The Genetic Structure and History of Africans and African Americans

Sarah A. Tishkoff,[1,2]* Floyd A. Reed,[1]†‡ Françoise R. Friedlaender,[3]‡ Christopher Ehret,[4] Alessia Ranciaro,[1,2,5]§ Alain Froment,[6]§ Jibril B. Hirbo,[1,2] Agnes A. Awomoyi,[1]‖ Jean-Marie Bodo,[7] Ogobara Doumbo,[8] Muntaser Ibrahim,[9] Abdalla T. Juma,[9] Maritha J. Kotze,[10] Godfrey Lema,[11] Jason H. Moore,[12] Holly Mortensen,[1]¶ Thomas B. Nyambo,[11] Sabah A. Omar,[13] Kweli Powell,[1]# Gideon S. Pretorius,[14] Michael W. Smith,[15] Mahamadou A. Thera,[8] Charles Wambebe,[16] James L. Weber,[17] Scott M. Williams[18]

Africa is the source of all modern humans, but characterization of genetic variation and of relationships among populations across the continent has been enigmatic. We studied 121 African populations, four African American populations, and 60 non-African populations for patterns of variation at 1327 nuclear microsatellite and insertion/deletion markers. We identified 14 ancestral population clusters in Africa that correlate with self-described ethnicity and shared cultural and/or linguistic properties. We observed high levels of mixed ancestry in most populations, reflecting historical migration events across the continent. Our data also provide evidence for shared ancestry among geographically diverse hunter-gatherer populations (Khoesan speakers and Pygmies). The ancestry of African Americans is predominantly from Niger-Kordofanian (~71%), European (~13%), and other African (~8%) populations, although admixture levels varied considerably among individuals. This study helps tease apart the complex evolutionary history of Africans and African Americans, aiding both anthropological and genetic epidemiologic studies.

Modern humans originated in Africa ~200,000 years ago and then spread across the rest of the globe within the past ~100,000 years (*1*). Thus, modern humans have existed continuously in Africa longer than in any other geographic region and have maintained relatively large effective population sizes, resulting in high levels of within-population genetic diversity (*1*, *2*). Africa contains more than 2000 distinct ethnolinguistic groups representing nearly one-third of the world's languages (*3*). Except for a few isolates that show no clear relationship with other languages, these languages have been classified into four major macrofamilies: Niger-Kordofanian (spoken across a broad region of Africa), Afroasiatic (spoken predominantly in Saharan, northeastern, and eastern Africa), Nilo-Saharan (spoken predominantly in Sudanic, Saharan, and eastern Africa), and Khoesan (languages containing click-consonants, spoken by San in southern Africa and by Hadza and Sandawe in eastern Africa) (fig. S1) (*4*).

Despite the importance of African population genetics, the pattern of genome-wide nuclear genetic diversity across geographically and ethnically diverse African populations is largely uncharacterized (*1*, *2*, *5*). Because of considerable environmental diversity, African populations show a range of linguistic, cultural, and phenotypic variation (*1*, *2*, *4*). Characterizing the pattern of genetic variation among ethnically diverse African populations is critical for reconstructing human evolutionary history, clarifying the population history of Africans and African Americans, and determining the proper design and interpretation of genetic disease association studies (*1*, *6*), because substructure can cause spurious results (*7*). Furthermore, variants associated with disease could be geographically restricted as a result of new mutations, genetic drift, or region-specific selection pressures (*1*). Thus, our in-depth characterization of genetic structure in Africa benefits research of biomedical relevance in both African and African-diaspora populations.

We genotyped a panel of 1327 polymorphic markers, consisting of 848 microsatellites, 476 indels (insertions/deletions), and three SNPs (single-nucleotide polymorphisms), in 2432 Africans from 113 geographically diverse populations (fig. S1), 98 African Americans, and 21 Yemenites (table S1). To incorporate preexisting African data and to place African genetic variability into a worldwide context, we integrated these data with data from the panel of markers genotyped in 952 worldwide individuals from the CEPH-HGDP (Centre d'Étude du Polymorphisme Humain–Human Genome Diversity Panel) (*8*–*10*) in 432 individuals of Indian descent (*11*) and in 10 Native Australians (tables S1 and S2).

African variation in a worldwide context. African and African American populations, with the exception of the Dogon of Mali, show the highest levels of within-population genetic diversity ($\theta = 4N_e\mu$, where θ is the level of genetic diversity based on variance of microsatellite allele length, N_e is the effective population size, and μ is the microsatellite mutation rate) (figs. S2 and S3). In addition, genetic diversity declines with distance from Africa (fig. S2, A to C), consistent with proposed serial founder effects resulting from the migration of modern humans out of Africa and across the globe (*9*, *11*–*13*). Within Africa, genetic diversity estimated from expected heterozygosity significantly correlates with estimates from microsatellite variance (fig. S4) (*4*) and varies by linguistic, geographic, and subsistence classifications (fig. S5). Three hunter-gatherer populations (Baka Pygmies, Bakola Pygmies, and San) were among the five populations with the highest levels of genetic diversity based on variance estimates (fig. S2A) (*4*). In addition, more private alleles exist in Africa than in other regions (fig. S6A). Consistent with bidirectional gene flow (*14*), African and Middle Eastern populations shared the greatest number of alleles absent from all other populations (fig. S6B). Within Africa, the most private alleles were in southern Africa, reflecting those in southern African Khoesan (SAK) San and !Xun/Khwe populations (fig. S6C) (*12*). Eastern and Saharan Africans shared the most alleles absent from other African populations examined (fig. S6D).

The proportion of genetic variation among all studied African populations was 1.71% (table S3). In comparison, Native American and Oceanic populations showed the greatest proportion of genetic variation among populations (8.36% and 4.59%, respectively), most likely due to genetic drift (*9*, *15*, *16*). Distinct patterns of the distribution of variation among African populations classified by geography, language, and subsistence were also observed (*4*).

---

[1] Department of Biology, University of Maryland, College Park, MD 20742, USA. [2] Departments of Genetics and Biology, University of Pennsylvania, Philadelphia, PA 19104, USA. [3] Independent researcher, Sharon, CT 06069, USA. [4] Department of History, University of California, Los Angeles, CA 90095, USA. [5] Dipartimento di Biologia ed Evoluzione, Università di Ferrara, 44100 Ferrara, Italy. [6] UMR 208, IRD-MNHN, Musée de l'Homme, 75116 Paris, France. [7] Ministère de la Recherche Scientifique et de l'Innovation, BP 1457, Yaoundé, Cameroon. [8] Malaria Research and Training Center, University of Bamako, Bamako, Mali. [9] Department of Molecular Biology, Institute of Endemic Diseases, University of Khartoum, 15-13 Khartoum, Sudan. [10] Department of Pathology, Faculty of Health Sciences, University of Stellenbosch, Tygerberg 7505, South Africa. [11] Department of Biochemistry, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania. [12] Departments of Genetics and Community and Family Medicine, Dartmouth Medical School, Lebanon, NH 03756, USA. [13] Kenya Medical Research Institute, Center for Biotechnology Research and Development, 54840-00200 Nairobi, Kenya. [14] Division of Human Genetics, Faculty of Health Sciences, University of Stellenbosch, Tygerberg 7505, South Africa. [15] Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702, USA. [16] International Biomedical Research in Africa, Abuja, Nigeria. [17] Marshfield Clinic Research Foundation, Marshfield, WI 54449, USA. [18] Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37232, USA.

*To whom correspondence should be addressed. E-mail: tishkoff@mail.med.upenn.edu

†Present address: Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany.

‡These authors contributed equally to this work.

§These authors contributed equally to this work.

‖Present address: Department of Internal Medicine, Ohio State University, Columbus, OH 43210, USA.

¶Present address: Office of Research and Development, National Center for Computational Toxicology, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA.

#Present address: College of Education, University of Maryland, College Park, MD 20742, USA.

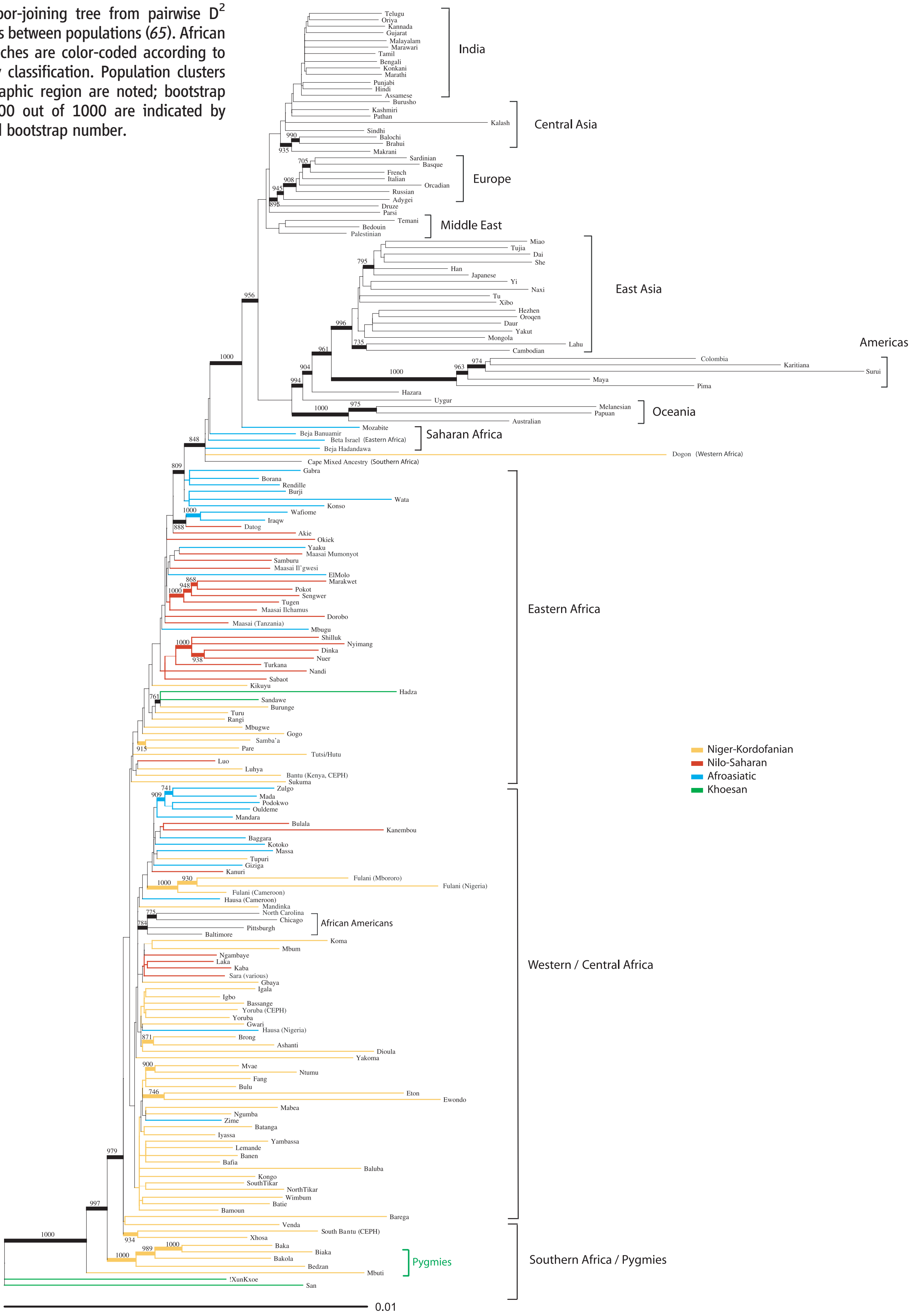


Fig. 1. Neighbor-joining tree from pairwise D^2 genetic distances between populations (*65*). African population branches are color-coded according to language family classification. Population clusters by major geographic region are noted; bootstrap values above 700 out of 1000 are indicated by thicker lines and bootstrap number.

Phylogenetic trees constructed from genetic distances between populations generally showed clustering by major geographic region, both on a global scale and within Africa (Fig. 1 and figs. S7 and S8). Within Africa, the two SAK populations cluster together and are the most distant from other populations, consistent with mitochondrial DNA (mtDNA), Y chromosome, and autosomal chromosome diversity studies, indicating that SAK populations have the most diverged genetic lineages (*12*, *17–21*). The Pygmy populations cluster near the SAK populations in the tree constructed from D^2 genetic distances (Fig. 1), whereas the Hadza and Sandawe cluster near the SAK populations in the tree constructed from R_{ST} genetic distances (fig. S8) (*4*). Note that population clustering in the tree may reflect common ancestry and/or admixture. African populations with high levels of non-African admixture [e.g., the Cape Mixed Ancestry (CMA) population, commonly referred to as "Cape Coloured" in South Africa] cluster in positions that are intermediate between Africans and non-Africans, whereas the African American populations, which are relatively less admixed with non-Africans, cluster more closely with West Africans. Additionally, populations with high levels of genetic drift (i.e., the Americas, Oceania, and Pygmy, Hadza, and SAK hunter-gatherers) have longer branch lengths.

Geographic distances (great circle routes) and genetic distances $(\delta\mu)^2$ between population pairs were significantly correlated, consistent with an isolation-by-distance model (figs. S9 to S11 and table S4) (*13*). A heterogeneous pattern of correlations across global regions was observed, consistent with a previous study (*16*); the strongest correlations were in Europe and the Middle East (Spearman's $\rho = 0.88$ and 0.83, respectively; $P \leq 0.0001$ for both), followed by Africa (Spearman's $\rho = 0.40$; $P < 0.0001$). Correlations were not significant for central Asia or India. Within Africa, the strongest correlations between genetic and geographic distances were in Saharan Africa and central Africa (Spearman's $\rho = 0.76$ and 0.55, respectively; $P < 0.0001$ for both) (fig. S11 and table S4). The smallest correlation was observed in eastern Africa ($\rho = 0.19$; $P < 0.0001$).

Genetic structure on a global level. Global patterns of genetic structure and individual ancestry were inferred by principal components analysis (PCA) (*22*) (Fig. 2A) and a Bayesian model-based clustering approach with STRUCTURE (*23*) (Figs. 3 and 4 and figs. S12 to S14). Worldwide, 72 significant principal components (PCs) were identified by PCA ($P < 0.05$) (*22*). PC1 (accounting for 19.5% of the extracted variation) distinguishes Africans from non-Africans. The CMA and African American individuals cluster between Africans and non-Africans, reflecting both African and non-African ancestry. PC2 (5.01%) distinguishes Oceanians, East Asians, and Native Americans from others. PC3 (3.5%) distinguishes the Hadza hunter-gatherers from others. The remaining PCs each extract less than 3% of the variation, and the 22nd to 72nd PCs extract less than 1% combined, with some minor PCs corresponding to regional and/or ethnically defined populations, consistent with STRUCTURE results below.

