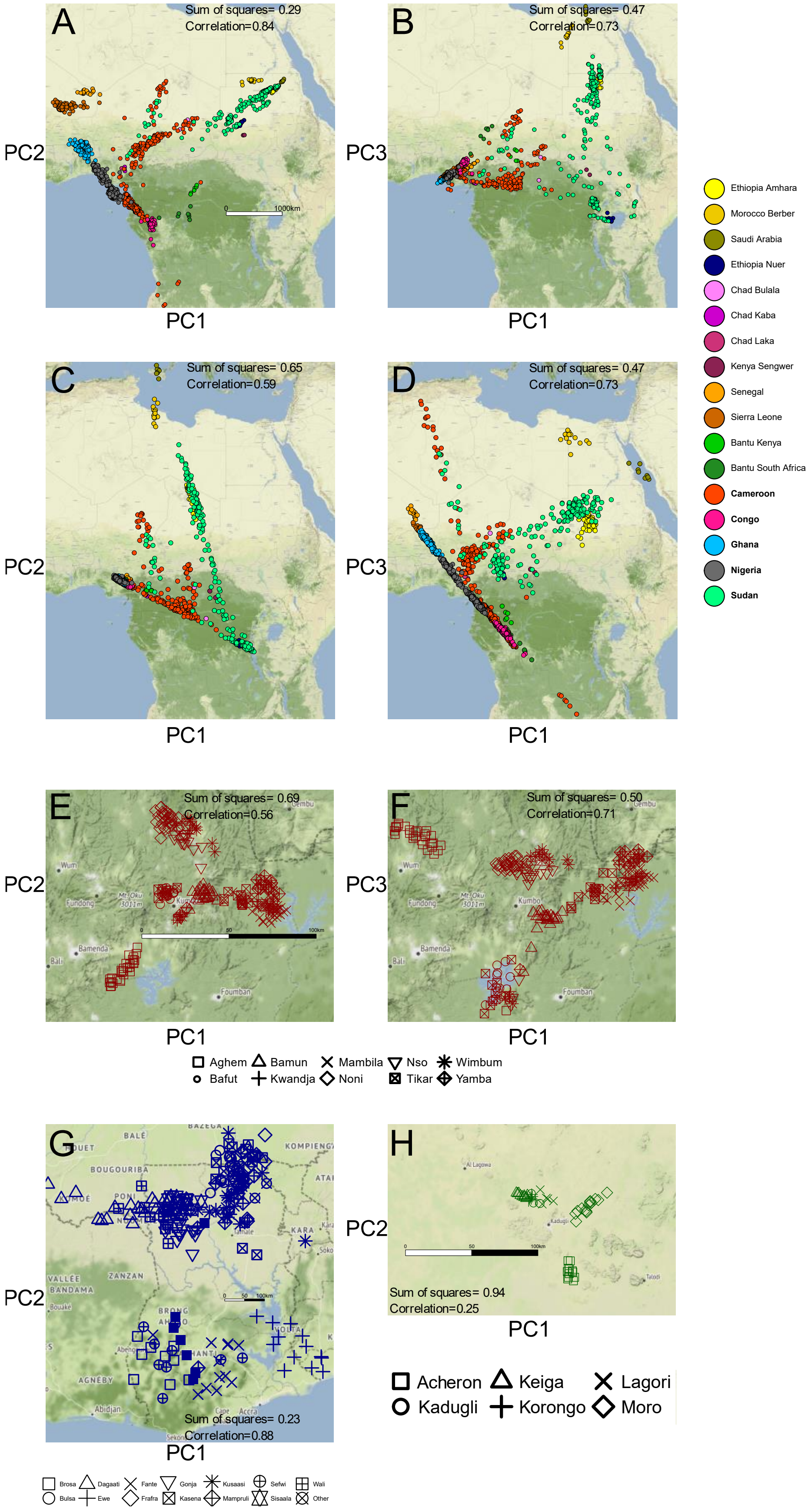

A
Sum of squares= 0.29
Correlation=0.84
PC2
PC1
0
1000km
B
Sum of squares= 0.47
Correlation=0.73
PC3
PC1
Ethiopia Amhara
Morocco Berber
Saudi Arabia
Ethiopia Nuer
Chad Bulala
Chad Kaba
Chad Laka
Kenya Sengwer
Senegal
Sierra Leone
Bantu Kenya
Bantu South Africa
Cameroon
Congo
Ghana
Nigeria
Sudan
C
Sum of squares= 0.65
Correlation=0.59
PC2
PC1
D
Sum of squares= 0.47
Correlation=0.73
PC3
PC1
E
Sum of squares= 0.69
Correlation=0.56
PC2
PC1
0
50
100km
F
Sum of squares= 0.50
Correlation=0.71
PC3
PC1
Aghem
Bamun
Mambila
Nso
Wimbum
Bafut
Kwandja
Noni
Tikar
Yamba
G
PC2
Sum of squares= 0.23
Correlation=0.88
PC1
0
50
100km
Brosa
Dagaati
Fante
Gonja
Kusaasi
Sefwi
Wali
Bulsa
Ewe
Frafra
Kasena
Mamprul
Sisaala
Other
H
PC2
0
50
100km
Sum of squares= 0.94
Correlation=0.25
PC1
Acheron
Keiga
Lagori
Kadugli
Korongo
Moro

Fig. S5.

Procrustes analyses inferring the correlation between principal components and geography. (A) and (B) Procrustes analyses of the haplotype-based PCA in Fig. 2, (C) and (D) Procrustes analyses of the smartPCA PCA in Fig. S4, all showing the rotated PCs over a map, and the sum of squares and the Pearson correlation between PCs and the mean coordinates of each individual. For PC2 and PC3, both the haplotype-based PCA and smartPCA do equally well. However, for PC1 and PC2, the haplotype-based PCA has a much stronger correlation with geography.

(E), (F), (G) and (H) show Procrustes analyses for the regions shown in Fig. 4, using a haplotype-based PCA. For the Grassfields, correlation with PC3 is better than PC2. Ghana shows a very high correlation between PCs and geography, while the South Kordofan region of Sudan has a low correlation.

See separate pdf.

Fig. S6.
ADMIXTURE analysis from K=2 to K=9 of the newly reported samples from Cameroon, Republic of the Congo, Ghana, Nigeria and Sudan. Reference populations on the left include: Norwegian, Orcadian, Palestinian, Morocco Berber, Ethiopia Amhara, Ethiopia Nuer, Senegal Mandinka, Kenya Bantu, South Africa Bantu and Mbuti Pygmy. The lowest cross validation error was at K=7.

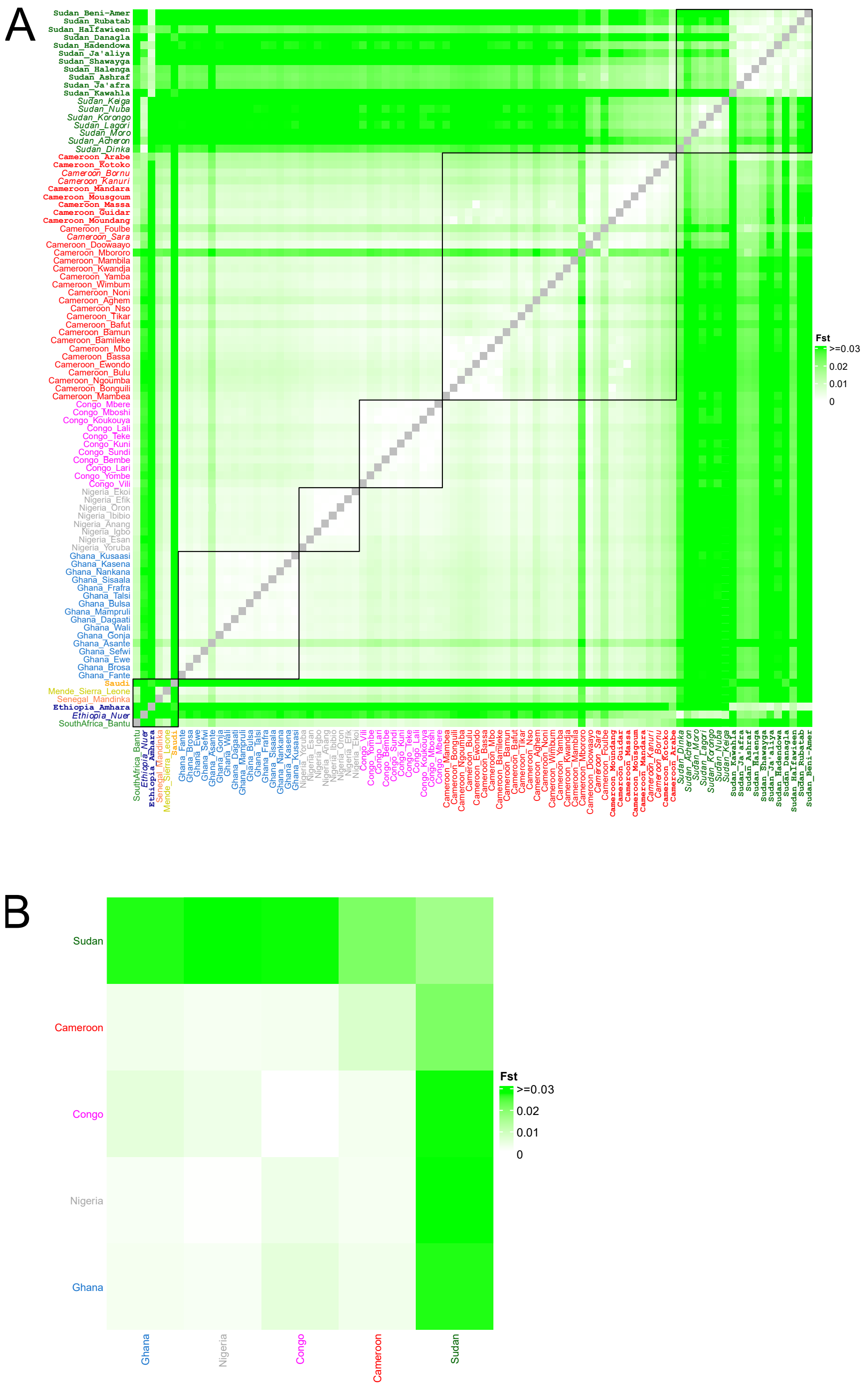


Fig. S7.

Heatmaps showing the average genetic similarity (F_{st}) between (A) every pair of ethnic groups sampled from Cameroon, Republic of the Congo, Ghana, Nigeria and Sudan, as well as populations included in the PCA in Fig. 2 and (B) each country. Ethnic groups in bold speak Afro-Asiatic languages, and those in italics speak Nilo-Saharan or Kordofanian languages. All other groups speak a Niger-Congo language.