STRUCTURE analysis revealed 14 ancestral population clusters ($K = 14$) on a global level (Figs. 3 and 4) (*4*). Middle Eastern and Oceanic populations exhibit low levels of East African ancestry up to $K = 8$, consistent with possible gene flow into these regions and with studies suggesting early migration of modern humans into southern Asia and Oceania (*16*, *24*). The Hadza, and to a lesser extent the Pygmy, SAK, and Sandawe hunter-gatherers, are distinguished at $K = 5$. The 11th cluster ($K = 11$) distinguishes Mbuti Pygmy and SAK individuals, indicating common ancestry of these geographically distant hunter-gatherers. A number of Africans (predominantly CMA, Fulani, and eastern Afroasiatic speakers) exhibit low to moderate levels of European–Middle Eastern ancestry, consistent with possible gene flow from those regions. We found more African substructure on a global level (nine clusters) than previously observed (*9–12*, *20*). A phylogenetic tree of genetic distances from inferred ancestral clusters (fig. S14) indicates that within Africa, the Pygmy and SAK associated ancestral clusters (AACs) form a clade, as do the Hadza and Sandawe AACs and the Nilo-Saharan and Chadic AACs, reflecting their ancient common ancestries.

Genetic structure within Africa. PCA of genetic variation within Africa indicated the presence of 43 significant PCs ($P < 0.05$ with a Tracy-Widom distribution). PC1 (10.8% of the extracted variation) distinguishes eastern and Saharan Africa from western, central, and southern Africa (Fig. 2B). The second PC (6.1%) distinguishes the Hadza; the third PC (4.9%) distinguishes Pygmy and SAK individuals from other Africans. The fourth PC (3.7%) is associated with the Mozabites, some Dogon, and the CMA individuals, who show ancestry from the European–Middle Eastern cluster. The fifth PC (3.1%) is associated with SAK speakers. The 10th PC was of particular interest (2.2%) because it associates with the SAK, Sandawe, and some Dogon individuals, suggesting shared ancestry.

We incorporated geographic data into a Bayesian clustering analysis, assuming no admixture (TESS software) (*25*) and distinguished six clusters within continental Africa (Fig. 5A). The most geographically widespread cluster (orange) extends from far Western Africa (the Mandinka) through central Africa to the Bantu speakers of South Africa (the Venda and Xhosa) and corresponds to the distribution of the Niger-Kordofanian language family, possibly reflecting the spread of Bantu-speaking populations from near the Nigerian/Cameroon highlands across eastern and southern Africa within the past 5000 to 3000 years (*26*, *27*). Another inferred cluster includes the Pygmy and SAK populations (green), with a noncontiguous geographic distribution in central and southeastern Africa, consistent with the STRUCTURE (Fig. 3) and phylogenetic analyses (Fig. 1). Another geographically contiguous cluster extends across northern Africa (blue) into Mali (the Dogon), Ethiopia, and northern Kenya. With the exception of the

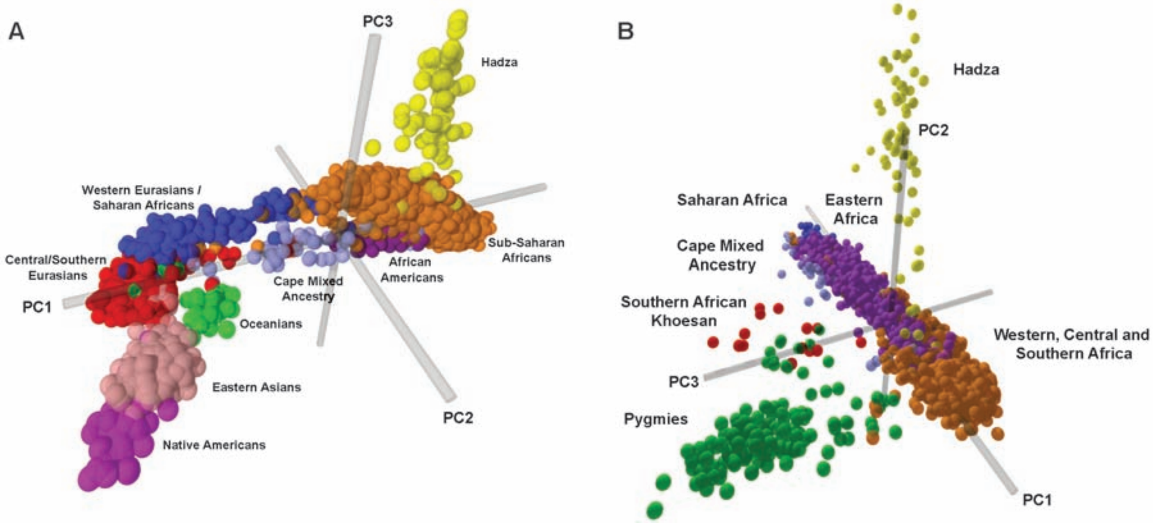


Fig. 2. Principal components analysis (*22*) created on the basis of individual genotypes. (**A**) Global data set and (**B**) African data set.

Dogon, these populations speak an Afroasiatic language. Chadic-speaking and Nilo-Saharan–speaking populations from Nigeria, Cameroon, and central Chad, as well as several Nilo-Saharan–speaking populations from southern Sudan, constitute another cluster (red). Nilo-Saharan and Cushitic speakers from the Sudan, Kenya, and Tanzania, as well as some of the Bantu speakers from Kenya, Tanzania, and Rwanda (Hutu/Tutsi),

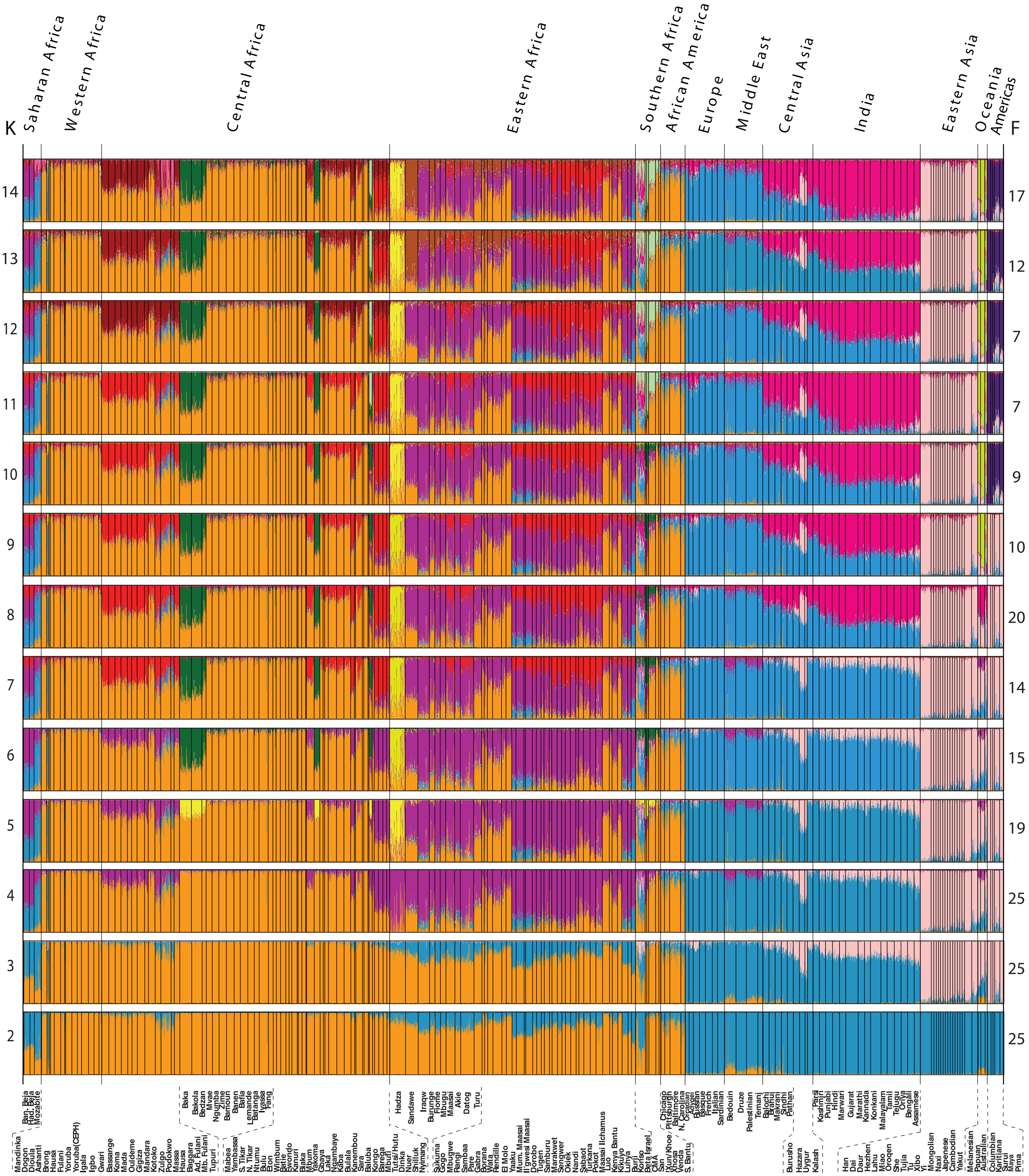


Fig. 3. STRUCTURE analysis of the global data set with 1327 markers genotyped in 3945 individuals. Each vertical line represents an individual. Individuals were grouped by self-identified ethnic group (at bottom) and ethnic groups are clustered by major geographic region (at top). Colors represent the inferred ancestry from K ancestral populations. STRUCTURE results for K = 2 to 14 (left) are shown with the number of similar runs (F) for the primary mode of 25 STRUCTURE runs at each K value (right).

constitute another cluster (purple), reflecting linguistic evidence for gene flow among these populations over the past ~5000 years (*28*, *29*). Finally, the Hadza are the sole constituents of a sixth cluster (yellow), consistent with their distinctive genetic structure identified by PCA and STRUCTURE.

STRUCTURE analysis of the Africa data set indicated 14 ancestral clusters (Fig. 5, B and C, and figs. S15 to S18). Analyses of subregions within Africa indicated additional substructure (figs. S19 to S29). At low *K* values, the Africa-wide STRUCTURE results (fig. S15) recapitulated the PCA and worldwide STRUCTURE results. However, as *K* increased, additional population clusters were distinguished (*4*): the Mbugu [who speak a mixed Bantu and Cushitic language (*30*), shown in dark purple]; Cushitic-speaking individuals of southern Ethiopian origin (light purple); Nilotic Nilo-Saharan–speaking individuals (red); central Sudanic Nilo-Saharan–speaking individuals (tan); and Chadic-speaking and Baggara individuals (maroon). At *K* = 14, subtle substructure between East African Bantu speakers (light orange) and West Central African Bantu speakers (medium orange), and individuals from Nigeria and farther west, who speak various non-Bantu Niger-Kordofanian languages (dark orange), was also apparent (Fig. 5, B and C). Bantu speakers of South Africa (Xhosa, Venda) showed substantial levels of the SAK and western African Bantu AACs and low levels of the East African Bantu AAC (the latter is also present in Bantu speakers from Democratic Republic of Congo and Rwanda). Our results indicate distinct East African Bantu migration into southern Africa and are consistent with linguistic and archeological evidence of East African Bantu migration from an area west of Lake Victoria (*28*) and the incorporation of Khoekhoe ancestry into several of the Southeast Bantu populations ~1500 to 1000 years ago (*31*).

High levels of heterogeneous ancestry (i.e., multiple cluster assignments) were observed in nearly all African individuals, with the exception of western and central African Niger-Kordofanian speakers (medium orange), who are relatively homogeneous at large *K* values (Fig. 5C and fig.

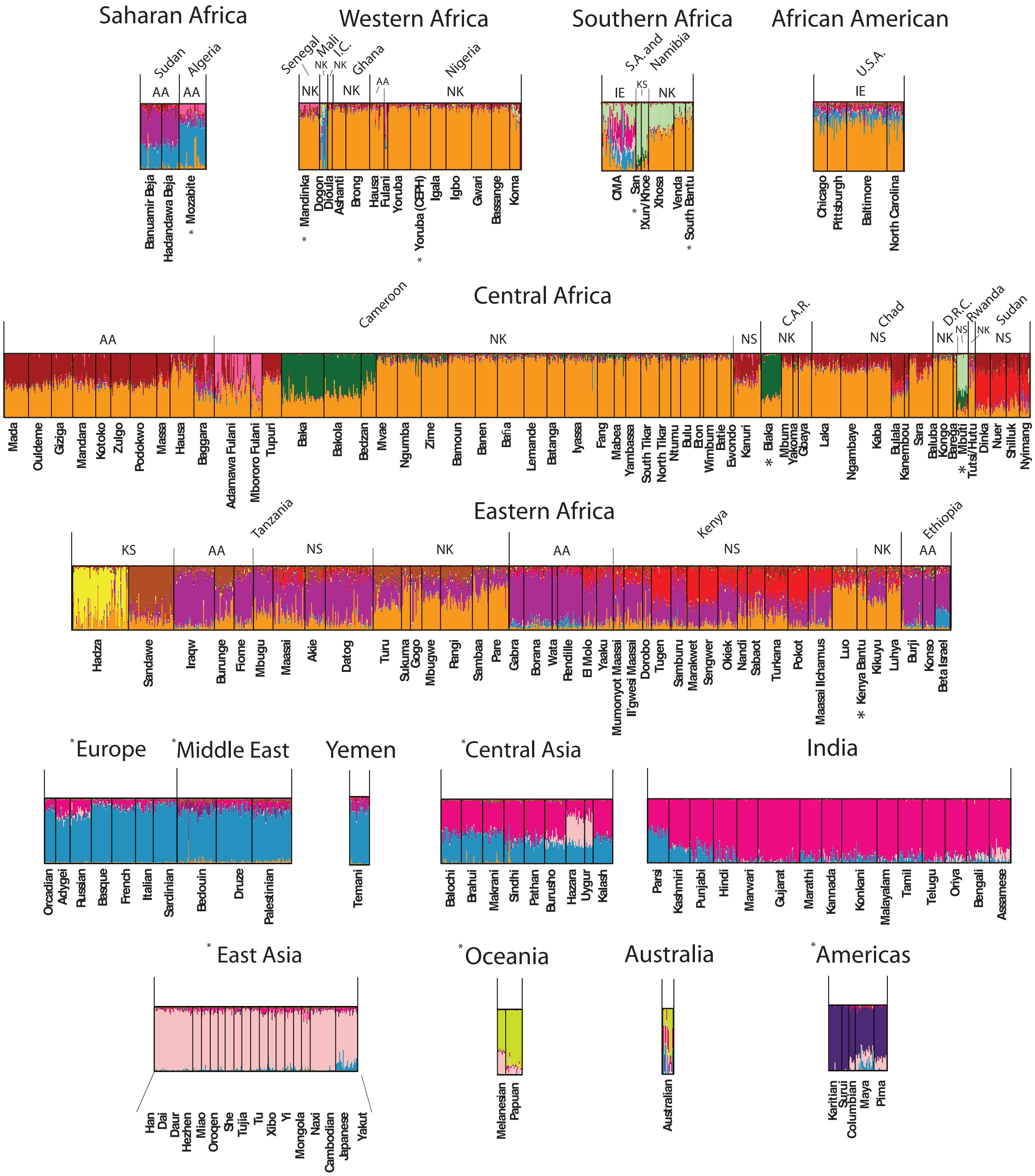
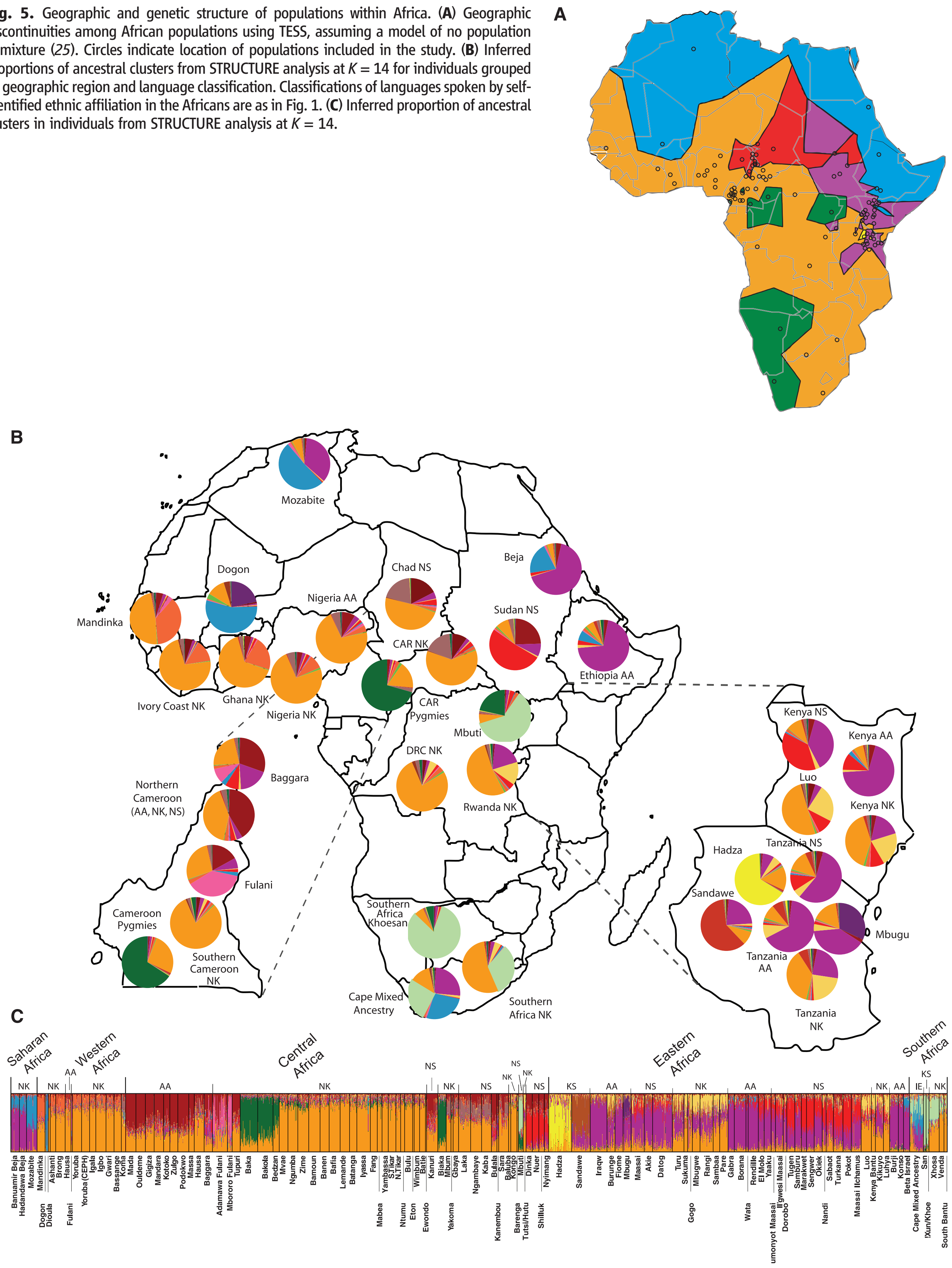


Fig. 4. Expanded view of STRUCTURE results at $K = 14$. Populations from the CEPH diversity panel are identified by asterisks. Languages spoken by populations are classified as Niger-Kordofanian (NK), Nilo-Saharan (NS), Afroasiatic (AA), Khoesan (KS), or Indo-European (IE).

Fig. 5. Geographic and genetic structure of populations within Africa. (**A**) Geographic discontinuities among African populations using TESS, assuming a model of no population admixture (*25*). Circles indicate location of populations included in the study. (**B**) Inferred proportions of ancestral clusters from STRUCTURE analysis at $K = 14$ for individuals grouped by geographic region and language classification. Classifications of languages spoken by self-identified ethnic affiliation in the Africans are as in Fig. 1. (**C**) Inferred proportion of ancestral clusters in individuals from STRUCTURE analysis at $K = 14$.



S15). Considerable Niger-Kordofanian ancestry (shades of orange) was observed in nearly all populations, reflecting the recent spread of Bantu speakers across equatorial, eastern, and southern Africa (*27*) and subsequent admixture with local populations (*28*). Many Nilo-Saharan–speaking populations in East Africa, such as the Maasai, show multiple cluster assignments from the Nilo-Saharan (red) and Cushitic (dark purple) AACs, in accord with linguistic evidence of repeated Nilotic assimilation of Cushites over the past 3000 years (*32*) and with the high frequency of a shared East African–specific mutation associated with lactose tolerance (*33*).

Our data support the hypothesis that the Sahel has been a corridor for bidirectional migration between eastern and western Africa (*34*–*36*). The highest proportion of the Nilo-Saharan AAC was observed in the southern and central Sudanese populations (Nuer, Dinka, Shilluk, and Nyimang), with decreasing frequency from northern Kenya (e.g., Pokot) to northern Tanzania (Datog, Maasai) (Fig. 5, B and C, and fig. S15). Additionally, all Nilo-Saharan–speaking populations from Kenya, Tanzania, southern Sudan, and Chad clustered with west central Afroasiatic Chadic–speaking populations in the global analysis at $K \leq 11$ (Fig. 3), which is consistent with linguistic and archeological data suggesting bidirectional migration of Nilo-Saharans from source populations in Sudan within the past ~10,500 to 3000 years (*4*, *29*). The proposed migration of proto-Chadic Afroasiatic speakers ~7000 years ago from the central Sahara into the Lake Chad Basin may have resulted in a Nilo-Saharan to Afroasiatic language shift among Chadic speakers (*37*). However, our data suggest that this shift was not accompanied by large amounts of Afroasiatic gene flow. Other populations of interest, including the Fulani (Nigeria and Cameroon), the Baggara Arabs (Cameroon), the Koma (Nigeria), and Beja (Sudan), are discussed in (*4*).

Genetic structure in East Africa. East Africa, the hypothesized origin of the migration of modern humans out of Africa, has a remarkable degree of ethnic and linguistic diversity, as reflected by the greatest level of regional substructure in Africa (figs. S15, S16, and S19 to S21). The diversity among populations from this region reflects the proposed long-term presence of click-speaking Hadza and Sandawe hunter-gatherers and successive waves of immigration of Cushitic, Nilotic, and Bantu populations within the past 5000 years (*4*, *29*, *32*, *38*, *39*). Within eastern Africa, including southern and central Sudan, clustering is primarily associated with language families, including Niger-Kordofanian, Afroasiatic, Nilo-Saharan, and two click-speaking hunter-gatherer groups: the Sandawe and Hadza (figs. S19 to S21). However, individuals from the Afroasiatic Cushitic Iraqw and Gorowa (Fiome) and the Nilo-Saharan Datog, who are in close geographic proximity, also cluster. Additionally, several hunter-gatherer populations were distinct, including the Okiek, Akie, and Yaaku and El Molo. Of particular interest is the common ancestry of the Akie (who have remnants of a Cushitic language) and the Eastern Cushitic El Molo and Yaaku at $K = 9$, consistent with linguistic data suggesting that these populations originated from southern Ethiopia and migrated into Kenya and Tanzania within the past ~4000 years (*4*, *29*, *32*, *39*).

Origins of hunter-gatherer populations in Africa. Our analyses demonstrate potential shared ancestry of a number of populations who practice (or until recently practiced) a traditional hunting and gathering lifestyle. For example, we observed a Hadza AAC (yellow) at $K = 5$ and $K = 3$ in the global and African STRUCTURE analyses, respectively (Fig. 3 and fig. S15), which is at moderate levels (0.18 to 0.32) in the SAK and Pygmy populations and at low levels (0.03 to 0.04) in the Sandawe and neighboring Burunge with whom the Sandawe have admixed (tables S8 and S9). The SAK and Pygmies continue to cluster at higher K values (Fig. 3 and fig. S15) and in the TESS (Fig. 5A) and phylogenetic (Fig. 1) analyses, consistent with an exclusively shared Y chromosome lineage (B2b4) (*40*). Additionally, we observed clustering of the SAK, Sandawe, and Hadza in the R_{ST} phylogenetic tree (fig. S8) and of the SAK, Sandawe, and Mbuti Pygmies at low K values in the secondary modes of Africa STRUCTURE analyses (fig. S16), consistent with observed low frequency of the Khoesan-specific mitochondrial haplotype (L0d) in the Sandawe (*18*, *19*), the presence of Khoesan-related rock art near the Sandawe homeland (*41*), and similarities between the Sandawe and SAK languages (*42*). These results suggest the possibility that the SAK, Hadza, Sandawe, and Pygmy populations are remnants of a historically more widespread proto–Khoesan-Pygmy population of hunter-gatherers. Analyses of mtDNA and Y chromosome lineages in the Khoesan-speaking populations suggest that divergence may be >35,000 years ago (*4*, *17*–*19*). The shared ancestry, identified here, of Khoesan-speaking populations with the Pygmies of central Africa suggests the possibility that Pygmies, who lost their indigenous language, may have originally spoken a Khoesan-related language, consistent with shared music styles between the SAK and Pygmies (*4*, *43*).

Shared ancestry of western and eastern Pygmies, who do not become differentiated until larger K values in STRUCTURE analyses (Fig. 3 and fig. S15), was also supported by the phylogenetic trees (Fig. 1 and figs. S7 and S8), consistent with mtDNA and autosomal studies indicating that the western and eastern Pygmies diverged >18,000 years ago (*44*–*47*). Western Pygmy populations usually clustered (Fig. 3 and fig. S15), consistent with a proposed recent common ancestry within the past ~3000 years (*48*). However, subtle substructure within the western Pygmies was apparent in the analysis of central Africa (fig. S24), probably due to recent geographic isolation and genetic drift. Asymmetric Bantu gene flow into Pygmy populations was also observed, with Bantu ancestry ranging from 0.13 in Mbuti to 0.54 in the Bedzan (table S8), consistent with prior studies (*40*, *44*, *49*, *50*).

The Hadza, with a census size of ~1000, were genetically distinct on a global level with STRUCTURE, PCA, and TESS (Figs. 2 to 5), consistent with linguistic data indicating that the Hadza language is divergent from or unrelated to other Khoesan languages (*42*, *51*, *52*). The Hadza, who have maintained a traditional hunter-gatherer lifestyle, show low levels of asymmetric gene flow from neighboring populations, whereas the Sandawe, with a census size of >30,000 (*39*), show evidence of bidirectional gene flow with neighboring populations, from whom they may have adopted mixed farming technologies (Figs. 3 to 5 and fig. S15). In fact, we observed high levels of the Sandawe AAC in northern Tanzania and low levels in northern Kenya and southern Ethiopia (Fig. 3 and fig. S15) ($K = 8$ to 13), consistent with linguistic and genetic data suggesting that Khoesan populations may once have extended from Somalia through eastern Africa and into southern Africa (*28*, *38*, *53*–*55*). Although the Hadza and Sandawe show evidence of common ancestry (Fig. 1 and figs. S7, S8, S14, S18, and S21), we observe no evidence of recent gene flow between them despite their geographic proximity, consistent with mtDNA and Y chromosome studies indicating divergence >15,000 years ago (*19*). The origins of other African hunter-gatherer populations (Dorobo, Okiek, Yaaku, Akie, El Molo, and Wata) are discussed in (*4*).

Origins of human migration within and out of Africa. The geographic origin for the expansion of modern humans was inferred, as in (*13*), from the correlation between genetic diversity and geographic position of populations (r) (figs. S30 and S31). Both the point of origin of human migration and waypoint for the out-of-Africa migration were optimized to fit a linear relationship between genetic diversity and geographic distance (*4*). This analysis indicates that modern human migration originated in southwestern Africa, at 12.5°E and 17.5°S, near the coastal border of Namibia and Angola, corresponding to the current San homeland, with the waypoint in northeast Africa at 37.5°E, 22.5°N near the midpoint of the Red Sea (figs. S2C, S30, and S31). However, the geographic distribution of genetic diversity in modern populations may not reflect the distribution of those populations in the past, although our waypoint analysis is consistent with other studies suggesting a northeast African origin of migration of modern humans out of Africa (*1*, *56*).

Correlation between genetic and linguistic diversity in Africa. Genetic clustering of populations was generally consistent with language classification, with some exceptions (Fig. 1 and fig. S32). For example, the click-speaking Hadza and Sandawe, classified as Khoesan, were separated from the SAK populations in the D^2 and $(\delta\mu)^2$ phylogenetic trees (Fig. 1 and fig. S7).

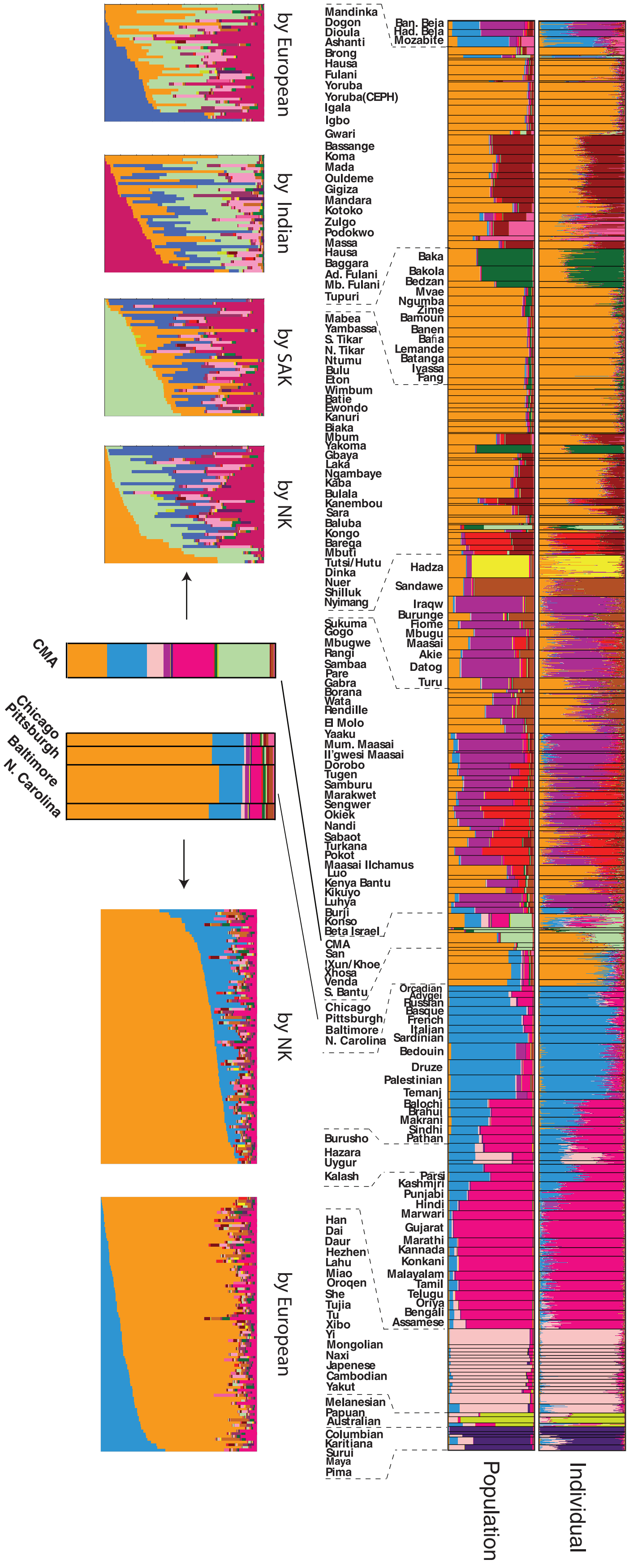


Fig. 6. Analyses of Cape Mixed Ancestry (CMA) and African American populations. Frequencies of inferred ancestral clusters are shown for $K = 14$ with the global data set for individuals (top row) and proportion of AACs in self-identified populations (bottom row). The proportions of AACs in the CMA and African American populations are highlighted in the center bottom row; proportions of AACs in individuals, sorted by Niger-Kordofanian, European, SAK, and/or Indian ancestry, are shown to the left and right, bottom row.