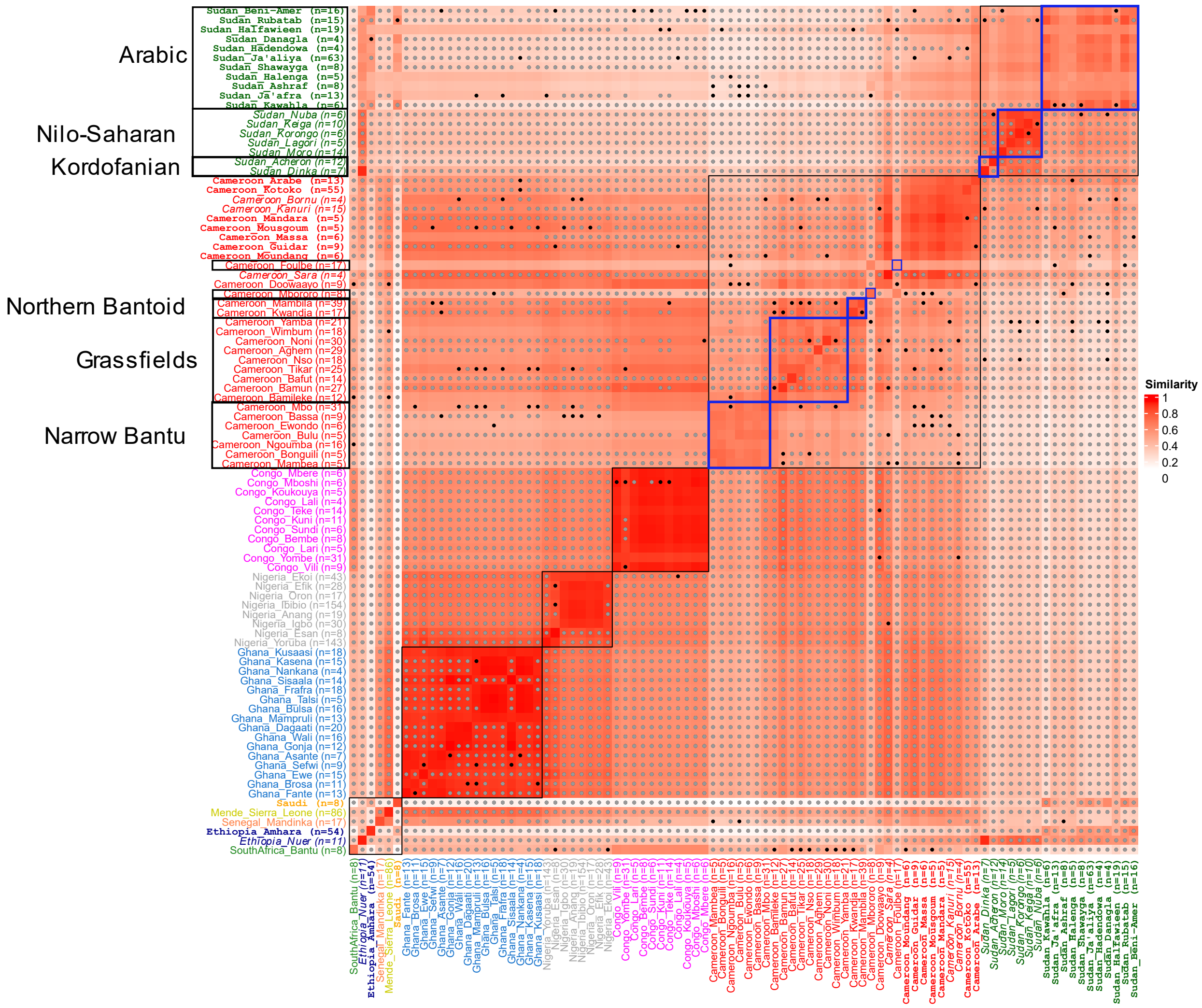


Fig. S8.

Heatmap showing the average genetic similarity (calculated as 1- TVD, Methods) between pairs of individuals from every pairing of ethnic groups sampled from Cameroon, Republic of the Congo, Ghana, Nigeria and Sudan, as well as populations included in the PCA in Fig. 2. Ethnic groups in bold speak Afro-Asiatic languages, and those in italics speak Nilo-Saharan or Kordofanian languages. All other groups speak a Niger-Congo language. Dots denote pairs of groups for which we can reject the null hypothesis that individuals from group A (rows) are more genetically similar to each other than individuals from A are to those from B (columns) at a per-test threshold of p<0.01 (black) or p<0.001 (grey) (Methods). These tests are not symmetric; group A can be distinguishable from group B while group B is not distinguishable from group A if (e.g.) group B has high genetic diversity. We conclude that the two groups can be distinguished if there is significance in either direction.

Annotations have been added to demonstrate certain results, with language groups shown at the sides and blue boxes placed around ethnic groups that speak that language. Firstly, there is low genetic similarity both within and between Arabic-speaking Sudanese groups, with only some ethnic groups being differentiable. This is in contrast with Nilo-Saharan and Kordofanian-speaking Sudanese who have a higher genetic similarity within language group and many differentiable ethnic groups.

Next, both Fulani ethnic groups (self-described as Foulbe and Mbororo) are distinguishable from almost all other ethnic groups, while there is very little structure in the rest of northern Cameroon.

In southern Cameroon, the three main language groups (northern Bantoid, Grassfields and Narrow Bantu) are highlighted. Many ethnic groups are differentiable within northern Bantoid and Grassfields speakers, while Narrow Bantu speaker have only one pair of ethnic groups that are differentiable and a relatively low genetic similarity within and between ethnic groups.

Lastly, Ghana show relatively high genetic similarity among groups and low genetic similarity with groups outside Ghana. In Nigeria, structure is mostly present between western (Yoruba and Esan) and eastern ethnic groups.

Note: Some individuals self-described as Tikar speak a Grassfields language and some speak a Northern Bantoid language.

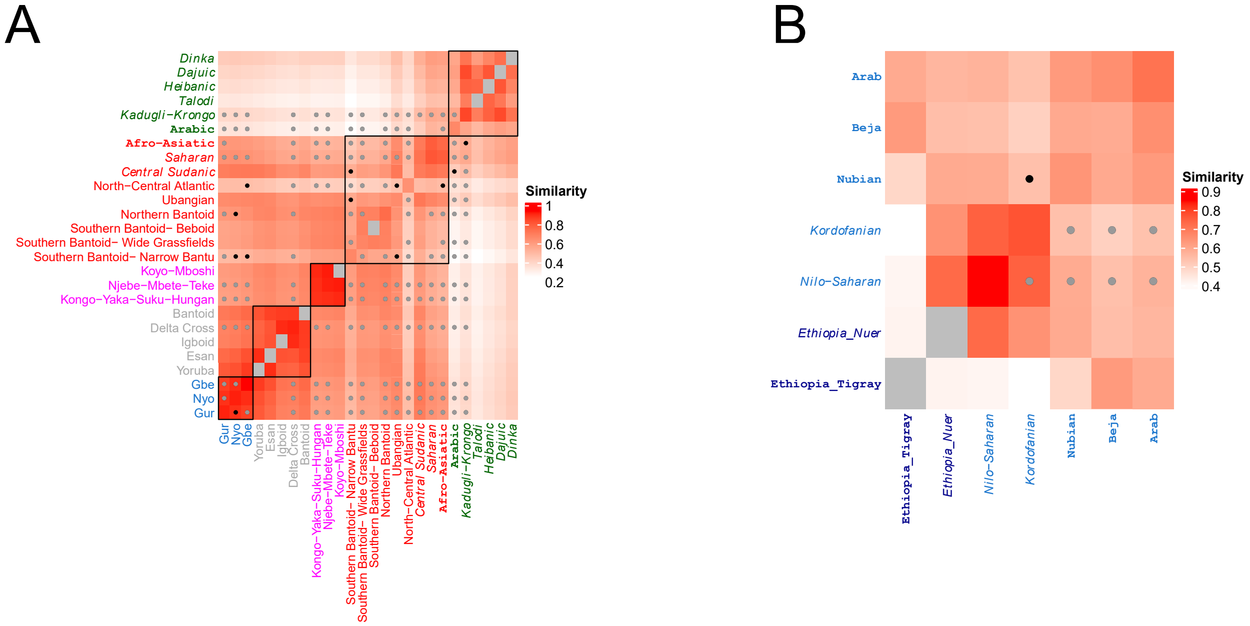


Fig. S9.

Heatmaps showing the average genetic similarity (calculated as 1- TVD) between pairs of individuals from different language classifications. The language classifications in (A) are given in Data S1, and described in more detail in Text S1. (B) shows language classifications within Sudan, grouping the languages of individuals from south Kordofan into Kordofanian and Nilo-Saharan. Sudanese Arabic speakers are classified into three ethnicities (Arab, Beja, Nubian) based on their pre-Arabic language (see Hollfelder et al. (*16*)). Ethiopian Nilo-Saharan and Afro-Asiatic speaking ethnic groups are included for reference. Languages in bold are Afro-Asiatic, and those in italics are Nilo-Saharan or Kordofanian. All other groups speak a Niger-Congo language. Dots denote pairs of groups for which we can reject the null hypothesis that individuals from group A (rows) are more genetically similar to each other than individuals from A are to those from B (columns) at a per-test threshold of p<0.01 (black) or p<0.001 (grey) (Methods). These tests are not symmetric; group A can be distinguishable from group B while group B is not distinguishable from group A if (e.g.) group B has high genetic diversity. We conclude that the two groups can be distinguished if there is significance in either direction. In the case where a language group comprised a single ethnic group, TVD could not be calculated as only comparisons between different ethnic groups were included (grey squares, see Methods). In these cases, permutation tests could not be carried out.