However, this observation is consistent with linguistic studies indicating that these Khoesan languages are highly divergent (*42*, *51*) and may reflect gene flow between the Hadza and Sandawe with neighboring populations in East Africa subsequent to divergence from the SAK. Additionally, the Afroasiatic Chadic–speaking populations from northern Cameroon cluster close to the Nilo-Saharan–speaking populations from Chad, rather than with East African Afroasiatic speakers (Fig. 1), consistent with a language replacement among the Chadic populations.

Other divergences between genetic and linguistic classifications include the Pygmies, who lost their indigenous language and adopted the neighboring Niger-Kordofanian language (*27*), and the Fulani, who speak a West African Niger-Kordofanian language but cluster near the Chadic- and Central Sudanic–speaking populations in the phylogenies (Fig. 1 and figs. S7 and S8), consistent with Y chromosome studies (*34*). Additionally, the Nilo-Saharan–speaking Luo of Kenya show predominantly Niger-Kordofanian ancestry in the STRUCTURE analyses (orange) (Figs. 3 and 4, Fig. 5, B and C, and fig. S15) and cluster together with eastern African Niger-Kordofanian–speaking populations in the phylogenetic trees (Fig. 1 and figs. S7 and S8).

Both language and geography explained a significant proportion of the genetic variance, but differences exist between and within the language families (table S5 and fig. S33, A to C) (*4*). For example, among the Niger-Kordofanian speakers, with or without the Pygmies, more of the genetic variation is explained by linguistic variation ($r^2 = 0.16$ versus 0.11, respectively; $P < 0.0001$ for both) than by geographic variation ($r^2 = 0.02$ for both; $P < 0.0001$ for both), consistent with recent long-range Bantu migration events. The reverse was true for Nilo-Saharan speakers ($r^2 = 0.06$ for linguistic distance versus 0.21 for geographic distance; $P < 0.0001$ for both), possibly due to admixture among Nilo-Saharan–, Cushitic-, and Bantu-speaking populations in eastern Africa, which might reduce the variation explained by language. The Afroasiatic family had the highest r^2 for both linguistic and geographic distances (0.20 and 0.34, respectively). However, when subfamilies were analyzed independently, the Chadic-speaking populations showed a strong association between geography and genetic variation (0.39), but not between linguistic and genetic variation (0.0012), as expected on the basis of a possible language replacement, whereas the Cushitic-speaking populations were significant for both (0.29 and 0.27, respectively) (*4*).

Genetic ancestry of African Americans and CMA populations. In contrast to prior studies of African Americans (*57*–*61*), we inferred African American ancestry with the use of genome-wide nuclear markers from a large and diverse set of African populations. African American populations from Chicago, Baltimore, Pittsburgh, and North Carolina showed substantial ancestry from the African Niger-Kordofanian AAC, most common in western Africa (means 0.69 to 0.74), and from the European–Middle Eastern AAC (means 0.11 to 0.15) (Fig. 6 and tables S6 and S8), consistent with prior genetic studies and the history of the slave trade (*4*, *57–62*). European and African ancestry levels varied considerably among individuals (Fig. 6). We also detected low levels of ancestry from the Fulani AAC (means 0.0 to 0.03, individual range 0.00 to 0.14), Cushitic AAC (means 0.02, individual range 0.00 to 0.10), Sandawe AAC (means 0.01 to 0.03, individual range 0.0 to 0.12), East Asian AAC (means 0.01 to 0.02, individual range 0.0 to 0.08), and Indian AAC (means 0.04 to 0.06, individual range 0.01 to 0.17) (table S6) (*4*). We observed very low levels of Native American ancestry, although other U.S. regions may reveal Native American ancestry (*57*).

Supervised STRUCTURE analysis (fig. S34) (*4*) was used to infer African American ancestry from global training populations, including both Bantu (Lemande) and non-Bantu (Mandinka) Niger-Kordofanian–speaking populations (fig. S34 and table S7). These results were generally consistent with the unsupervised STRUCTURE analysis (table S6) and demonstrate that most African Americans have high proportions of both Bantu (~0.45 mean) and non-Bantu (~0.22 mean) Niger-Kordofanian ancestry, concordant with diasporas originating as far west as Senegambia and as far south as Angola and South Africa (*62*). Thus, most African Americans are likely to have mixed ancestry from different regions of western Africa. This observation, together with the subtle substructure observed among Niger-Kordofanian speakers, will make it a challenge to trace the ancestry of African Americans to specific ethnic groups in Africa, unless considerably more markers are used.

The CMA population shows the highest levels of intercontinental admixture of any global population, with nearly equal high levels of SAK ancestry (mean 0.25, individual range 0.01 to 0.48), Niger-Kordofanian ancestry (mean 0.19, individual range 0.01 to 0.71), Indian ancestry (mean 0.20, individual range 0.0 to 0.69), and European ancestry (mean 0.19, individual range 0.0 to 0.86) (Fig. 6 and tables S6 and S8). The CMA population also has low levels of East Asian ancestry (mean 0.08, individual range 0.0 to 0.21) and Cushitic ancestry (mean 0.03, individual range 0.0 to 0.40). These results are consistent with the supervised STRUCTURE analyses (fig. S34 and table S7) and with the history of the CMA population (*4*, *63*).

The genetic, linguistic, and geographic landscape of Africa. The differentiation observed among African populations is likely due to ethnicity, language, and geography, as well as technological, ecological, and climatic shifts (including periods of glaciation and warming) that contributed to population size fluctuations, fragmentations, and dispersals in Africa (*1*, *4*, *34*, *64*). We observed significant associations between genetic and geographic distance in all regions of Africa, although their strengths varied. We also observed significant associations between genetic and linguistic diversity, reflecting the concomitant spread of languages, genes, and often culture [e.g., the spread of farming during the Bantu expansion (*28*)]. Of interest for future anthropological studies are the cases in which populations have maintained their culture in the face of extensive genetic introgression (e.g., Maasai and Pygmies) and populations that have maintained both cultural and genetic distinction (e.g., Hadza).

Given the extensive amount of ethnic diversity in Africa, additional sampling—particularly from underrepresented regions such as North and Central Africa—is important. Because of the extensive levels of substructure in Africa, ethnically and geographically diverse African populations need to be included in resequencing, genome-wide association, and pharmacogenetic studies to identify population- or region-specific functional variants associated with disease or drug response (*1*). The high levels of mixed ancestry from genetically divergent ancestral population clusters in African populations could also be useful for mapping by admixture disequilibrium. Future large-scale resequencing and genotyping of Africans will be informative for reconstructing human evolutionary history, for understanding human adaptations, and for identifying genetic risk factors (and potential treatments) for disease in Africa.

References and Notes

1. M. C. Campbell, S. A. Tishkoff, *Annu. Rev. Genomics Hum. Genet.* **9**, 403 (2008).
2. F. A. Reed, S. A. Tishkoff, *Curr. Opin. Genet. Dev.* **16**, 597 (2006).
3. Ethnologue (www.ethnologue.com).
4. See supporting material on *Science* Online.
5. S. A. Tishkoff, S. M. Williams, *Nat. Rev. Genet.* **3**, 611 (2002).
6. G. Sirugo *et al*., *Hum. Genet.* **123**, 557 (2008).
7. J. K. Pritchard, M. Stephens, N. A. Rosenberg, P. Donnelly, *Am. J. Hum. Genet.* **67**, 170 (2000).
8. H. M. Cann *et al*., *Science* **296**, 261 (2002).
9. N. A. Rosenberg *et al*., *Science* **298**, 2381 (2002).
10. N. A. Rosenberg *et al*., *PLoS Genet.* **1**, e70 (2005).
11. N. A. Rosenberg *et al*., *PLoS Genet.* **2**, e215 (2006).
12. M. Jakobsson *et al*., *Nature* **451**, 998 (2008).
13. S. Ramachandran *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15942 (2005).
14. P. Forster, V. Romano, *Science* **316**, 50 (2007).
15. S. Wang *et al*., *PLoS Genet.* **3**, e185 (2007).
16. J. S. Friedlaender *et al*., *PLoS Genet.* **4**, e19 (2008).
17. D. M. Behar *et al*., *Am. J. Hum. Genet.* **82**, 1130 (2008).
18. M. K. Gonder, H. M. Mortensen, F. A. Reed, A. de Sousa, S. A. Tishkoff, *Mol. Biol. Evol.* **24**, 757 (2007).
19. S. A. Tishkoff *et al*., *Mol. Biol. Evol.* **24**, 2180 (2007).
20. J. Z. Li *et al*., *Science* **319**, 1100 (2008).
21. Y. S. Chen *et al*., *Am. J. Hum. Genet.* **66**, 1362 (2000).
22. N. Patterson, A. L. Price, D. Reich, *PLoS Genet.* **2**, e190 (2006).
23. J. K. Pritchard, M. Stephens, P. Donnelly, *Genetics* **155**, 945 (2000).
24. P. Forster, S. Matsumura, *Science* **308**, 965 (2005).
25. O. Francois, S. Ancelet, G. Guillot, *Genetics* **174**, 805 (2006).
26. C. Ehret, *Int. J. Afr. Hist. Stud.* **34**, 5 (2001).
27. K. A. Klieman, *"The Pygmies Were Our Compass": Bantu and Batwa in the History of West Central Africa, Early Times to c. 1900 C.E.* (Heinemann, Portsmouth, NH, 2003).
28. C. Ehret, *An African Classical Age: Eastern and Southern Africa in World History, 1000 B.C. to A.D. 400* (Univ. Press of Virginia, Charlottesville, VA, 1998).
29. C. Ehret, in *Culture History in the Southern Sudan*, J. Mack, P. Robertshaw, Eds. (British Institute in Eastern Africa, Nairobi, 1983), pp. 19–48.
30. C. Ehret, *The Historical Reconstruction of Southern Cushitic Phonology and Vocabulary* (Reimer, Berlin, 1980).
31. C. Ehret, *Southern African Humanities* **30**, 7 (2008).
32. C. Ehret, *Ethiopians and East Africans: The Problem of Contacts* (East African Publishing House, Nairobi, 1974).
33. S. A. Tishkoff *et al*., *Nat. Genet.* **39**, 31 (2007).
34. H. Y. Hassan, P. A. Underhill, L. L. Cavalli-Sforza, M. E. Ibrahim, *Am. J. Phys. Anthropol.* **137**, 316 (2008).
35. R. E. Bereir *et al*., *Eur. J. Hum. Genet.* **15**, 1183 (2007).
36. V. Cerny, A. Salas, M. Hajek, M. Zaloudkova, R. Brdicka, *Ann. Hum. Genet.* **71**, 433 (2007).
37. C. Ehret, in *West African Linguistics: Studies in Honor of Russell G. Schuh,* P. Newman, L. Hyman, Eds. (Ohio State Univ., Columbus, OH, 2006), pp. 56–66.
38. S. H. Ambrose, in *The Archaeological and Linguistic Reconstruction of African History*, C. Ehret, M. Posnansky, Eds. (Univ. of California Press, Berkeley, CA, 1982), pp. 104–157.
39. J. L. Newman, *The Peopling of Africa* (Yale Univ. Press, New Haven, CT, 1997).
40. E. T. Wood *et al*., *Eur. J. Hum. Genet.* **13**, 867 (2005).
41. L. Lim, thesis, Brown University (1992).
42. C. Ehret, *Sprache Gesch. Afrika* **7**, 105 (1986).
43. A. Lomax *et al*., in *Folk Song Style and Culture* (National Association for the Advancement of Science, Washington, DC, 1968), pp. 16–18, 26, 91–92.
44. L. Quintana-Murci *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1596 (2008).
45. C. Batini *et al*., *Mol. Phylogenet. Evol.* **43**, 635 (2007).
46. G. Destro-Bisol *et al*., *Am. Nat.* **163**, 212 (2004).
47. E. Patin *et al*., *PLoS Genet.* **5**, e1000448 (2009).
48. P. Verdu *et al*., *Curr. Biol.* **19**, 312 (2009).
49. G. Destro-Bisol *et al*., *Mol. Biol. Evol.* **21**, 1673 (2004).
50. V. Coia *et al*., *Am. J. Hum. Biol.* **16**, 57 (2004).
51. B. Sands, in *Language, Identity and Conceptualization Among the Khoisan*, M. Schladt, Ed. (Rudiger Kupper, Köln, Germany, 1998), vol. Bd 15, pp. 266–283.
52. E. D. Elderkin, *Sprache Gesch. Afrika* **4**, 67 (1982).
53. R. Scozzari *et al*., *Am. J. Hum. Genet.* **65**, 829 (1999).
54. O. Semino, A. S. Santachiara-Benerecetti, F. Falaschi, L. L. Cavalli-Sforza, P. A. Underhill, *Am. J. Hum. Genet.* **70**, 265 (2002).
55. L. L. Cavalli-Sforza, A. Piazza, P. Menozzi, *History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ, 1994).
56. R. C. Walter *et al*., *Nature* **405**, 65 (2000).
57. E. J. Parra *et al*., *Am. J. Phys. Anthropol.* **114**, 18 (2001).
58. A. Salas *et al*., *Am. J. Phys. Anthropol.* **128**, 855 (2005).
59. J. M. Lind *et al*., *Hum. Genet.* **120**, 713 (2007).
60. M. W. Smith *et al*., *Am. J. Hum. Genet.* **74**, 1001 (2004).
61. E. J. Parra *et al*., *Am. J. Hum. Genet.* **63**, 1839 (1998).
62. Trans-Atlantic Slave Trade Database (www.slavevoyages.org/tast/index.faces).
63. G. T. Nurse, J. S. Weiner, T. Jenkins, *The Peoples of Southern Africa and Their Affinities* (Oxford Univ. Press, New York, 1985).
64. P. Mellars, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 9381 (2006).
65. J. B. Reynolds *et al*., *Genetics* **105**, 767 (1983).

66. We thank the thousands of people who donated DNA samples used in this study. We thank D. Bygott, S. J. Deo, D. Guracha, J. Hanby, D. Kariuki, P. Lufungulo, A. Mabulla, A. A. Mohamed, W. Ntandu, L. A. Nyindodo, C. Plowe, and A. Tibwitta for assisting with sample collection; K. Panchapakesan and L. Pfeiffer for assistance in sample preparation; S. Dobrin for assistance with genotyping; J. Giles, J. Bartlett, N. Kodaman, and J. Jarvis for assistance with analyses; and N. Rosenberg, J. Pritchard, A. Brooks, J. S. Friedlaender, J. Jarvis, C. Lambert, B. Payseur, N. Patterson, and J. Plotkin for

helpful suggestions and discussions. Conducted in part using the ACCRE computing facility at Vanderbilt University, Nashville, TN. Supported by L. S. B. Leakey and Wenner Gren Foundation grants, NSF grants BCS-0196183, BSC-0552486, and BCS-0827436, NIH grants R01GM076637 and 1R01GM083606-01, and David and Lucile Packard and Burroughs Wellcome Foundation Career Awards (S.A.T.); NIH grant F32HG03801 (F.A.R.); and NIH grant R01 HL65234 (S.M.W. and J.H.M.). Genotyping was supported by the NHLBI Mammalian Genotyping Service. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government. The project included in this manuscript has been funded in part with federal funds from the National Cancer Institute under contract N01-CO-12400. Original genotype data are available at http://research.marshfieldclinic.org/genetics/genotypingData_Statistics/humanDiversityPanel.asp. Data used for analyses in the current manuscript are available at www.med.upenn.edu/tishkoff/Supplemental/files.html and at http://chgr.mc.vanderbilt.edu/page/supplementary-data.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1172257/DC1
Materials and Methods
SOM Text
Figs. S1 to S35
Tables S1 to S9
References

13 February 2009; accepted 17 April 2009
Published online 30 April 2009;
10.1126/science.1172257
Include this information when citing this paper.

Dispersion of the Excitations of Fractional Quantum Hall States

Igor V. Kukushkin,[1,2] Jurgen H. Smet,[1]* Vito W. Scarola,[3,4]
Vladimir Umansky,[5] Klaus von Klitzing[1]

The rich correlation physics in two-dimensional (2D) electron systems is governed by the dispersion of its excitations. In the fractional quantum Hall regime, excitations involve fractionally charged quasi particles, which exhibit dispersion minima at large momenta referred to as rotons. These rotons are difficult to access with conventional techniques because of the lack of penetration depth or sample volume. Our method overcomes the limitations of conventional methods and traces the dispersion of excitations across momentum space for buried systems involving small material volume. We used surface acoustic waves, launched across the 2D system, to allow incident radiation to trigger these excitations at large momenta. Optics probed their resonant absorption. Our technique unveils the full dispersion of such excitations of several prominent correlated ground states of the 2D electron system, which has so far been inaccessible for experimentation.

In two-dimensional electron systems (2DESs) exposed to a strong perpendicular magnetic field B, interaction effects give rise to a remarkable set of quantum fluids. When all electrons reside in the lowest electronic Landau level, the kinetic energy is quenched and the Coulomb interaction then dominates. The strong repulsive interaction gives rise to the incompressible fractional quantum Hall fluids at rational fillings ν_p of the lowest Landau level of the form $\nu_p = p/[2p \pm 1]$, $p = 1, 2, 3, \ldots$ (*1*). The appearance of these fluids may also be understood as a result of Landau quantization of a Fermi sea, which forms at filling factor $\nu_{p\to\infty} = 1/2$ and is composed of quasi particles referred to as composite fermions (*2*–*4*). At this filling, these composite fermions experience a vanishing effective magnetic field B_{eff}. When moving away from half filling, the composite fermions are sent into circular cyclotron orbits that they execute with frequency $\omega_{c,\text{CF}} \propto |B_{\text{eff}}|$. Landau quantization of these composite fermion orbits and the successive depopulation of the associated Landau levels give rise to the incompressible fractional quantum Hall fluids. The lowest energy-neutral excitation of these fluids involves a negatively charged quasi particle with a fractional charge of $e/(2p \pm 1)$, where e is the charge on the electron (*5*–*7*), and a positively charged quasi hole that is left behind. This excitation requires an energy that, in the weakly interacting picture, corresponds to the energy gap separating adjacent composite fermion Landau levels (*4*, *8*). According to theory, these neutral excitations at fractional filling ν_p possess an energy dispersion with p minima at large wave vectors q on the order of the inverse of the magnetic length $l_B = \sqrt{e/hB}$, where h is Planck's constant, or about 10^8 m^{-1} for typical densities of gallium arsenide–based 2DESs (*1*, *9*–*14*).

The minima are referred to as magneto-roton minima and are analogous to the roton minimum in the excitation dispersion that was introduced by Landau (*15*) to account for the anomalous heat capacity observed in superfluid He-II (*16*). The magneto-roton minima govern the low-

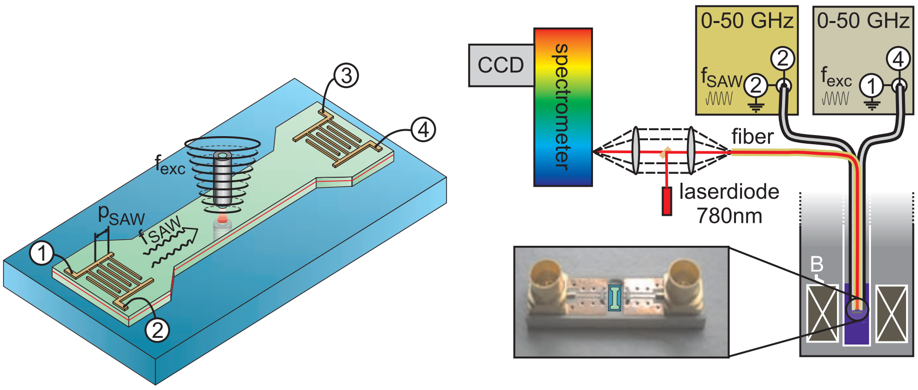


Fig. 1. Experimental arrangement for the detection of resonant microwave absorption at large wave vectors. (Left) Sample geometry consisting of a 0.1-mm-wide and 1-mm-long mesa. At its ends, the mesa widens and hosts two interdigital transducers with period p_{SAW}. High-frequency radiation drives the left transducer. The transducer launches SAWs across the sample. In the active-device region, light from a 780-nm laserdiode triggers a luminescence signal. This region of the sample is also irradiated with a quasi-monochromatic microwave by using a second high-frequency generator. Electrodes 1 and 4, which belong to transducers on opposite sides of the mesa, serve as a dipole antenna. (Right) Schematic of the cryostat configuration and the high-frequency chip carrier.

---

[1] Max Planck Institute for Solid State Research, D-70569 Stuttgart, Germany. [2] Institute of Solid State Physics, Russian Academy of Science, Chernogolovka 142432, Russia. [3] Department of Chemistry and Pitzer Center for Theoretical Chemistry, University of California at Berkeley, Berkeley, CA 94720, USA. [4] Theoretische Physik, Eidgenössische Technische Hochschule Zürich, 8093 Zürich, Switzerland. [5] Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot 76100, Israel.

*To whom correspondence should be addressed. E-mail: j.smet@fkf.mpg.de

Originally published 30 April 2009; corrected 21 May 2009

www.sciencemag.org/cgi/content/full/1172257/DC1

Supporting Online Material for

The Genetic Structure and History of Africans and African Americans

Sarah A. Tishkoff,* Floyd A. Reed, Françoise R. Friedlaender, Christopher Ehret,
Alessia Ranciaro, Alain Froment, Jibril B. Hirbo, Agnes A. Awomoyi, Jean-Marie Bodo,
Ogobara Doumbo, Muntaser Ibrahim, Abdalla T. Juma, Maritha J. Kotze, Godfrey Lema,
Jason H. Moore, Holly Mortensen, Thomas B. Nyambo, Sabah A. Omar, Kweli Powell,
Gideon S. Pretorius, Michael W. Smith, Mahamadou A. Thera, Charles Wambebe,
James L. Weber, Scott M. Williams

*To whom correspondence should be addressed. E-mail: tishkoff@mail.med.upenn.edu

Published 30 April 2009 on *Science* Express
DOI: 10.1126/science.1172257

This PDF file includes:

Materials and Methods

SOM Text

Figs. S1 to S35

Tables S1 to S9

References

Correction (21 May 2009): Two phylogenetic analyses have been added (new figs. S7 and S8), and succeeding figure numbers have been correspondingly increased. Also, text was added to the sections “Phylogenetic Analyses” (pages 3 and 4) and “Comparison of Phylogenetic Trees” (page 12).

METHODS

Sample collection and genotyping: IRB approval for this project was obtained from the University of Maryland at College Park, the University of Pennsylvania, and Vanderbilt University. Written informed consent was obtained from all participants and research/ethics approval and permits were obtained from the following institutions prior to sample collection: COSTECH and NIMR in Dar es Salaam, Tanzania; KEMRI in Nairobi, Kenya; the University of Khartoum in Sudan; Regional Hospital Sunyani, Ghana; the Nigerian Institute for Research and Pharmacological Development, Abuja, Nigeria; the Ministry of Health and National Committee of Ethics, Cameroon; the University of Bamako Faculty of Medicine, Pharmacy, and Odonto-stomatology (FMPOS) Ethics Committee, Mali; for the South African samples, ethical clearance was obtained and a material transfer agreement signed with the University of Stellenbosch for use of the samples in the current study and approval was also obtained from Dr. Roger Chennels and the Working Group of Indigenous Minorities in Southern Africa (WIMSA). Self-identified ethnicity, parent and grandparent information was recorded. Use of the African American samples was approved by the Protocol Review Office of the Institutional Review Board of the National Cancer Institute and informed consent was obtained from all individuals at the study sites prior to sample collection. Ethnic groups, sample size, language classification, and subsistence classification are given in **Table S1**. For the Kenyan, Sudanese, Nigerian, and Cameroonian samples, white cells were isolated in the field from whole blood with a salting out procedure modified from (*S1*) and DNA was extracted in the lab with a Purgene™ DNA extraction kit (Gentra Systems Inc., Minneapolis, MN). Dogon samples were obtained from blood spots donated by participants in a cohort study of malaria incidence in Bandiagara, Mali. Ghanaian DNA was extracted onsite from whole blood, with the Purgene™ DNA extraction kit. The Ju-speaking !Xun (*a.k.a.*Vasekela) and Khoe-speaking Khwe samples were collected from individuals in the area of Schmidtsdrift in the North-West Cape of South Africa. The Cape Mixed Ancestry (CMA) population, commonly referred to as Cape Coloured in South Africa, was collected in the Western Cape Province. The Dogon sample was collected in Bandiagara, Mali. Nigerian samples were collected in Abuja and Adamawa State districts. Cameroon samples were collected from the Eastern Province (Baka Pygmies and neighboring Bantu groups), Southern and Ocean Provinces (Bakola Pygmies and neighboring Bantu groups and coastal groups), Center Provinces (Medzan Pygmies and neighboring groups, mostly Bantu populations), Western Province (Bamileke and Mbororo Fulani groups), Extreme North Province (Mandara mountains and northern plains; Fulani and Afroasiatic/Nilo-Saharan speaking populations). All Cameroonians were sampled in their native village; the Hausa sample (a population who emigrated mainly from the Kano area two generations ago) was sampled in the city of Yaounde. Samples from Chad, CAR, Congo, DRC and Rwanda were obtained from individuals who recently immigrated to Cameroon. Tanzanian DNA samples were collected from individuals residing in the Arusha and Dodoma provinces of Tanzania. Samples from Kenyan populations of southern Ethiopian origin (Burji, Konso) were collected in the Rift valley, Nyanza, and Eastern provinces of Kenya. Sudanese samples were collected in the Khartoum and Kasala provinces of the Sudan. The Yemenite Temani and Ethiopian Beta Israel samples were purchased from the National Laboratory for the Genetics of Israeli Populations (*S2*). The South African !Xun/Khoe, Xhosa, Vende, Cape Mixed Ancestry (CMA), Yemenite Temani, Beta Israel, and the Malian Dogon DNA samples were amplified by Whole Genome Amplification (WGA) with Illustra GenomiPhi HY™ kits provided by GE-Healthcare (Buckinghamshire, UK). It should be noted that the DNA for the Dogon population extracted from blood spots appeared to be of lower quality and microsatellite markers did not amplify as well as other samples obtained from whole blood (43% of markers had missing data).

Data description: The full panel of Marshfield markers from screening sets #16 and #54 (microsatellites) and #101 (indels) were initially genotyped in 3,325 individuals (3,194 Africans, 109 African Americans, 22 Yemenites) sent to the Marshfield clinic for genotyping (a total of 40ug DNA/individual at a concentration of ~20ng/ul was used for genotyping). Genotyping data is available at (*S3*). Relatives were identified and removed as described below. Approximately 1% of the markers were removed because they consistently failed to amplify. **Table S1** gives the average percent of genotypes that were missing due to either failed genotyping or to not having the marker in all datasets. Of the original 3,325 individuals genotyped, we removed the following individuals from subsequent analyses: 737 individuals who were inferred to be third degree or more closely related relatives (see below), 33 individuals with low genotyping success, two single individual representatives of population samples (Fante/Ewe and Nuba), and one individual who was an outlier in the PCA analysis (see below). The remaining dataset used for subsequent analyses consisted of 2,432 individuals from 113 African populations, 21 individuals from one Yemenite population and 98 African Americans from four locations in the United States (**Table S1).** Data coverage was 95.7% (4.3% of the sample/marker combinations yielded missing data).

Data integration: To combine our dataset with preexisting data and place African genetic variability into a worldwide context, our data were integrated with data previously genotyped in the H952 subset of 1,048 samples in the CEPH-HGDP (*S4, S5*), 432 samples of Indian descent from 15 populations (*S6*), and 10 Native Australian samples of unknown ethnic population origin (provided by European Collection of Cell Cultures, Salisbury, UK). A table of marker size adjustments (to account for changes in the PCR primers used in different Marshfield screening sets) and allele size adjustments are shown in **Table S2**. The combined dataset contains a total of 1,327 genotyped markers (consisting of 848 microsatellites, 476 indels and 3 SNPs) and 3,945 DNA samples (of which 2,566 are from Africa). The overall overlap of markers genotyped in the combined dataset is 80% (not all markers were genotyped in all populations), resulting in total of ~4,000,000 genotypes.