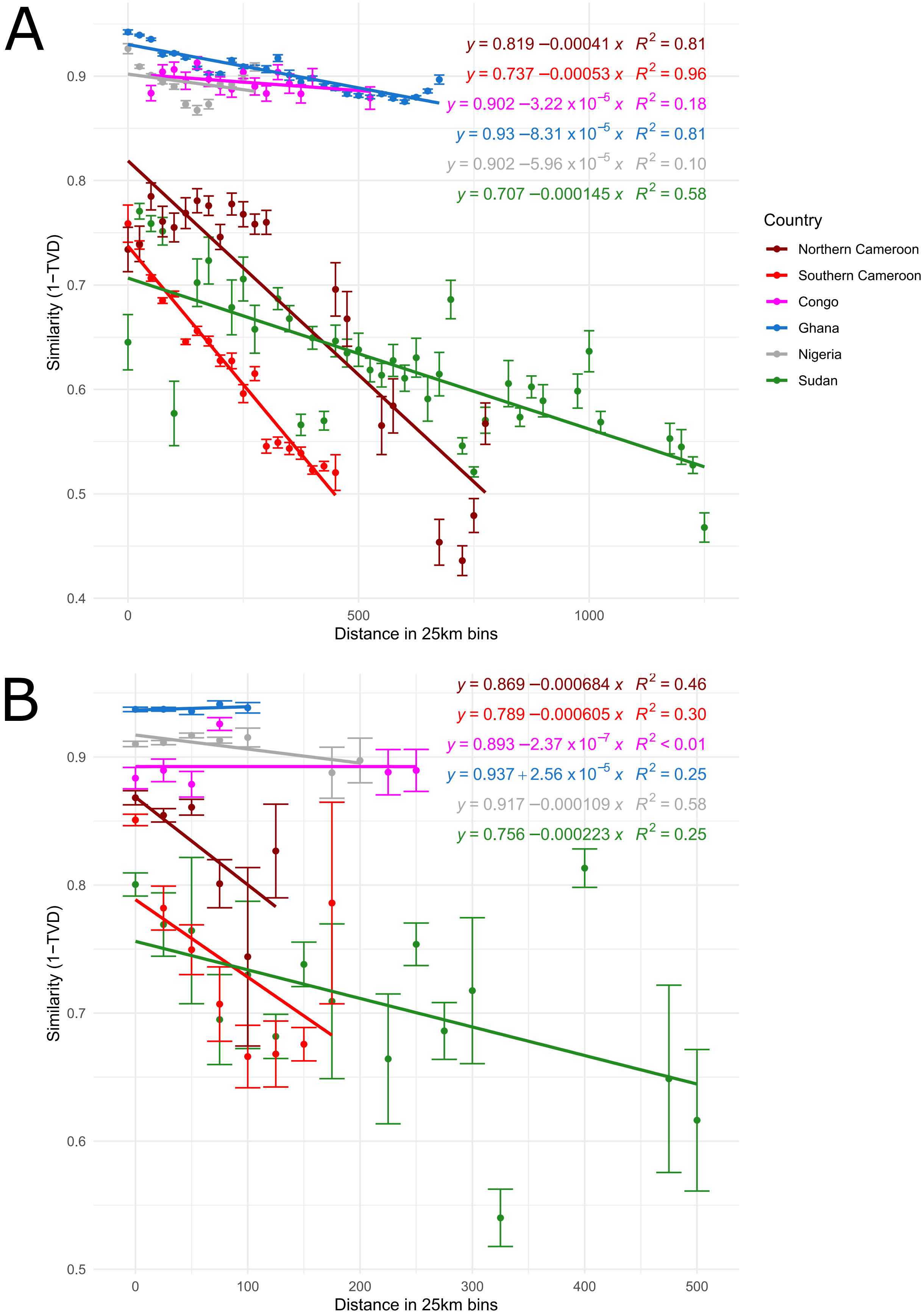

A
$y = 0.819 - 0.00041\,x$ $R^2 = 0.81$
$y = 0.737 - 0.00053\,x$ $R^2 = 0.96$
$y = 0.902 - 3.22 \times 10^{-5}\,x$ $R^2 = 0.18$
$y = 0.93 - 8.31 \times 10^{-5}\,x$ $R^2 = 0.81$
$y = 0.902 - 5.96 \times 10^{-5}\,x$ $R^2 = 0.10$
$y = 0.707 - 0.000145\,x$ $R^2 = 0.58$
Similarity (1−TVD)
0.9
0.8
0.7
0.6
0.5
0.4
0
500
1000
Distance in 25km bins
Country
Northern Cameroon
Southern Cameroon
Congo
Ghana
Nigeria
Sudan
B
$y = 0.869 - 0.000684\,x$ $R^2 = 0.46$
$y = 0.789 - 0.000605\,x$ $R^2 = 0.30$
$y = 0.893 - 2.37 \times 10^{-7}\,x$ $R^2 < 0.01$
$y = 0.937 + 2.56 \times 10^{-5}\,x$ $R^2 = 0.25$
$y = 0.917 - 0.000109\,x$ $R^2 = 0.58$
$y = 0.756 - 0.000223\,x$ $R^2 = 0.25$
Similarity (1−TVD)
0.9
0.8
0.7
0.6
0.5
0
100
200
300
400
500
Distance in 25km bins

Fig. S10.

Plots of genetic similarity (1-TVD) among individuals pairs against distance (km) in 25km bins, split by country. Cameroonian individuals are split into north and south based on fineSTRUCTURE results and geography (see Fig. 1 and Fig. 3). (A) considers only individual pairings from different ethnic groups, while (B) considers only pairs of individuals from the same ethnic group. The analysis in (A) removes ethnic groups effects, as individuals of the same ethnic group are likely to have high similarity and live at short distances from one another. Hence this gives us a more conservative estimate of an isolation by distance effect. The analysis in (B) reveals that isolation by distance can still be seen in some cases even amongst individuals of the same ethnic groups living at different distances from one another. Individuals whose maternal grandmother and paternal grandfather lived more than 150km apart were removed for this analysis. For (A) only bins with more than 100 comparisons were included; for (B) this threshold was 20. Both plots have been filtered to remove bins where all comparisons involve one particular individual.

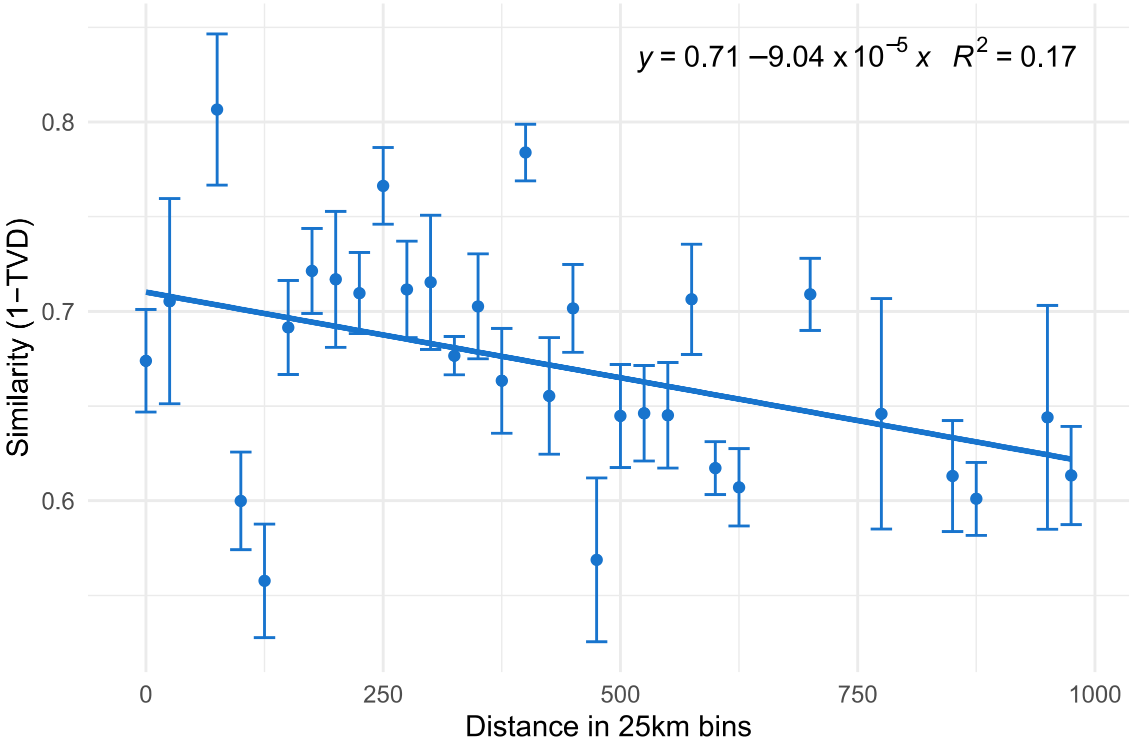


Fig. S11.
Plots of genetic similarity (1-TVD “internal” analysis) against distance (km) in 25km bins between each pair of individuals from different ethnic groups who live on the Nile in Sudan (i.e. excluding south Kordofan and coastal individuals). There is a weak negative correlation between geographic distance and genetic similarity. Individuals whose maternal grandmother and paternal grandfather lived more than 150km apart were removed for this analysis. Only bins with more than 20 comparisons were included and bins where all comparisons involve one particular individual were removed.