Ascertainment Bias (AB) in the microsatellite marker set: 374 of the Marshfield markers (the GATA microsatellite repeats) were ascertained in individuals of recent European ancestry and 953 non-GATA markers were ascertained in a world-wide panel. Therefore, we tested for the possibility of ascertainment bias in the GATA markers. A significant deviation between the European and non-European samples in the ratio of GATA to non-GATA population variances was found (Mann-Whitney *U* test, U=1030, Z=2.97, p=1.49 x 10^{-3}). This deviation is in the direction of higher relative levels of variability in the Europeans with European ascertained GATA repeat markers compared to the non-GATA markers that were ascertained in a world-wide panel, consistent with an effect of population ascertainment bias in a subset of these markers (**Fig. S35**). Because of the potential effects of AB, these markers were removed from the estimates of genetic diversity on the basis of microsatellite allele variance and heterozygosity.

Detection of relative pairs: Relative pairs and duplicated samples in the dataset were inferred from the pattern of shared genotypes and population allele frequencies with RELPAIR 2.0.1 (*S6*-*8*) . Because the inclusion of closely related individuals can impact population genetic inferences (*e.g.* (*S9*)), we took the conservative approach of excluding individuals inferred to be third degree or more closely related, including inferred relative pairs between regional ethnic populations (*e.g.* all Tanzanian populations). An exception was made in the case of the Dogon as it is difficult to reliably infer relative pairs in a small sample and the Dogon are highly distinctive and could not be readily merged with other populations to improve allele frequency estimates. Merging the Dogon with other non-Pygmy West African populations inferred four unrelated individuals in the sample, but this may be overly conservative given the distinctiveness of the Dogon sample from other West Africans. Also, the Dogon are the only representatives from Mali in our study and since the sample size is already small we did not want to further reduce the sample size in the analyses, especially if the relative pair estimates were questionable. Therefore, RELPAIR inferred relative pairs among the Dogon were not excluded. In total 737 individuals were removed. Networks of relatives, which in some cases were quite complex, were plotted with neato from the GraphViz software package (*S10*), which was used to select the minimum number of individuals to exclude to break up networks of relative pairs.

Phylogenetic analyses: Pairwise distance matrices between populations were calculated using microsatellite frequencies. Three measures of distance, D^2, R_{ST}, and $(\delta\mu)^2$, were used, each sensitive to different underlying models of evolution. However, all three measures were highly correlated (r =0.92, 0.91, and 0.78 for the R_{ST}-$(\delta\mu)^2$, $(\delta\mu)^2$ - D^2. and R_{ST} - D^2 comparisons, respectively).

The distance measure D^2 is based on the Reynolds coancestry coefficient (*S11*), often referred to as θ_W, and is an estimator of F_{ST}. It is based on a model in which genetic drift is the only force influencing allele frequency differences across populations (*S11*). In cases where microsatellite alleles do not follow a stepwise mutation process, or where genetic drift and/or gene flow have stronger influence on shaping diversity than mutation, this distance measure gives more reliable phylogenetic results as compared to R_{ST} and $(\delta\mu)^2$ (*S12, S13*). D^2 genetic distances were estimated using PHYLIP (PHYLIP version 3.6 (*S14*) software package). The equation below for D^2 given in the PHYLIP documentation is derived from (*S11*).

$$D^2 = \frac{\sum_m \sum_i (p_{1mi} - p_{2mi})^2}{2\sum_m [1 - \sum_i p_{1mi} p_{2mi}]}$$

where *m* is summed over loci, *i* over alleles at the *m*-th locus, and where p_{1mi} is the frequency of the *i*-th allele at the *m*-th locus in population 1.

R_{ST} is an analogue to F_{ST} for microsatellite data, and is based on the assumptions of the stepwise mutation model (*S15*). For equal sample sizes (and equal weighting among loci) R_{ST} can be calculated as,

$$R_{ST} = \frac{\overline{S} - S_W}{\overline{S}},$$

where $\overline{S}$ is twice the repeat unit variance across populations and S_W is twice the repeat unit variance within populations (*S15*). Weighting was used so that each population sample contributed equally to $\overline{S}$ despite unequal sample sizes (*S15*).

$(\delta\mu)^2$ (*S16*) was developed specifically for microsatellite markers, and, like R_{ST}, assumes a stepwise mutation model (*S16*). This measure works best for phylogenetic reconstruction when taxa have been separated for long time periods and is expected to be linear with time (*S16*).

$$(\delta\mu)^2 = (m_x - m_y)^2$$

where m_x and m_y are the mean allele sizes in populations x and y, respectively. Data for multiple loci are combined by averaging the single-locus values of $(\delta\mu)^2$.

Allele frequency datasets were re-sampled 1000 times to generate multiple distance matrices. Unrooted neighbor-joining trees were constructed from these matrices using *neighbor* (PHYLIP version 3.6 software package) (*S14*). For the D^2 analysis, a consensus tree, obtained using *consense* (PHYLIP), was used as an input for *contml* (PHYLIP) to generate branch lengths from allele frequency data using a maximum likelihood algorithm. Nodes that were supported by bootstrap values of at least 70% were labeled. For the R_{ST} and $(\delta\mu)^2$ trees, a single tree from *neighbor* was generated with branch lengths, and the bootstrap values, calculated using *consense*, were placed on that tree. The nodes supported by bootstrap values of at least 50% were labeled.

The topology of inferred trees depends upon the genetic distance measures used that are based on different evolutionary models. These models differ in regard to sensitivity to evolutionary forces such as mutation, genetic drift, and migration. For example, when gene flow is reduced, the effect of mutation may become an important factor in population differentiation, whereas when levels of gene flow are high, F_{ST} models may outperform distance measures based on stepwise mutation models (*S13*). Both the R_{ST} and $(\delta\mu)^2$ measures assume that mutation is an important factor in population differentiation and, therefore, may be more informative for detecting relationships between older, geographically isolated populations. However, there may be deviations from a stepwise mutation model for some microsatellite loci. Additionally, even when the strictest conditions of the stepwise mutation model are met, R_{ST} and $(\delta\mu)^2$ have higher variances than F_{ST} measures (although the variance is less for large numbers of loci), resulting in low bootstrap values for phylogenetic analyses of closely related taxa (*S12, S16, S17*).

Trees were plotted with TreeViewX (*S18*). All trees were built as unrooted but the San population is displayed as an outgroup in the D^2 and R_{ST} trees for ease of visualization; prior studies have indicated that they branch from the root of the human phylogenetic tree (*S19-22*).

Estimates of genetic diversity from the average microsatellite repeat unit length variance: Theta was estimated as twice the variance in repeat length units, $\theta = 2\sigma^2$, (under a stepwise mutation model) (*S23*). Prior to analyses GATA repeats were excluded due to possible effects of ascertainment bias, discussed above.

Heterozygosity: The expected heterozygosity was computed from the microsatellites with the GDA software (*S24*), with the sample-size corrected estimator, as in (*S25*). From Ohta and Kimura (*S26*), under a stepwise model, the expected relationship between θ and heterozygosity (H) is

$$H = 1 - \frac{1}{\sqrt{1+2\theta}},$$

which rearranges to

$$\theta = \frac{1}{2}\left(\frac{1}{(H-1)^2} - 1\right).$$

For autosomal loci in an ideal population, θ is defined as $\theta = 4N_e\mu$, where N_e is defined as the effective population size and μ is the per generation mutation rate. Prior to analyses GATA repeats were excluded due to possible effects of ascertainment bias, discussed above.

Estimates of shared and private alleles: The ADZE software (*S27*) was used to estimate the number of private and shared alleles within and between populations and geographic regions. The ADZE software implements a rarefaction approach for counting alleles private to combinations of populations by evaluating the number of alleles found in each of a set of populations but absent in all remaining populations, considering equal-sized subsamples from each population (*S27*).

Analysis of Molecular Variance (AMOVA) : Variance components were estimated for the 848 microsatellites, as in equation 5.3 from (*S28*), with GDA (*S24*), with 95% confidence intervals created on the basis of 1000 bootstraps across loci.

Mapping population geographic coordinates: Approximate coordinates of the traditional range for each ethnic group were determined from Ethnologue's language maps (*S29*) or as reported previously for the CEPH diversity panel and Indian datasets (*S6, S30, S31*). The average latitude and longitude was used when a range of coordinates was given for a single population or for samples composed of a few individuals from different neighboring populations. For data from previously published population samples, the reported geographic coordinates were used (*S30, S31*). A complete listing of coordinates can be found in **Table S1**.

Serial founder effect analysis: The geographic origin of modern humans and the route of migration out of Africa south of the Sahara to northern Africa and Eurasia were estimated from the data under a serial founder effect model (*S31*). This inference was created on the basis of the geographic position of sampled populations and estimated levels of genetic diversity for each population (θ estimated from the average microsatellite repeat unit length variance). Under this model, the highest levels of modern genetic variation are expected to occur in the geographic region of origin of modern humans. As modern human populations migrated away from this point, founder effects steadily reduced levels of genetic diversity with geographic distance (*S31*). Thus, the point of origin is predicted to have the most negative correlation (r) between genetic diversity and distance from the origin, *i.e.* diversity decreases with distance from this point, yielding a negative slope. A grid of all points of origin was evaluated and the resulting correlation between diversity and great circle distance (*i.e.* along the curvature of the earth) from these points was plotted. Furthermore, rather than fixing a single waypoint of migration for populations in Northern Africa and Eurasia [*cf.* (*S31*)], this point was also varied over a grid, simultaneously with the point of origin, and the best fit (according to a least squares linear regression) was found.

The grid of potential points of origin, excluding Oceania and the Americas, was evaluated, and the resulting correlation (r) between diversity and great circle distance (*i.e.* along the curvature of the earth) from these points was plotted (**Fig. S30**). The portion of the globe encompassing Africa and Eurasia (-40 to 70 degrees latitude, and -22.5 to 187.5 degrees longitude) was divided into a grid with units of 2.5 degrees, creating a 45x85 grid with 3,825 total points. However, because Oceania and the oceans were excluded as possible points of origin or waypoints out of Africa, only 1,495 grid points were plotted. 10,000 bootstrap replicates were created where populations were randomly re-sampled, with replacement, from the full set of sampled populations (thus, some populations were included more than once and some not at all in individual bootstrap replicates) and the procedure was repeated for each bootstrap replicate. This procedure was used to determine the geographic confidence region for both the origin and the waypoint (**Fig. S31**). The points with the best fit in each replicate were chosen as the origin and waypoint.

Also, within this serial founder effect framework, the predicted regional levels of expected genetic variation were estimated (**Fig. S2C**). To do this, we used a linear regression weighted by geographic distance away from each point of evaluation (*i.e.* nearby levels of variation contribute more to the correlation than points further away) and solved for the intercept (*i.e.* the predicted level of variation at that point, where distance is equal to zero, on the basis of the relationship between genetic variation and distance away from each point).

TESS analysis: The genetic clustering of individuals in the presence of a spatial geographic network was inferred with TESS (*S32*). Because only a single geographic point is used for each population, the individuals within a population were randomly assigned geographic positions over a narrow range, ±0.1 degree N-S and E-W, centered on the population value (this avoids identical geographic placements that can obscure heterogeneous clustering results within population samples). The no-admixture model was used and the interaction parameter was set to one, ψ=1 (lower values for this parameter result in the identification of additional clusters). The maximum number of clusters was set to K=10 (of which the program estimated the presence of K=6 clusters). Five primary run sets of five runs each (25 total primary runs) were run for 2,000 steps, then each of the five sets were extended in a single secondary run (five total secondary runs) of 2,000 steps starting from the highest likelihood endpoint of the five primary runs in each set. Finally, the highest likelihood result at the end of these five secondary runs was chosen to display the resulting geographic clustering pattern.

Principal Components Analysis: The EIGENSOFT software package (*S9, S33*) was used for individual principal components analyses. *Smartpca* (of the EIGENSOFT package) was recompiled and the makefile edited to substitute a gfortran library call for the f2c library call. The microsatellite data was converted into a false SNP format by scoring the presence or absence of each of *n*-1 alleles (where *n* is the number of alleles in the sample). The in/del data were converted to a binary coding for the presence or absence of a reference allele.

Unsupervised STRUCTURE analysis: Population structure was inferred with a Bayesian clustering approach implemented in the STRUCTURE software package (*S34, S35*). This program identifies groups of individuals with similar allele frequency profiles (*S35*). This clustering approach avoids *a priori* population classifications, instead estimating the shared population ancestry of individuals based solely on their genotypes under an assumption of Hardy-Weinberg equilibrium and linkage equilibrium in ancestral populations. It infers individual proportions of ancestry from *K* clusters, where *K* is specified in advance and corresponds to the number of posited ancestral populations and can be varied across independent runs. Individuals can be assigned admixture estimates from multiple ancestral populations, with the admixture estimates summing to 1 across these population clusters (*S36*).

STRUCTURE Version 2.2.3 was used for unsupervised STRUCTURE runs assuming the *F* model of correlated allele frequencies among the ancestral clusters (*S37*) with a 20,000 step burn-in and 10,000 step chain, with a separate α estimated for each population (POPALPHAS = 1). The latter allows for asymmetric patterns of admixture amongst the inferred populations.

We analyzed the global dataset, the African dataset, and African regional data subsets separately with an unsupervised STRUCTURE analysis. 25 replicates were run for each *K* (number of ancestral clusters assumed) and each dataset. We ran up to $K = 15$ for the global, African, and East African datasets, $K = 10$ for the Middle African and Western African datasets, and $K = 5$ for the Saharan African and South African datasets. The structure outputs were processed with CLUMPP (*S38*) and a G-statistic greater than 90% was used to assign groups of runs to a common clustering pattern. The maximum *K* value was determined on the basis of: (1) the *K* value at which the likelihood distribution reached a maximum and began to plateau or decrease; (2) high stability of clustering patterns between runs (the primary mode was observed in at least 60% of the 25 runs) and; (3) from the Kmax value at which Kmax + 1 no longer refines the clusters (i.e. Kmax + 1 no longer splits the cluster distinguished at Kmax). The run with the highest likelihood of the data given the parameter values for the predominant clustering pattern (i.e. the mode) at each *K* was used for plotting with DISTRUCT (*S4, S39*).

Supervised STRUCTURE analysis: With STRUCTURE Version 2.2.3 (*S34*), individuals from populations having a high frequency of distinct ancestral clusters inferred with the unsupervised structure analyses (Papua New Guineans, Pima Native

Americans, Han Chinese, French, Indian, Iraqw, Hadza, Baka Pygmies, San, Dinka, Fulani, Mandinka, Mada, Lemande) were used as 14 training populations and a supervised analysis was performed in order to determine membership coefficients in the African Americans and Cape Mixed Ancestry individuals. For this analysis, we included both Bantu (Lemande) and non-Bantu (Mandinka) Niger-Kordofanian populations with highest frequencies of distinct ancestral clusters on the basis of the unsupervised STRUCTURE analysis of western Africa (**Fig. S26**). Ten replicates were done which gave similar results and the mean of these replicates was used for plotting (**Fig. S34)**.

NJ tree inferred from inferred ancestral clusters in STRUCTURE: Unrooted NJ trees were constructed with *Phylip* software version 3.66 (*S14*) from the average pairwise nucleotide distance between ancestral clusters (*S37*).

Geographic vs genetic distance analyses: Analyses were performed to assess the relationship between geographic distance and genetic differences at multiple scales: within major global regions and within Africa. Great circle geographic distances were calculated with the Haversine method, as described (*S31*). Prior to analyses, $(\delta\mu)^2$ genetic distances were tested for normality, with the Shapiro Wilkes test (*S40*). All populations deviated from normality ($p \leq 0.01$) except the Middle Eastern ($p=0.26$) and European (0.06) populations. Because of the highly significant deviation from normality in all other populations, correlation analysis was done, with the non-parametric Spearman's Rho test, which was calculated at the following levels: all Africa, Middle East, Central Asia, East Asia, India, and Europe. Within Africa regions were divided into Eastern Africa, Southern Africa, Central Africa, Saharan Africa, and Western Africa. To test for heterogeneity of Spearman's Rho, Fisher *r*-to-*z* transformations of the Spearman correlation coefficients was performed (*S41*). All analyses were performed with SAS software (*S42*).

Genetic/linguistic diversity correlations: In a historical linguistic classification, languages are attributed to particular clades and subclades in a language family on the basis of their common possession of particular linguistic 'mutations,' i.e., unique innovations in lexicon, lexical meanings, phonology, and grammar. Such shared innovations are analogous to genetic mutations, in that they attest to the descent of the languages in question from earlier single ancestral languages, in which the innovations ('mutations') took place. In every case this kind of evidence forms the primary basis for constructing the family trees of the African language families (**Fig S32**). Where the evidence is extensive, complex and detailed trees of successive language divergences can be constructed, as the case of the Nilo-Saharan family illustrates (**Fig. S32H**) (*S43*). For Afroasiatic, an outline presentation of the innovations on which its primary branchings are based can be found in (*S44*). A variety of other sources present the testimony of linguistic innovation for the internal subgrouping of the Chadic branch (*S45*), the Omotic branch (*S46*), and the Cushitic branch (*S47, S48*), along with its Eastern Cushitic (*S49*) and Southern Cushitic (*S50*) subclades. For the classification of Niger-Kordofanian into its major clades and subclades, the work of Williamson and Blench (*S51*) has been followed. A detailed Bantu subclassification based on phonological, pronominal, and lexical innovations, combining the findings of a number of scholars, can be accessed at (*S52*).

The language distance measures in this study rest on a different kind of evidence, lexicostatistics. Its results turn out to be generally congruent with the evidence of lexical, phonological, and grammatical innovation. The data come from a large number of sources—for Bantu (*S53*), non-Bantu Niger-Kordofanian subclades (*S54-59*), Nilo-Saharan (*S46, S60-66*), Afroasiatic (*S46, S50, S67, S68*), and Khoesan (*S69, S70*) (see also (*S52*)). The values indicated at each node in the language relationship trees are calculated medians of the distributions of attested pairwise percentages of cognation for languages whose ancestral forms diverged from each other at the particular node.

Divergence times between related languages were estimated with archeological dates and glottochronological methods (*S71, S72*). However, these age estimates were not used for the correlations with genetic distance. Rather, levels of shared cognates between population pairs were used to infer linguistic similarities (scaled from 0-1; **Fig. S32**). The relationship between linguistic distance (1- language similarity) and genetic distances $(\delta\mu)^2$ was then analyzed within language families. We restricted our analyses to within language families because of the more rapid decay of linguistic as compared to genetic similarities. The sampled Pygmy populations, which speak Niger-Kordofanian languages today, are a well known case of recent language replacement. This, taken together with the Pygmies genetic dissimilarity to other Niger-Kordofanian populations, led us to analyze the pairs within the Niger-Kordofanian family with and without the Pygmies. Differences between results can be used to infer the relationship between language and genetic distance in Pygmy populations relative to other NK populations.

Regression analyses were performed with $(\delta\mu)^2$ as the dependent variable and language distance (1- the similarity index) and geographic distance as independent variables, with SAS (Version 9.1). Since language distances were not normally distributed (Shapiro Wilks test, $p < 0.001$) the language variables were transformed with an inverse transformation (1/language distance). Regression analyses were performed on the transformed variable and then adjusted for geographic distance. In addition, an interaction term (language similarity X geographic distance) was added to the model and tested for significance. A regression of $(\delta\mu)^2$ on geographic distance was also analyzed and adjusted for language. The reduced models were: genetic distance $((\delta\mu)^2) = \mu + \alpha$ (language similarity) + ε and genetic distance = $\mu + \beta$(geographic distance) + ε. The full model was: genetic distance = $\mu + \beta$(geographic distance) + α (language similarity) + ε. The full model was analyzed with and without an interaction term. Only within language family comparisons were performed. These included the AA, NS, and NK with and without Pygmies. Adjustment for the second dependent variable is only possible when the two variables are not co-linear. This was examined, and was not an issue for the analyses performed with the exception of the Cushitic speakers (see below).

Statistical significance: Unless otherwise noted, we considered a statistic significant if the p-value for the test was less than or equal to 0.05.

SUPPLEMENTARY TEXT

African language classification

There are over 2,000 distinct ethno-linguistic groups in Africa, speaking nearly a third of the world's languages (*S29*). Except for a few unallocated languages, the languages of Africa have been classified into four macro-families: Niger-Kordofanian, Afroasiatic, Nilo-Saharan, and Khoesan. Here we use the spelling of Afroasiatic as originally defined by Joseph Greenberg in the 1950s (*S73*) rather than the derivative "Afro-Asiatic" spelling. Of the four African language families, Niger-Kordofanian is the most wide-spread. Its languages are spoken across half of Africa. Some of the West African Niger-Kordofanians sampled in our study belonged to urbanized societies, with centralized political structures, social stratification, and elaborate ritual institutions, and were historically involved in complex systems of long-distance trade (e.g. the Yoruba and other southern Nigerians, as well as the Asante and Brong, for the past 1,500 years, and the Malinke (Mende) and their ancestors since 3,500-3,000 ya). Many of the Niger-Kordofanians of Central and southeastern Africa sampled in this study belonged to chiefdom-size polities since 1,500-1,000 ya, but others, most notably the Kongo, formed powerful kingdoms with widespread trade connections across the Congo basin (*S74*).

The Afroasiatic family includes a great variety of societies. Many Chadic peoples of West Africa lived in chiefdom-sized agricultural communities, but the Hausa in particular for the past 1,000 years have formed an urbanized society with cities of up to 50,000 inhabitants, engaged in manufacture and widespread commercial activities. The Amazigh-speaking Mzab of the Sahara formed a commercial enclave with similarly widespread commercial relations. Many of the Cushitic populations, in contrast, were organized into village-scale farming or agro-pastoral communities; others such as the Beja pastoralists and Oromo mixed farmers formed alliances of up to several thousand people, on the basis of either clan confederacies or on age-grade institutions of governance (*S74*). This language family also includes the ancient Egyptians and one group, the Semites, who reside in northern Africa and the Middle East (*S75*).

The Nilo-Saharan family extends across another wide expanse of Africa, from the Songay of Mali in the west to the Nilotes of southern Sudan and Eastern Africa. The Central Sudanic groups of the Chad Basin formed both chiefdoms and kingdoms in recent centuries, while the Kanuri and Kanembu over the past 1,000 years belonged to powerful kingdoms with wide-reaching commercial connections. States as early as 3,500 BC along the Nubian Nile also spoke Nilo-Saharan languages (*S67, S74*). In contrast, the Nilotic peoples often belonged to large alliances, clan-based as among the Dinka and Nuer of southern Sudan or from age-grade institutions as among the Maasai of East Africa (*S74*).

The fourth family, Khoesan, is highly unusual in being composed today primarily of residual hunter-gatherer populations (e.g. San, !Xun/Khoe, Hadza, Sandawe), with only the Sandawe being recently established farmers and herders, albeit with strong ideological attachments to hunting remaining (*S76*). However, it should be noted that the classification of Hadza and Sandawe as Khoesan is a contentious issue (*S77*) since they are highly divergent from each other and from SAK languages. Linguistic data indicates that the Hadza language, in particular, is especially divergent from, or perhaps unrelated to, other Khoesan languages (*S78, S79*).

African classification by subsistence mode

African populations were classified based on subsistence modes historically practiced by the ethnic populations listed in **Table S1**. These include populations that practice a diverse and complex array of subsistence modes, including animal domestication (herding), plant cultivation (farming), plant cultivation and herding (mixed farming) and hunting and gathering. Although the majority of the newly collected African samples included in this study originated from individuals living in rural populations that continue to practice farming- or herding-based economies, some of the populations (e.g. Mende, Xhosa, Yoruba, among others) include individuals living in cities who no longer engage in farming or herding practices. African non-forager populations include the numerous stratified, commercially active, and politically centralized societies, as well as the up-till-recently large clan- or age-grade-based societies and the smaller village-scale societies.

Comparison of genetic diversity (θ) inferred from microsatellite variance versus heterozygosity

Estimates of genetic diversity ($\theta=4N_e\mu$) from the sample variance in repeat length of the microsatellite alleles compared to (θ) on the basis of heterozygosity are shown in **Fig. S2**. The correlation between variance and heterozygosity estimates ($R^2 = 0.927$, $p = 1.19\text{x}10^{-104}$) is shown in **Fig. S4**. Three African hunter-gatherer populations are included in the five populations with highest measured levels of genetic diversity on the basis of variance of microsatellite allele length; the Baka and Bakola Pygmies from Cameroon and the San from Namibia (the other two populations with highest diversity are the Ntumu from Cameroon and the Burunge from Tanzania) (**Fig. S2A**). In contrast, the five populations with highest diversity from heterozygosity estimates are the Burunge, Turu, Gogo, and Sukuma from Tanzania (all agriculturalist or agro-pastoralist populations), and the African Americans from Baltimore (**Fig. S2B**). The San and the Hadza hunter-gatherers are amongst the African populations with lowest levels of heterozygosity.

Prior studies from simulations as well as empirical data have indicated that the ratio of θ inferred from variance relative to θ inferred from heterozygosity can be informative for inferring past population expansion and bottleneck events (*S80, S81*). Specifically, a population bottleneck followed by population expansion causes an imbalance between estimates of θ with allele size variance and those with heterozygosity (*S80*), because the variance estimate is transiently higher than expected under equilibrium conditions, resulting in the variance θ to heterozygosity θ ratio being greater than one. By contrast, populations which have recently expanded in the absence of a strong bottleneck event exhibit ratios of variance to heterozygosity less than one. This is because after expansion, both variance and heterozygosity increase as the population approaches a new mutation-drift equilibrium. If the expansion is sufficiently recent with mutation events largely restricted to the post expansion period and distributed among many lineages, the initial effect is to increase heterozygosity from new mutations more than allele size variance, since the latter quantity is more sensitive to mutations that distinguish the oldest lineages and requires more mutational events to recover (*S81*). As a result, heterozygosity approaches its new limit value faster than genetic variance (*S81*).

The ratios of θ inferred from variance and heterozygosity for the current dataset are shown in **Fig. S3**. The ratio of variance and heterozygosity is the largest in Native

American populations, followed by Oceanic and East Asian populations, all with values greater than one, intermediate in most European, Middle Eastern, and Indian populations, with values near one, and with values less than one in most African populations and a few Middle Eastern and European populations. This observation is consistent with previous findings, suggesting a strong bottleneck followed by a recent and rapid expansion in Native Americans and Australo-Melanesians, and expansion but lack of a recent strong bottleneck in Africans (*S80, S82*). Interestingly, the San and the Hadza hunter-gatherers have the highest ratio of variance relative to heterozygosity among almost all African populations, with a ratio value slightly greater than 1.0 (**Fig. S3**). The only African population with a larger ratio is the Dogon. The Hadza and San are also apparent outliers in the plot of θ inferred from heterozygosity shown in **Fig. S2B**. These results are consistent with relatively stable small population sizes in these hunter-gatherer populations, although simulations will be required to obtain detailed demographic parameter values.

Comparison of Phylogenetic Trees

Three measures of genetic distance were used, each based on and sensitive to different underlying models of evolution (see methods). The D^2 genetic distance assumes that population differentiation is due to genetic drift (*S11*). Therefore, this measure is most sensitive to recent differentiation events, such as among the Niger-Kordofanian speaking populations (**Fig. 1**). R_{ST} is similar to F_{ST}, but is based on the fraction of the total variance in allele size between subpopulations, assuming a stepwise mutation model (*S15*). $(\delta\mu)^2$ is based on differences in the means of microsatellite allele sizes, also assuming a stepwise mutation model (*S16*). Therefore, both R_{ST} and $(\delta\mu)^2$ may be more sensitive to older differentiation events, where mutation is expected to play a large role. It is interesting to compare the results of the phylogenetic analyses using these three genetic distance measures. As expected, the tree constructed from D^2 (**Fig. 1**) shows clustering of closely related Niger-Kordofanian speaking populations (e.g. the two Yoruba populations from Nigeria and the two Tikar populations from Cameroon). In this tree, the Pygmies cluster near the SAK, with the Mbuti appearing closest to the SAK, consistent with the STRUCTURE, TESS, and PCA results (**Figs. 2, 3, 4, 5**). In the D^2 neighbor-joining tree (**Fig. 1**), the Hadza and Sandawe populations cluster near the neighboring East African populations with whom they have admixed. By contrast, in the R_{ST} tree (**Fig. S7**), the Hadza and Sandawe populations cluster near the SAK, close to other East African populations. In both the R_{ST} and $(\delta\mu)^2$ trees (**Figs. S7 and S8**), the Pygmies cluster closest to the Niger-Kordofanian Bantu-speaking populations with whom they have admixed. The clustering of the SAK with the Hadza/Sandawe in the R_{ST} tree, and the clustering of the SAK with the Pygmies in the D^2 tree may indicate that these population differentiation events are quite old, and that patterns of population relationships have been influenced by subsequent demographic events, including admixture with local populations. Additionally, these patterns may reflect ancient admixture events between the ancestors of the SAK (whose distribution may have extended as far north as Ethiopia (*S83-87*)) and the ancestors of the Hadza, Sandawe, and Pygmies, whose geographic ranges may have overlapped. The genetic similarity between the SAK and Pygmies, as indicated in the TESS and PCA analyses (**Figs. 2, 5A**), and at high K values in the STRUCTURE analyses (**Figs. 3, 4, 5B, 5C, S15**), raises the possibility that the SAK and Pygmies (particularly the Mbuti) may share more recent common ancestry and/or gene flow.