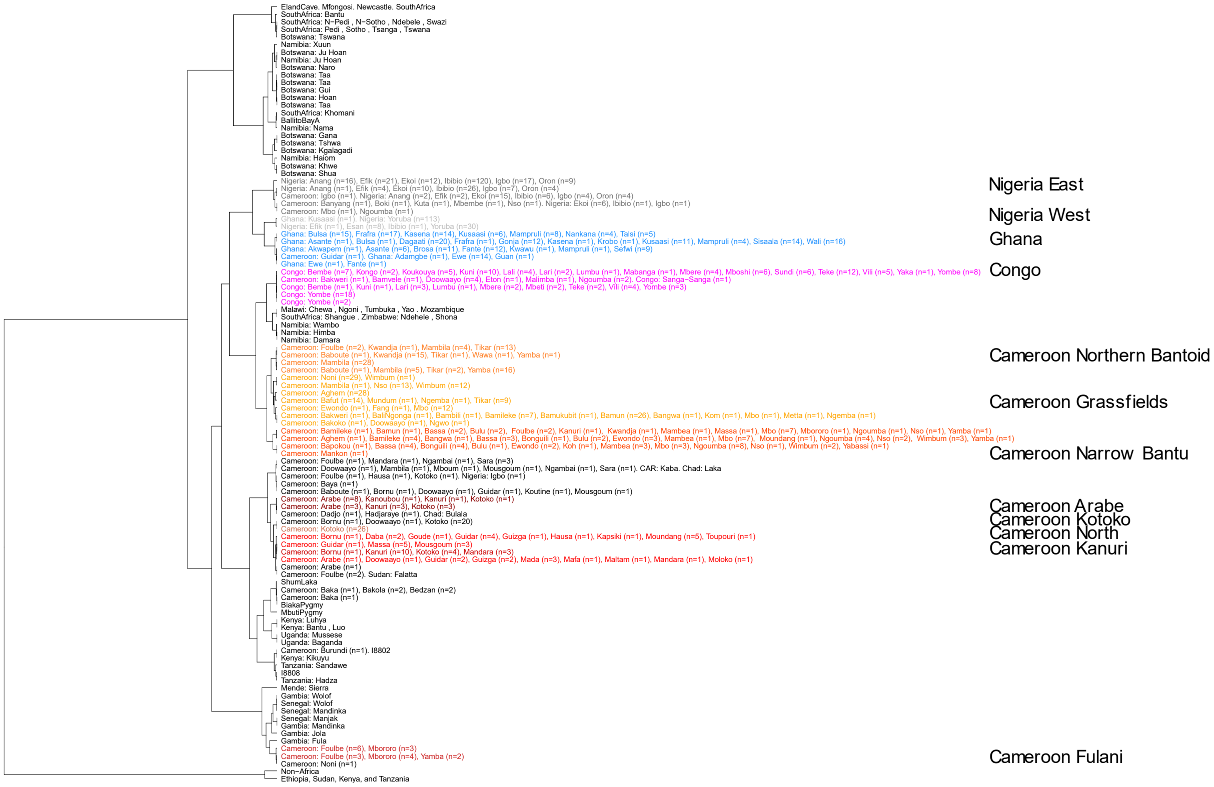

ElandCave. Mfongosi. Newcastle. SouthAfrica
SouthAfrica: Bantu
SouthAfrica: N−Pedi , N−Sotho , Ndebele , Swazi
SouthAfrica: Pedi , Sotho , Tsanga , Tswana
Botswana: Tswana
Namibia: Xuun
Botswana: Ju Hoan
Namibia: Ju Hoan
Botswana: Naro
Botswana: Taa
Botswana: Taa
Botswana: Gui
Botswana: Hoan
Botswana: Taa
SouthAfrica: Khomani
BallitoBayA
Namibia: Nama
Botswana: Gana
Botswana: Tshwa
Botswana: Kgalagadi
Namibia: Haiom
Botswana: Khwe
Botswana: Shua
Nigeria: Anang (n=16), Efik (n=21), Ekoi (n=12), Ibibio (n=120), Igbo (n=17), Oron (n=9)
Nigeria East
Nigeria: Anang (n=1), Efik (n=4), Ekoi (n=10), Ibibio (n=26), Igbo (n=7), Oron (n=4)
Cameroon: Igbo (n=1). Nigeria: Anang (n=2), Efik (n=2), Ekoi (n=15), Ibibio (n=6), Igbo (n=4), Oron (n=4)
Cameroon: Banyang (n=1), Boki (n=1), Kuta (n=1), Mbembe (n=1), Nso (n=1). Nigeria: Ekoi (n=6), Ibibio (n=1), Igbo (n=1)
Cameroon: Mbo (n=1), Ngoumba (n=1)
Nigeria West
Ghana: Kusaasi (n=1). Nigeria: Yoruba (n=113)
Nigeria: Efik (n=1), Esan (n=8), Ibibio (n=1), Yoruba (n=30)
Ghana: Bulsa (n=15), Frafra (n=17), Kasena (n=14), Kusaasi (n=6), Mampruli (n=8), Nankana (n=4), Talsi (n=5)
Ghana
Ghana: Asante (n=1), Bulsa (n=1), Dagaati (n=20), Frafra (n=1), Gonja (n=12), Kasena (n=1), Krobo (n=1), Kusaasi (n=11), Mampruli (n=4), Sisaala (n=14), Wali (n=16)
Ghana: Akwapem (n=1), Asante (n=6), Brosa (n=11), Fante (n=12), Kwawu (n=1), Mampruli (n=1), Sefwi (n=9)
Cameroon: Guidar (n=1). Ghana: Adamgbe (n=1), Ewe (n=14), Guan (n=1)
Ghana: Ewe (n=1), Fante (n=1)
Congo: Bembe (n=7), Kongo (n=2), Koukouya (n=5), Kuni (n=10), Lali (n=4), Lari (n=2), Lumbu (n=1), Mabanga (n=1), Mbere (n=4), Mboshi (n=6), Sundi (n=6), Teke (n=12), Vili (n=5), Yaka (n=1), Yombe (n=8)
Congo
Cameroon: Bakweri (n=1), Bamvele (n=1), Doowaayo (n=4), Eton (n=1), Malimba (n=1), Ngoumba (n=2). Congo: Sanga−Sanga (n=1)
Congo: Bembe (n=1), Kuni (n=1), Lari (n=3), Lumbu (n=1), Mbere (n=2), Mbeti (n=2), Teke (n=2), Vili (n=4), Yombe (n=3)
Congo: Yombe (n=18)
Congo: Yombe (n=2)
Malawi: Chewa , Ngoni , Tumbuka , Yao . Mozambique
SouthAfrica: Shangue . Zimbabwe: Ndehele , Shona
Namibia: Wambo
Namibia: Himba
Namibia: Damara
Cameroon: Foulbe (n=2), Kwandja (n=1), Mambila (n=4), Tikar (n=13)
Cameroon: Baboute (n=1), Kwandja (n=15), Tikar (n=1), Wawa (n=1), Yamba (n=1)
Cameroon Northern Bantoid
Cameroon: Mambila (n=28)
Cameroon: Baboute (n=1), Mambila (n=5), Tikar (n=2), Yamba (n=16)
Cameroon: Noni (n=29), Wimbum (n=1)
Cameroon: Mambila (n=1), Nso (n=13), Wimbum (n=12)
Cameroon: Aghem (n=28)
Cameroon: Bafut (n=14), Mundum (n=1), Ngemba (n=1), Tikar (n=9)
Cameroon Grassfields
Cameroon: Ewondo (n=1), Fang (n=1), Mbo (n=12)
Cameroon: Bakweri (n=1), BaliNgonga (n=1), Bambili (n=1), Bamileke (n=7), Bamukubit (n=1), Bamun (n=26), Bangwa (n=1), Kom (n=1), Mbo (n=1), Metta (n=1), Ngemba (n=1)
Cameroon: Bakoko (n=1), Doowaayo (n=1), Ngwo (n=1)
Cameroon: Bamileke (n=1), Bamun (n=1), Bassa (n=2), Bulu (n=2), Foulbe (n=2), Kanuri (n=1), Kwandja (n=1), Mambea (n=1), Massa (n=1), Mbo (n=7), Mbororo (n=1), Ngoumba (n=1), Nso (n=1), Yamba (n=1)
Cameroon: Aghem (n=1), Bamileke (n=4), Bangwa (n=1), Bassa (n=3), Bonguili (n=1), Bulu (n=2), Ewondo (n=3), Mambea (n=1), Mbo (n=7), Moundang (n=1), Ngoumba (n=4), Nso (n=2), Wimbum (n=3), Yamba (n=1)
Cameroon: Bapokou (n=1), Bassa (n=4), Bonguili (n=4), Bulu (n=1), Ewondo (n=2), Koh (n=1), Mambea (n=3), Mbo (n=3), Ngoumba (n=8), Nso (n=1), Wimbum (n=2), Yabassi (n=1)
Cameroon: Mankon (n=1)
Cameroon Narrow Bantu
Cameroon: Foulbe (n=1), Mandara (n=1), Ngambai (n=1), Sara (n=3)
Cameroon: Doowaayo (n=1), Mambila (n=1), Mboum (n=1), Mousgoum (n=1), Ngambai (n=1), Sara (n=1). CAR: Kaba. Chad: Laka
Cameroon: Foulbe (n=1), Hausa (n=1), Kotoko (n=1). Nigeria: Igbo (n=1)
Cameroon: Baya (n=1)
Cameroon: Baboute (n=1), Bornu (n=1), Doowaayo (n=1), Guidar (n=1), Koutine (n=1), Mousgoum (n=1)
Cameroon: Arabe (n=8), Kanoubou (n=1), Kanuri (n=1), Kotoko (n=1)
Cameroon Arabe
Cameroon: Arabe (n=3), Kanuri (n=3), Kotoko (n=3)
Cameroon: Dadjo (n=1), Hadjaraye (n=1). Chad: Bulala
Cameroon Kotoko
Cameroon: Bornu (n=1), Doowaayo (n=1), Kotoko (n=20)
Cameroon: Kotoko (n=26)
Cameroon North
Cameroon: Bornu (n=1), Daba (n=2), Goude (n=1), Guidar (n=4), Guizga (n=1), Hausa (n=1), Kapsiki (n=1), Moundang (n=5), Toupouri (n=1)
Cameroon: Guidar (n=1), Massa (n=5), Mousgoum (n=3)
Cameroon Kanuri
Cameroon: Bornu (n=1), Kanuri (n=10), Kotoko (n=4), Mandara (n=3)
Cameroon: Arabe (n=1), Doowaayo (n=1), Guidar (n=2), Guizga (n=2), Mada (n=3), Mafa (n=1), Maltam (n=1), Mandara (n=1), Moloko (n=1)
Cameroon: Arabe (n=1)
Cameroon: Foulbe (n=2). Sudan: Falatta
ShumLaka
Cameroon: Baka (n=1), Bakola (n=2), Bedzan (n=2)
Cameroon: Baka (n=1)
BiakaPygmy
MbutiPygmy
Kenya: Luhya
Kenya: Bantu , Luo
Uganda: Mussese
Uganda: Baganda
Cameroon: Burundi (n=1). I8802
Kenya: Kikuyu
Tanzania: Sandawe
I8808
Tanzania: Hadza
Mende: Sierra
Gambia: Wolof
Senegal: Wolof
Senegal: Mandinka
Senegal: Manjak
Gambia: Mandinka
Gambia: Jola
Gambia: Fula
Cameroon: Foulbe (n=6), Mbororo (n=3)
Cameroon: Foulbe (n=3), Mbororo (n=4), Yamba (n=2)
Cameroon Fulani
Cameroon: Noni (n=1)
Non−Africa
Ethiopia, Sudan, Kenya, and Tanzania

Fig. S12.

fineSTRUCTURE dendrogram of the whole dataset with populations from Cameroon (orange to red), Republic of the Congo (pink), Ghana (blue), or Nigeria (grey) clustered, and the rest of the dataset grouped into 'superindividuals' (Methods). The ethnic group composition of each cluster is displayed (and can be found in Data S3). This tree was used, along with information on self-reported ethnic group, to cluster individuals from Cameroon, Republic of the Congo, Ghana, and Nigeria (Methods). The super-groups used for admixture analyses are shown in different colours, corresponding to those in Fig. 3 (Data S5).

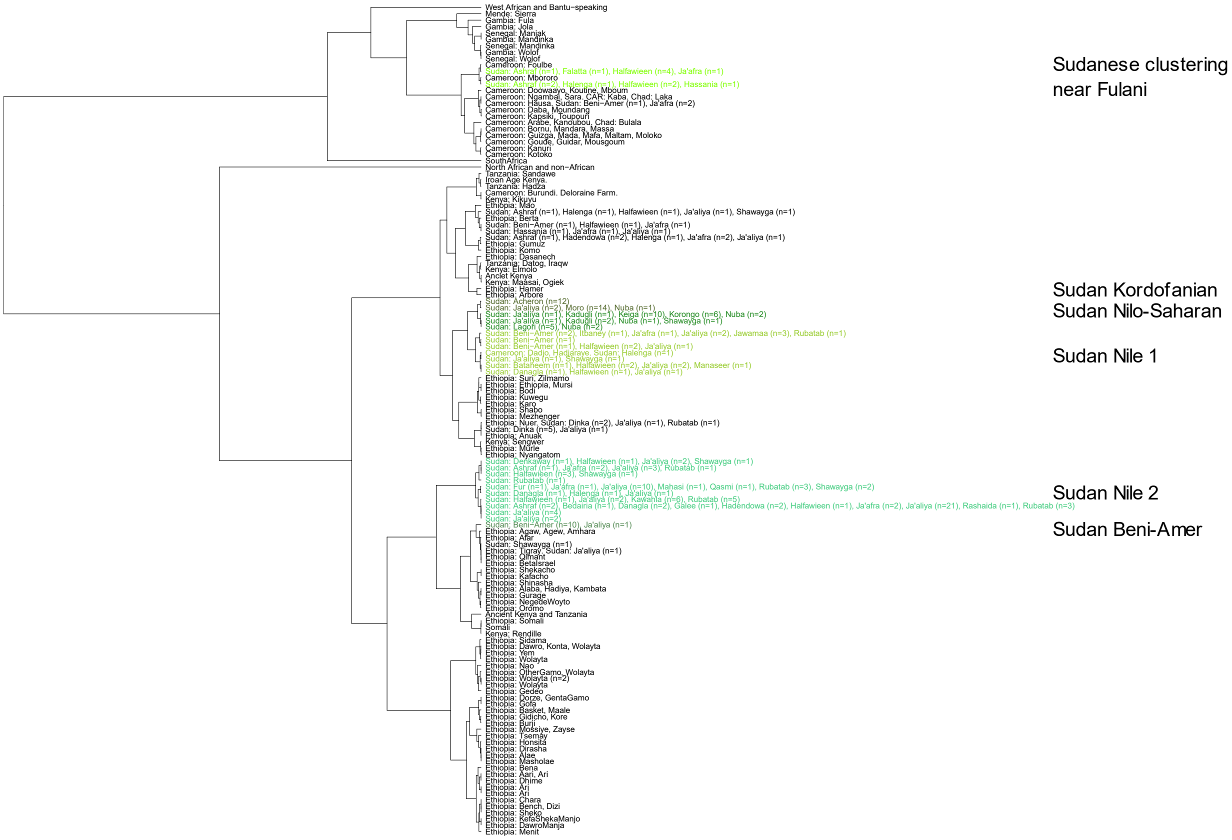

West African and Bantu−speaking
Mende: Sierra
Gambia: Fula
Gambia: Jola
Senegal: Manjak
Gambia: Mandinka
Senegal: Mandinka
Gambia: Wolof
Senegal: Wolof
Cameroon: Foulbe
Sudan: Ashraf (n=1), Falatta (n=1), Halfawieen (n=4), Ja'afra (n=1)
Cameroon: Mbororo
Sudan: Ashraf (n=2), Halenga (n=1), Halfawieen (n=2), Hassania (n=1)
Cameroon: Doowaayo, Koutine, Mboum
Cameroon: Ngambai, Sara. CAR: Kaba. Chad: Laka
Cameroon: Hausa. Sudan: Beni−Amer (n=1), Ja'afra (n=2)
Cameroon: Daba, Moundang
Cameroon: Kapsiki, Toupouri
Cameroon: Arabe, Kanoubou. Chad: Bulala
Cameroon: Bornu, Mandara, Massa
Cameroon: Guizga, Mada, Mafa, Maltam, Moloko
Cameroon: Goude, Guidar, Mousgoum
Cameroon: Kanuri
Cameroon: Kotoko
SouthAfrica
North African and non−African
Tanzania: Sandawe
Iroan Age Kenya.
Tanzania: Hadza
Cameroon: Burundi. Deloraine Farm.
Kenya: Kikuyu
Ethiopia: Mao
Sudan: Ashraf (n=1), Halenga (n=1), Halfawieen (n=1), Ja'aliya (n=1), Shawayga (n=1)
Ethiopia: Berta
Sudan: Beni−Amer (n=1), Halfawieen (n=1), Ja'afra (n=1)
Sudan: Hassania (n=1), Ja'afra (n=1), Ja'aliya (n=1)
Sudan: Ashraf (n=1), Hadendowa (n=2), Halenga (n=1), Ja'afra (n=2), Ja'aliya (n=1)
Ethiopia: Gumuz
Ethiopia: Komo
Ethiopia: Dasanech
Tanzania: Datog, Iraqw
Kenya: Elmolo
Anciet Kenya
Kenya: Maasai, Ogiek
Ethiopia: Hamer
Ethiopia: Arbore
Sudan: Acheron (n=12)
Sudan: Ja'aliya (n=2), Moro (n=14), Nuba (n=1)
Sudan: Ja'aliya (n=1), Kadugli (n=1), Keiga (n=10), Korongo (n=6), Nuba (n=2)
Sudan: Ja'aliya (n=1), Kadugli (n=2), Nuba (n=1), Shawayga (n=1)
Sudan: Lagori (n=5), Nuba (n=2)
Sudan: Beni−Amer (n=2), Itbaney (n=1), Ja'afra (n=1), Ja'aliya (n=2), Jawamaa (n=3), Rubatab (n=1)
Sudan: Beni−Amer (n=1)
Sudan: Beni−Amer (n=1), Halfawieen (n=2), Ja'aliya (n=1)
Cameroon: Dadjo, Hadjaraye. Sudan: Halenga (n=1)
Sudan: Ja'aliya (n=1), Shawayga (n=1)
Sudan: Bataheem (n=1), Halfawieen (n=2), Ja'aliya (n=2), Manaseer (n=1)
Sudan: Danagla (n=1), Halfawieen (n=1), Ja'aliya (n=1)
Ethiopia: Suri, Zilmamo
Ethiopia: Ethiopia, Mursi
Ethiopia: Bodi
Ethiopia: Kuwegu
Ethiopia: Karo
Ethiopia: Shabo
Ethiopia: Mezhenger
Ethiopia: Nuer. Sudan: Dinka (n=2), Ja'aliya (n=1), Rubatab (n=1)
Sudan: Dinka (n=5), Ja'aliya (n=1)
Ethiopia: Anuak
Kenya: Sengwer
Ethiopia: Murle
Ethiopia: Nyangatom
Sudan: Denkaway (n=1), Halfawieen (n=1), Ja'aliya (n=2), Shawayga (n=1)
Sudan: Ashraf (n=1), Ja'afra (n=2), Ja'aliya (n=3), Rubatab (n=1)
Sudan: Halfawieen (n=3), Shawayga (n=1)
Sudan: Rubatab (n=1)
Sudan: Fur (n=1), Ja'afra (n=1), Ja'aliya (n=10), Mahasi (n=1), Qasmi (n=1), Rubatab (n=3), Shawayga (n=2)
Sudan: Danagla (n=1), Halenga (n=1), Ja'aliya (n=1)
Sudan: Halfawieen (n=1), Ja'aliya (n=2), Kawahla (n=6), Rubatab (n=5)
Sudan: Ashraf (n=2), Bedairia (n=1), Danagla (n=2), Galee (n=1), Hadendowa (n=2), Halfawieen (n=1), Ja'afra (n=2), Ja'aliya (n=21), Rashaida (n=1), Rubatab (n=3)
Sudan: Ja'aliya (n=4)
Sudan: Ja'aliya (n=2)
Sudan: Beni−Amer (n=10), Ja'aliya (n=1)
Ethiopia: Agaw, Agew, Amhara
Ethiopia: Afar
Sudan: Shawayga (n=1)
Ethiopia: Tigray. Sudan: Ja'aliya (n=1)
Ethiopia: Qimant
Ethiopia: BetaIsrael
Ethiopia: Shekacho
Ethiopia: Kafacho
Ethiopia: Shinasha
Ethiopia: Alaba, Hadiya, Kambata
Ethiopia: Gurage
Ethiopia: NegedeWoyto
Ethiopia: Oromo
Ancient Kenya and Tanzania
Ethiopia: Somali
Somali
Kenya: Rendille
Ethiopia: Sidama
Ethiopia: Dawro, Konta, Wolayta
Ethiopia: Yem
Ethiopia: Wolayta
Ethiopia: Nao
Ethiopia: OtherGamo, Wolayta
Ethiopia: Wolayta (n=2)
Ethiopia: Wolayta
Ethiopia: Gedeo
Ethiopia: Dorze, GentaGamo
Ethiopia: Gofa
Ethiopia: Basket, Maale
Ethiopia: Gidicho, Kore
Ethiopia: Burji
Ethiopia: Mossiye, Zayse
Ethiopia: Tsemay
Ethiopia: Honsita
Ethiopia: Dirasha
Ethiopia: Alae
Ethiopia: Masholae
Ethiopia: Bena
Ethiopia: Aari, Ari
Ethiopia: Dhime
Ethiopia: Ari
Ethiopia: Ari
Ethiopia: Chara
Ethiopia: Bench, Dizi
Ethiopia: Sheko
Ethiopia: KefaShekaManjo
Ethiopia: DawroManja
Ethiopia: Menit
Sudanese clustering
near Fulani
Sudan Kordofanian
Sudan Nilo-Saharan
Sudan Nile 1
Sudan Nile 2
Sudan Beni-Amer

Fig. S13.
fineSTRUCTURE dendrogram of Sudanese individuals, with the rest of the dataset clustered into 'superindividuals' (Methods). The ethnic group composition of each cluster is displayed (and can be found in Data S3). This tree was used, along with information on self-reported ethnic group, to cluster individuals from Sudan (Methods). The five Sudanese 'super-groups' are highlighted in different shades of green starting from the bottom; Beni-Amer, Nile2, Nile1, Nilo-Saharan and Kordofanian. The final group at the top shows the Sudanese individuals who cluster with the Fulani from Cameroon. Differential admixture between individuals, as demonstrated in Fig. S15, is likely driving these clustering results.

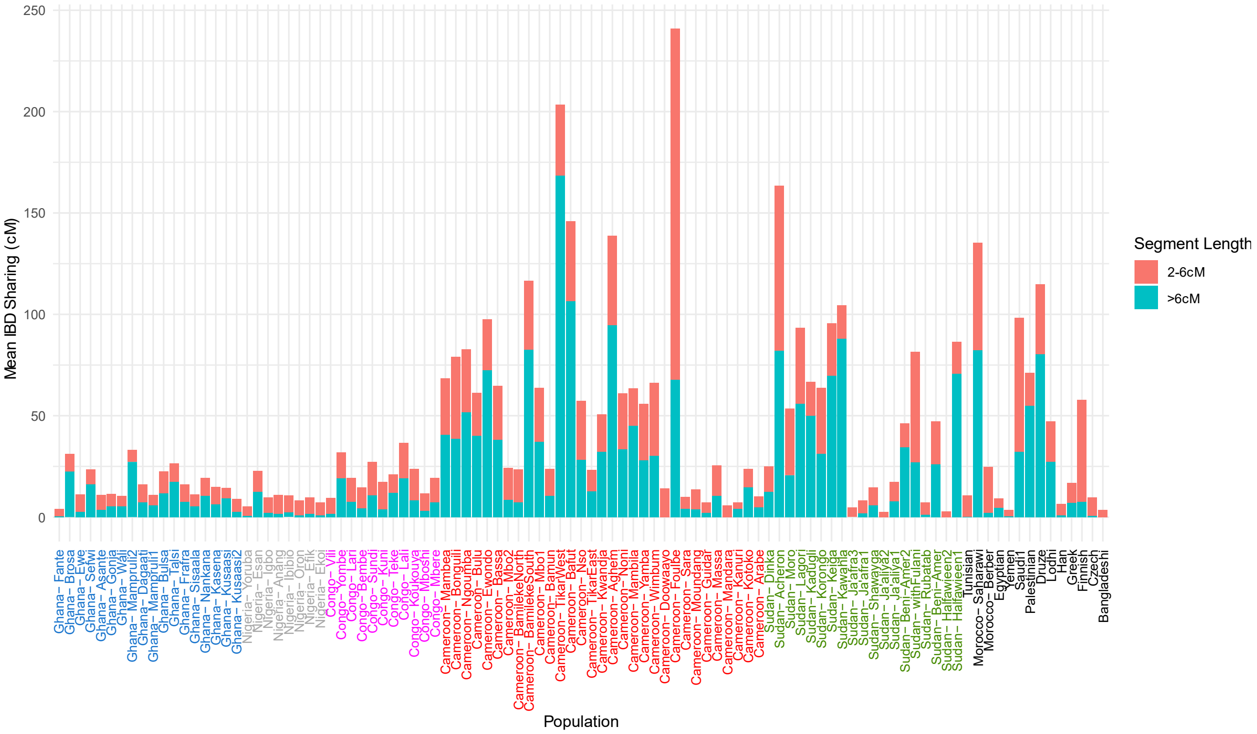


Fig. S14.

Mean IBD sharing between each pair of individuals within each cluster from Ghana, Nigeria, Republic of the Congo, Cameroon and Sudan, as well as a selection of North African and non-African clusters (black label). All clusters contain more than 3 individuals. Total IBD length has been divided into short (2-6cM) and long (>6cM) tracts.

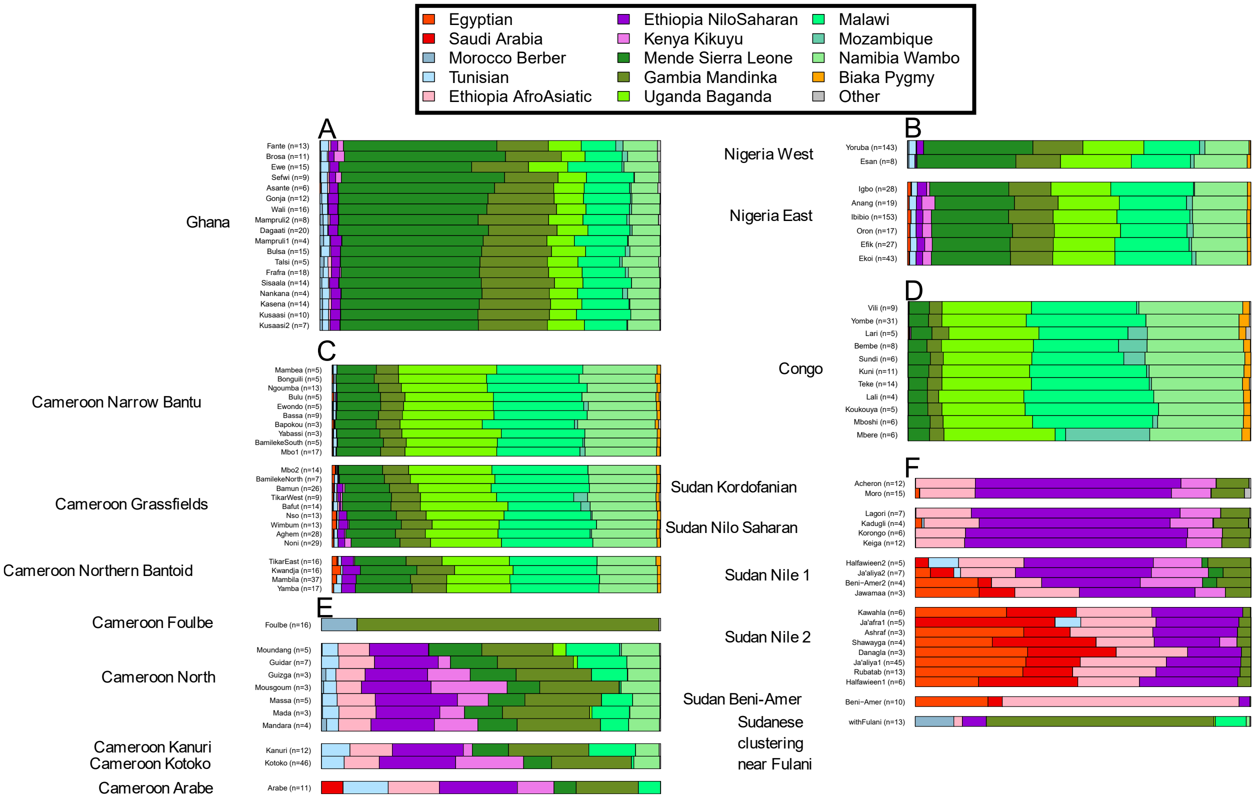


Fig. S15.

Ancestry modeled in SOURCEFIND for each cluster from Ghana (A), Nigeria (B), Cameroon (C) and (E), Republic of the Congo (D) and Sudan (F). Clusters are grouped into ‘super-groups’ as shown on the fineSTRUCTURE dendrograms in Fig. S12 and 13. Clusters within each super-group exhibit similar patterns of inferred recent ancestry sharing with reference populations. Reference populations are shown in the key and follow red=Arabic, blue=north African, pink=east African, green=west African and Bantu-speaking peoples, orange=rainforest hunter-gatherer. Note that the reference populations are themselves admixed, and so results should be interpreted with caution.

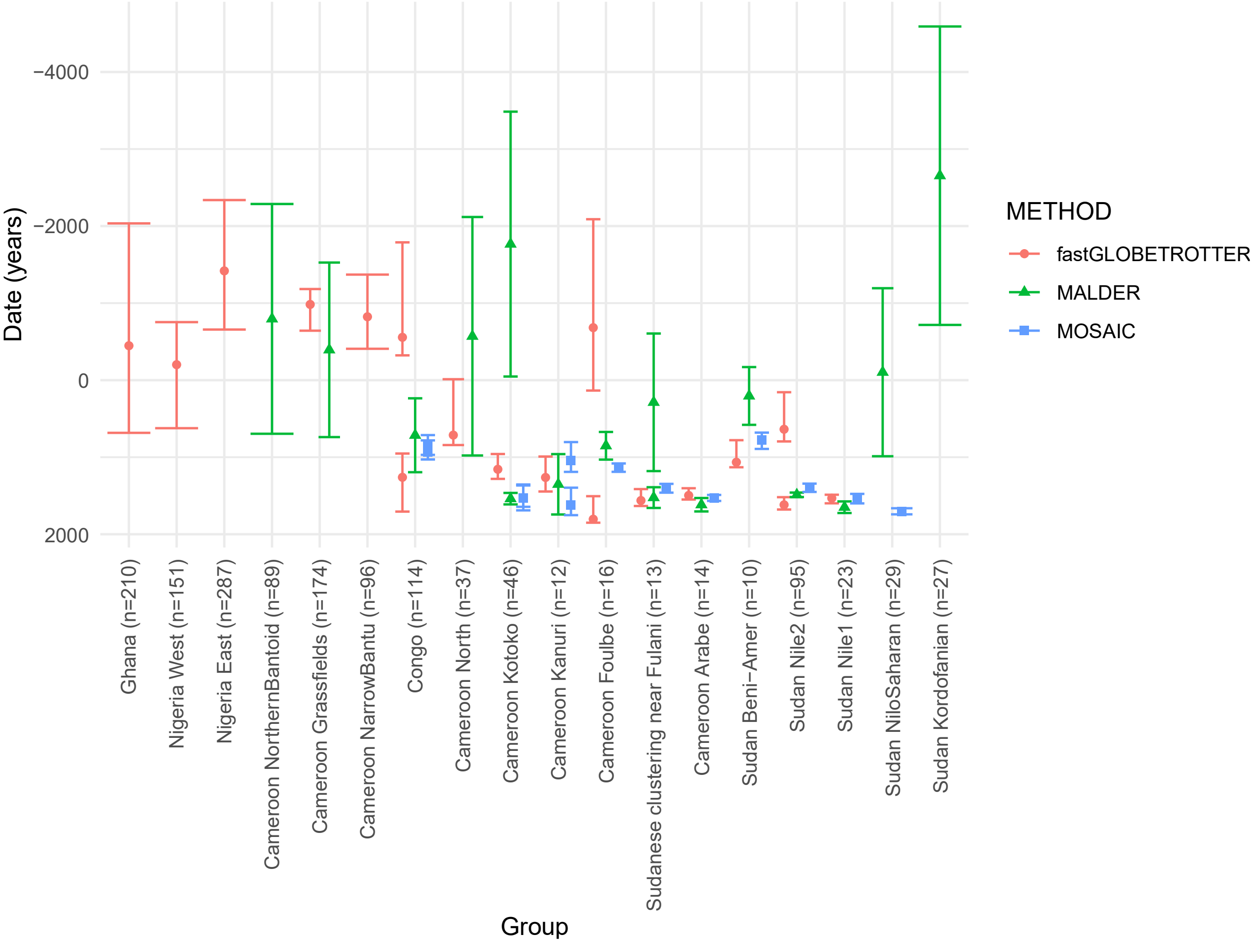


Fig. S16.

Admixture dates and their 95% confidence intervals inferred using fastGLOBETROTTER (red circle), MALDER (green triangle) and MOSAIC (blue square) for the 18 "super-groups" from Cameroon, Republic of the Congo, Ghana, Nigeria and Sudan. See Data S6 for inferred admixture sources. Inferred dates overlap for at least two methods in 10 out of 18 cases. We note that our sample sizes vary widely, which will impact our power to detect and date admixture.

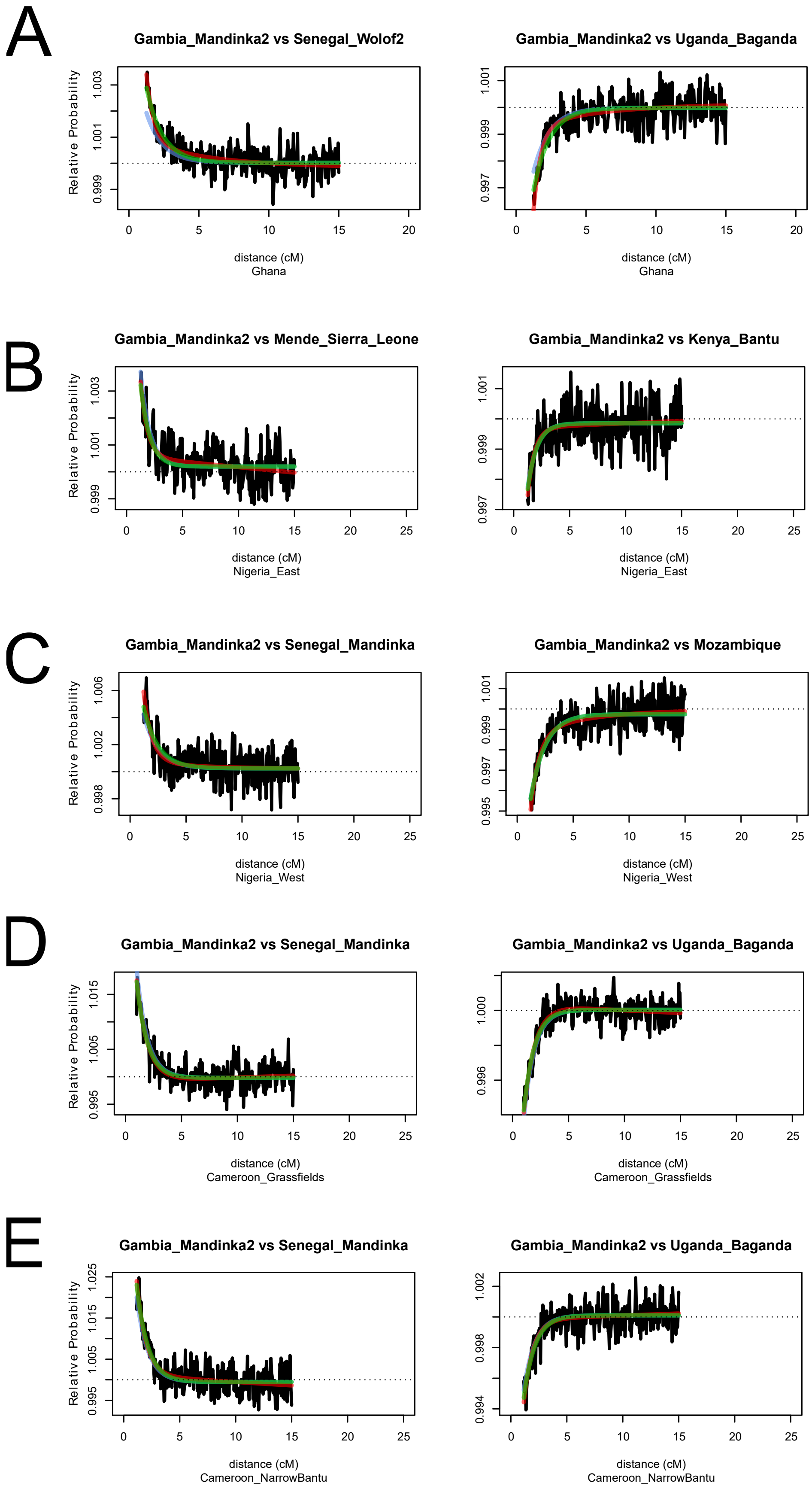

A
Gambia_Mandinka2 vs Senegal_Wolof2
Relative Probability
0.999
1.001
1.003
0
5
10
15
20
distance (cM)
Ghana
Gambia_Mandinka2 vs Uganda_Baganda
0.997
0.999
1.001
0
5
10
15
20
distance (cM)
Ghana
B
Gambia_Mandinka2 vs Mende_Sierra_Leone
Relative Probability
0.999
1.001
1.003
0
5
10
15
20
25
distance (cM)
Nigeria_East
Gambia_Mandinka2 vs Kenya_Bantu
0.997
0.999
1.001
0
5
10
15
20
25
distance (cM)
Nigeria_East
C
Gambia_Mandinka2 vs Senegal_Mandinka
Relative Probability
0.998
1.002
1.006
0
5
10
15
20
25
distance (cM)
Nigeria_West
Gambia_Mandinka2 vs Mozambique
0.995
0.997
0.999
1.001
0
5
10
15
20
25
distance (cM)
Nigeria_West
D
Gambia_Mandinka2 vs Senegal_Mandinka
Relative Probability
0.995
1.005
1.015
0
5
10
15
20
25
distance (cM)
Cameroon_Grassfields
Gambia_Mandinka2 vs Uganda_Baganda
0.996
1.000
0
5
10
15
20
25
distance (cM)
Cameroon_Grassfields
E
Gambia_Mandinka2 vs Senegal_Mandinka
Relative Probability
0.995
1.005
1.015
1.025
0
5
10
15
20
25
distance (cM)
Cameroon_NarrowBantu
Gambia_Mandinka2 vs Uganda_Baganda
0.994
0.998
1.002
0
5
10
15
20
25
distance (cM)
Cameroon_NarrowBantu

Fig. S17.

Example fastGLOBETROTTER coancestry curves for the admixture event inferred in (A) Ghana, (B) Nigeria East, (C) Nigeria West, (D) Cameroon Grassfields and (E) Cameroon Narrow Bantu. The black line depicts the (scaled) inferred probability that two DNA segments separated by distance X in a target population individual were inherited from sources related to (respectively) the two reference populations given in the plot's title. Curves decreasing with distance indicate the two reference populations represent the same admixing source, while increasing curves indicate the two reference populations represent different admixing sources. The green and red lines show the model fit assuming one or two pulses of admixture, respectively.

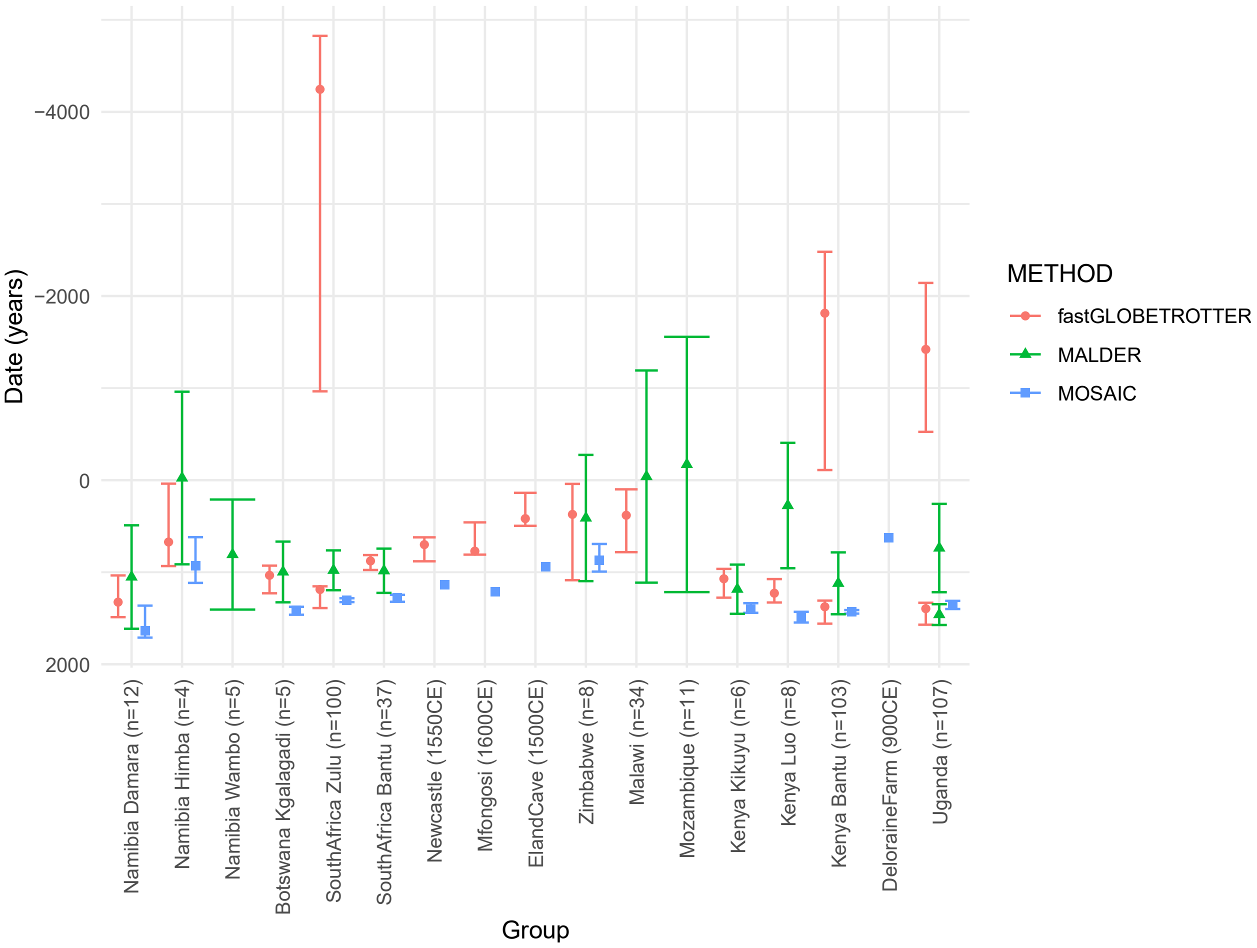


Fig. S18.

Inferred admixture dates and their 95% confidence intervals inferred using fastGLOBETROTTER (red circle), MALDER (green triangle) and MOSAIC (blue square) for the 13 groups and 4 ancient individuals with Bantu speaking-related ancestry. Dates for the ancient individuals have been adjusted using their carbon dates (Deloraine Farm 900CE, Newcastle 1550CE, Mfongosi 1600CE, Eland Cave 1500CE). MALDER cannot infer admixture dates using a single sample, and MOSAIC cannot infer date confidence intervals in such cases, so these results are not shown for the ancient individuals. See Data S8 for inferred admixture sources. Inferred dates overlap for at least two methods in 10 out of 13 cases. We note that sample sizes, shown in the axis labels, vary widely, which means power to infer admixture also varies widely. In groups with very large sample sizes (Zulu, Uganda, Kenya Bantu), multiple dates of admixture are inferred, potentially because of the increased power. Hence this means that in other groups, we may simply lack the power to detect additional admixture events.

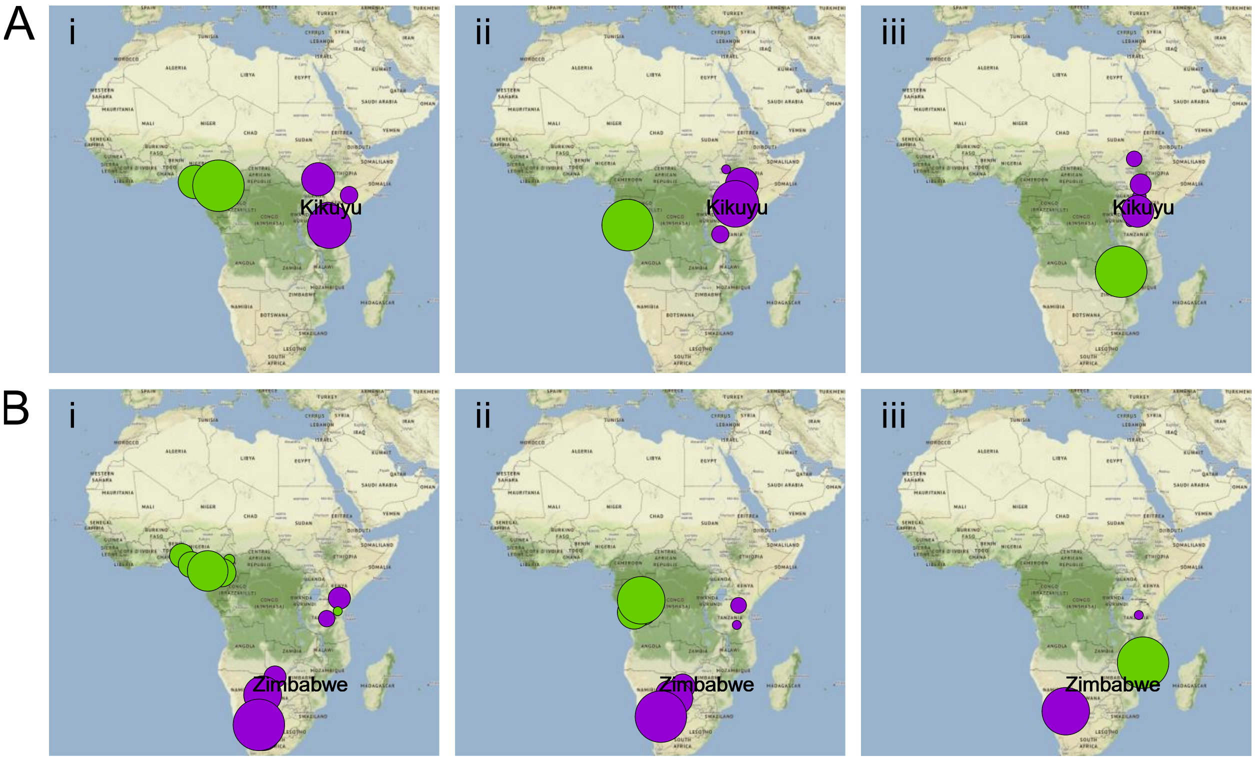


Fig. S19.

fastGLOBETROTTER results for (A) the Kikuyu and (B) Zimbabweans, when using non Bantu-speaking groups as surrogates plus (i) nothing else, (ii) Congolese Bantu speakers, and (iii) all Bantu-speakers in the dataset. This is analogous to analyses in (*5*, *15*, *23*, *30*)). The inferred Bantu-speaking-like admixing source is represented with green, and the other inferred admixing source is represented with purple. Size of point indicates the percentage contribution of that surrogate population to describe the genetic make-up each admixing source. These results provide evidence for the late split model, as Bantu-speaking-related ancestry (green) is more closely related to Congolese than Cameroonians, and more closely related to Malawians and Mozambicans than Congolese, consistent with the migration being first southwards and eastwards before splitting into two branches.

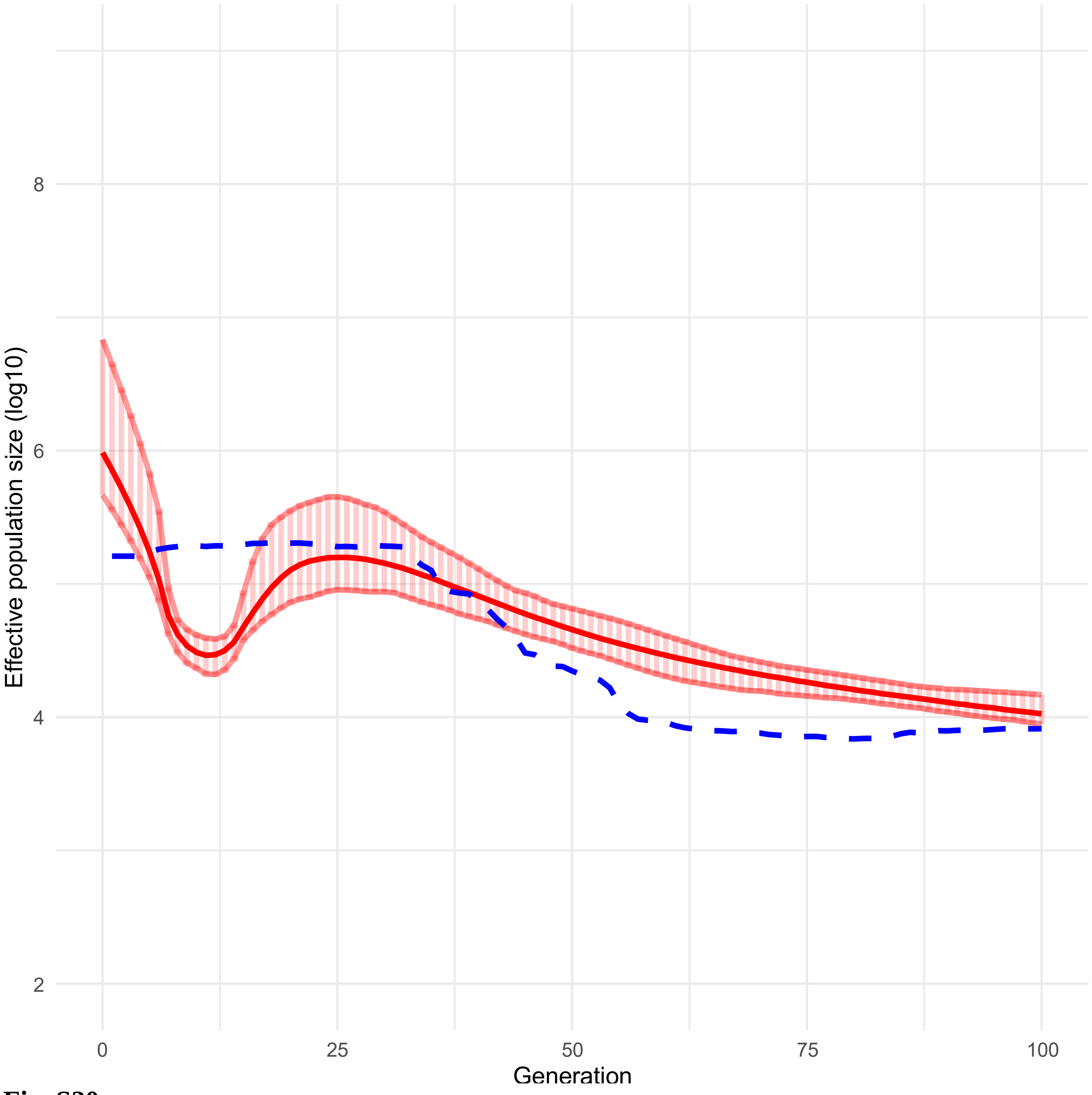


Fig. S20.

Effective population size (N_e) changes in the history of individuals from the Republic of the Congo, inferred using GONE (blue dotted) and IBDNe (red). IBDNe 95% confidence intervals are shown as transparent red lines. Results are shown for up to 100 generations ago, as inference any earlier than this can be unreliable.

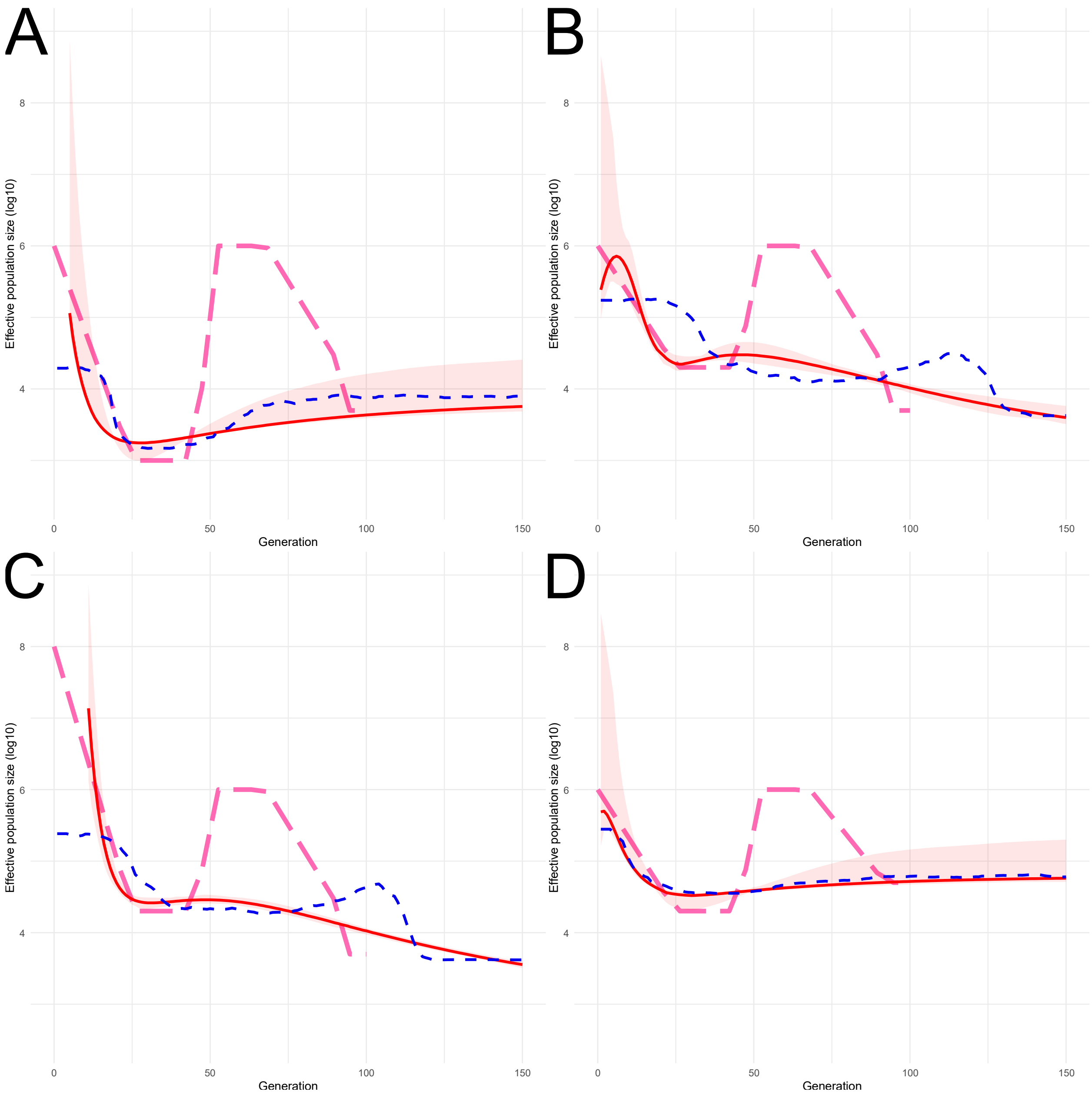


Fig. S21.

Simulations in msprime (Methods, text S4) of a population expansion, followed by a bottleneck, followed by an expansion, to mimic the proposed two wave expansion into the Congo basin. The four simulations (panels) vary by the strength of bottleneck and expansions, with the pink dashed line indicating the effective population size trajectory simulated. The red and blue dashed lines depict inference using IBDNe (considering inferred IBD segments > 2cM) and GONE, respectively, with shading showing 95% confidence intervals for IBDNe. Only N_e values and 95% confidence intervals between 10^2.5 and 10^9 are shown. In all scenarios (A)-(D), neither IBDNe nor GONE are able to capture the true population dynamics, with IBDNe often inferring relatively constant population size changes until recently.

Data S1. (separate file)
Samples published in this study, including self-described ethnic group, two levels of language classification (as defined by glottolog and ethnologue), number of individuals from that population, and the mean latitude and longitude for all individuals in that population (see Fig. S1 for map).

Data S2. (separate file)
Previously published data included in this study (shown in Fig. S2), including both present day and ancient samples.

Data S3. (separate file)
Raw fineSTRUCTURE results, showing the population composition of each cluster and which analysis they were inferred in.

Data S4. (separate file)
Clusters defined using both fineSTRUCTURE results and each individuals' self-described ethnic group (Methods), with the final eight columns showing which analyses each cluster acted as a source in.

Data S5. (separate file)
Composition of the 18 “super-groups” formed by merging clusters up the fineSTRUCTURE dendrogram, used for admixture and ancestry analysis (Fig. 3 and 5).

Data S6. (separate file)
fastGLOBETROTTER, MOSAIC, MALDER and SOURCEFIND results for the 18 super-groups.

Data S7. (separate file)
fastGLOBETROTTER, MALDER and MOSAIC results for extra analyses of Sudanese clustering near Fulani to test specific hypothesis of admixture between Cameroonian populations.

Data S8. (separate file)
fastGLOBETROTTER, MOSAIC, MALDER and SOURCEFIND results for the 13 groups with Bantu-speaking related ancestry and 4 ancient individuals.