Details of global STRUCTURE results

When two clusters are assumed in the STRUCTURE analysis (K = 2) (**Fig. 3**), individuals can primarily be assigned to African (orange) or non-African (blue) clusters, consistent with the PCA (**Fig. 2A)**. Individuals from Saharan and Eastern Africa show heterogeneous ancestry, reflecting descent from populations ancestral to non-Africans and/or gene flow from non-Africans into Africa. We also find evidence for low levels of African ancestry in several Middle Eastern and Oceanic populations. The latter observation is consistent with possible gene flow into these regions and with studies based on archeological and genetic data, indicating an early migration event of modern humans out of Africa, across southern Asia, and into Oceania (*S36, S88*). With three clusters (K = 3), the East Asian, Oceanic and Native American individuals become distinguished (pink) (with individuals from South and Central Asia showing heterogeneous ancestry), again consistent with the PCA. With a fourth cluster (K = 4), many Eastern African populations (purple), particularly the Hadza and Afroasiatic speakers, are distinguishable from other African populations. At K = 4, the Middle Eastern and Oceanic populations both show evidence for low levels of Eastern African ancestry. The fifth cluster (K = 5) distinguishes the Hadza hunter-gatherer population (yellow), consistent with PCA (**Fig. 2**), and to a lesser extent the Pygmy and SAK hunter-gatherer and the Sandawe (former hunter-gatherer) populations. The sixth cluster (K = 6), distinguishes the western Pygmies (dark green). The seventh cluster (K = 7) distinguishes African individuals who speak Chadic (a western Afroasiatic sub-family) and/or Nilo-Saharan languages (red). The eighth cluster (K =8) distinguishes the Indian individuals (dark pink) and the Oceanic populations, consistent with possible shared ancestry of these populations. The ninth cluster (K = 9) distinguishes the Oceanic individuals (light green). The tenth cluster (K = 10) distinguishes the Native American individuals (dark purple). The eleventh cluster (K = 11) distinguishes the Mbuti Pygmy and SAK individuals (light green), indicating shared common ancestry of these geographically distant hunter-gatherer populations. At K = 12, the Chadic and Nilo-Saharan speaking populations originating from northern Cameroon, Chad, and southern Sudan, become distinct (maroon). The thirteenth cluster (K = 13) distinguishes the Sandawe from Tanzania (brown) (former hunter-gatherers who adopted mixed farming an uncertain number of centuries ago) and the fourteenth cluster (K =14) distinguishes the nomadic pastoral Fulani populations (fuchsia). Several of the African populations (the CMA, Fulani, and eastern Afroasiatic-speakers) show low to moderate levels of the European/Middle Eastern cluster, consistent with possible gene flow from those regions. At K = 10-14, the non-African pattern essentially recaptures the results of prior studies of the CEPH-HGDP (*S4-6, S20, S89*) in which individuals cluster by major geographic regions. However, our data reveal considerably more substructure among Africans (nine at K = 14) than previously observed (*S4-6, S20, S89*).

Allele frequencies from inferred ancestral clusters derived from the STRUCTURE analysis at K = 14 were used to construct an un-rooted neighbor-joining tree (**Fig. S14**). African and non-African Associated Ancestral Clusters (AACs; labeled based on the populations showing the highest levels of ancestry for each inferred ancestral cluster) are highly divergent. The Oceanic AAC is the branch closest to the African AACs, followed by a clade formed by the European and Indian AACs, and finally a clade formed by the Asian and Native American AACs. Within Africa, the Pygmy and SAK AACs form a clade, as do the Hadza and Sandawe AACs, and the Nilo-Saharan and Chadic AACs. The Fulani and Cushitic (an eastern Afroasiatic subfamily) AACs, which likely reflect Saharan African and East African ancestry, respectively, are closest to the non-African AACs, consistent with an East African migration of modern humans out of Africa or a back-migration of non-Africans into Saharan and Eastern Africa.

Genetic variation within Africa

The proportion of variation among African populations classified based on geographic, linguistic, and subsistence classification was determined based on AMOVA analysis of the microsatellite data. It should be noted that sampling design influences inferences of variance among populations. Indeed, the proportion of variation among African populations inferred from AMOVA analysis of the microsatellite data is 2.59% for the CEPH-HGDP African populations but is 1.71% in our expanded set of Africans (**Table S3**), likely due to an over-representation of relatively isolated hunter-gatherer populations in the CEPH-HGDP. Within Africa, southern Africa shows the highest level of among population variation (2.13%), presumably reflecting the highly divergent SAK populations. Hunter-gatherer populations in general had the highest levels of among population variance (3.18%). Populations that speak Khoesan languages, most of whom are (or until recently were) hunter-gatherers were relatively variable (3.39%). In contrast, African herding and farming populations had the lowest levels of among population variance (0.94% and 0.97%, respectively), possibly due to population expansions, assimilations, and long range migrations over the past ~5,000 years (*S90, S91*) . Low levels of population variance were also observed among the Nilo-Saharan (1.13%) and Niger-Kordofanian (1.17%) language families, many of whom originated from herding and farming societies, respectively.

Details of Africa STRUCTURE results

The Africa-wide STRUCTURE result (**Fig. S15**) largely recapitulates the African PCA results. Specifically, the western and eastern African populations were distinguishable at K=2, several hunter-gatherer (or former hunter-gatherer) populations at K=3 (the Hadza, SAK, and Pygmies, and to a lesser extent the Sandawe and neighboring Burunge with whom the Sandawe have admixed, shown in yellow), and the Hadza (yellow) are distinguished from the SAK and Pygmies (dark green) at K=4. As K increases, the following population clusters are sequentially distinguished: the Nilo-Saharan and Chadic speaking individuals (shown in red) from K=5 upward; individuals with some European or Middle Eastern ancestry, as inferred from the global STRUCTURE analysis, from K=6 upward (shown in blue, consisting of eastern African Afroasiatic and Nilo-Saharan speaking populations, the Fulani, and CMA population); the SAK and Mbuti Pygmies (shown in light green) are distinguished from the western Pygmies (dark green) at K=7 and higher; the Sandawe (shown in brown) from K=8 and higher; the Fulani (pink) at K=9 and higher; the Mbugu (who speak a Southern Cushitic language with extreme Bantu admixture (*S50, S92*), shown in dark purple) at K=10 and higher; the Cushitic speaking populations of southern Ethiopian origin (Borana, Burji, Konso) and northern Kenya (Wata, Rendille and Gabra) at K = 11 and 13 (light purple); at K = 14, Nilotic Nilo-Saharan speaking populations (i.e. Maasai, Dorobo, Sengwer, Saboat, Tugen, Samburu, Marakwet, Sengwer, Okiek, Nandi, Saboat, Turkana, Pokot; red) are distinguished from the Central Sudanic Nilo-Saharans (Laka, Ngambaye, Kaba, Bulala, Kenembou, Sara; tan), and Chadic-speaking populations (Mada, Ouldeme, Giziga, Mandara, Kotoko, Zulgo, Podokwo, Masa, Hausa) and Semitic-speaking Baggara (maroon). The Bantu speakers of South Africa (Xhosa, Venda), had high proportions of the SAK and western African Bantu AACs, and low levels of the East African Bantu AAC (the latter is also present in Bantu speakers from Democratic Republic of Congo and Rwanda). These results demonstrate, with nuclear genetic markers, evidence for two sources of migration (from the East and West) of Bantu speakers into southern Africa. These results are consistent with linguistic and archeological evidence, suggesting a distinct East African Bantu migration event into southern Africa from a core area west of Lake Victoria ~2,000 ya (*S83*) and the incorporation of Khoekhoe ancestry into several of the Southeast-Bantu populations ~1,500 – 1,000 ya. The incorporation of a major Khoekhoe demic component in the proto-Sotho and proto-Nguni societies of ~1,500-1,000 ya in South Africa is demonstrated in both proto-languages by their separate borrowings of Khoekhoe loanword sets of the heavy intensive category (*S93*), a diagnostic marker of this kind of population history (*S94*).

Origins of Nilo-Saharan and Afroasiatic Cushitic speaking populations

The southern/central Sudanese show high levels of both the Nilo-Saharan (red) and Afroasiatic Cushitic (purple) AACs from K = 8-13 (**Fig S15**), consistent with linguistic arguments suggesting a long history of extensive contact and gene flow ~20,000 – 10,500 ya, along the western edges of the Ethiopian highlands (*S95*). The history of regional interactions between Nilo-Saharans and Cushites 5,000-1,000 ya in southwestern Sudan and adjacent parts of Uganda and Kenya (*S62*) were likely to have reinforced the genetic patterns observed in the STRUCTURE analyses.

Our data support the hypothesis based on linguistic, archeological, mtDNA, and Y chromosome data, that the Sahel has been a corridor for bi-directional migration between eastern and western Africa (*S96-98*). We observe the highest proportion of the “Nilo-Saharan AAC” in the southern/central Sudanese populations (Nuer, Dinka, Shilluk, Nyimang), with decreasing frequency from northern Kenya (*e.g.* Pokot) to northern Tanzania (Datog, Maasai). From K = 5-13, all Nilo-Saharan speaking populations from Kenya, Tanzania, southern Sudan, and Chad cluster with west-central Afroasiatic Chadic speaking populations (**Fig. S15**). These results are consistent with linguistic and archeological data, suggesting a possible common ancestry of Nilo-Saharan speaking populations from an eastern Sudanese homeland within the past ~10,500 years, with subsequent bi-directional migration westward to Lake Chad and southward into modern-day southern Sudan, and more recent migration eastward into Kenya and Tanzania ~3,000 ya (giving rise to Southern Nilotic speakers) and westward into Chad ~2,500 ya (giving rise to Central Sudanic speakers) (*S62, S65, S67, S74*). A proposed migration of proto-Chadic Afroasiatic speakers ~7,000 ya from the central Sahara into the Lake Chad Basin may have caused many western Nilo-Saharans to shift to Chadic languages (*S99*). Our data suggest that this shift was not accompanied by large amounts of Afroasiatic gene flow. Analyses of mtDNA provide evidence for divergence ~8,000 ya of a distinct mtDNA lineage present at high frequency in the Chadic populations and suggest an East African origin for most mtDNA lineages in these populations (*S100*).

Origins of the Fulani, Baggara Arabs, Koma, and Beja

The Saharan African Beja (Sudan) and Mozabite (Algeria) populations show high levels of Middle Eastern/European and eastern African Cushitic AACs (**Figs. 3, 4, 5B, 5C**), suggesting possible gene flow from those regions and/or common ancestry. Linguistic evidence indicates that the Afroasiatic language family originated in the Horn of Africa **(***S67, S75, S95*), consistent with high levels of the Afroasiatic AAC in the Beja populations (although the latter observation could also be due to reverse gene flow from the Middle East). When the Saharan African populations were analyzed separately, with STRUCTURE, the two Beja clan alliances (the Banuamir and the Hadandawa) were distinguished (**Fig. S22**).

Other groups of interest can be assessed with STRUCTURE analyses, including the Fulani, sampled from Nigeria and Cameroon, the Baggara sampled from northern Cameroon, and the Koma sampled from the Alantika Mountains in eastern Nigeria. The Fulani are nomadic pastoralists who speak a Niger-Kordofanian language (Atlantic Senegambian subfamily) and occupy a broad geographic range in central and western Africa. The Fulani show a number of morphological features that have led some anthropologists to suggest that they may have originated from East Africa or possibly Egypt or the Near East (*S101*). Mitochondrial DNA analysis indicates that Fulani have lineages of predominantly West African origin and that they cluster together and close to the Mandenka population from Senegal (*S101*). By contrast, Y chromosome analyses of Fulani sampled in the Sudan indicates shared ancestry with Nilo-Saharan and Afro-Asiatic speaking populations (*S97*). These results raise the possibility of differential patterns of male and female gene flow into this population. Our analysis, using genome-wide nuclear markers and STRUCTURE, indicates that the Fulani have distinctive ancestry (fuchsia) at $K = 14$ in the global analysis (**Figs. 3, 4**) and at $K = 9$ -14 in the Africa analysis (**Fig. S15**). Low to moderate levels of the Fulani AAC was also observed in the Mozabite and Mandinka populations in the global analysis (**Figs. 3 and 4**). The Fulani cluster with the Chadic and Central Sudanic speaking populations at $K \leq 13$ in the global analysis (**Fig. 3**; maroon) and at $K \leq 8$ in the Africa analysis (**Fig. S15;** red). They also cluster near the Chadic and Central Sudanic speaking populations in the NJ tree based on population genetic distances (**Figs. 1, S7 and S8**). In the global STRUCTURE analysis, the Fulani show low to moderate levels of European/Middle Eastern ancestry (blue), consistent with mtDNA (*S101*) and Y chromosome (*S97*) analyses, as well as the presence at low frequency of the -13910T mutation associated with lactose tolerance in Europeans in this population (*S102*). Additionally, we observe moderate to high levels of Niger- Kordofanian ancestry in the Fulani populations (**Figs. 3, 4, S15;Tables S8, S9**). These results do not enable us to determine the definitive origin of the Fulani, although they indicate shared ancestry with Saharan and Central Sudanic populations and suggest that the Fulani have admixed with local populations, and possibly adopted a Niger-Kordofanian language, during their spread across central and western Africa. The origin of European (possibly via the Iberian peninsula) and/or Middle Eastern ancestry in the Fulani requires further exploration with additional genetic markers.

The Baggara (or Baggara Arabs or Shuwa Arabs) are nomadic pastoralists who speak an Afroasiatic Semitic language and inhabit regions ranging from southern Sudan to Nigeria. They are thought to be descendants of tribes originating from the Arabian Peninsula. More specifically, the Baggara trace their ancestry to the Banu Judham tribe of Yemen. This tribe was among the first tribes to settle in Egypt during the Muslim conquest in mid 7th Century AD *(S103)* and during the Shiite Fatimid dynasty (AD 909 to 1171). After the fall of Fatimid and rise of other subsequent dynasties such as the Ayyubid (1171-1250) and Mamluk Turk Sultans of Bahri (1250-1517), they are thought to have moved southwest and subsequently settled in present day Chad and then subsequently into northern Cameroon and Nigeria within the past 300 years. The earliest records of their existence in Chad are from the late 14th century, although they may have been there before then (*S104-108*).

Consistent with their proposed history of migration from Arabia across eastern Africa, southern Sudan, and the Sahel, the Baggara show low levels of Middle Eastern/European associated ancestry (blue) and high to moderate levels of Cushitic (purple) and Nilo-Saharan (red) associated ancestry in the global and African STRUCTURE analyses (**Fig. 3, 4, S15**). They also show ancestry from the Niger-Kordofanian, Fulani, and Chadic AACs (**Fig. 5B,C**), suggesting that they admixed with local populations as they migrated westward, consistent with studies of mtDNA(S*96*). These results are consistent with the phylogenetic trees of population genetic distances where they cluster near the Chadic and Central Sudanic-speaking populations (**Figs. 1, S7, and S8**).

The Koma, a population previously uncharacterized at the genetic level, who currently reside at high elevations in the Alantika mountain range bordering Nigeria and Cameroon, speak a Niger-Kordofanian language (sub-classified as Adamawa-Ubangi) and descend from populations that retreated into the mountains during the 18th century to take refuge during Fulani invasions (*S109*). In the D^2 phylogenetic tree (**Fig. 1**), they form a clade with Mbum from C.A.R. (who also speak an Adamawa-Ubangi language) and cluster near the Central Sudanic-speaking populations. In the African STRUCTURE analysis (**Figs. 5B, 5C, S15**), the Koma show predominantly Niger-Kordofanian ancestry at most K values, but at K = 14 they show moderate levels of shared ancestry with the Central Sudanic AAC (tan).

Population history of East Africa

Here we characterize the genetic history of several eastern Africa populations whose origin was previously unknown. It is probable that the observed patterns from the STRUCTURE analyses are the result of hypothesized successive waves of migration into eastern Africa. The indigenous populations of East Africa are thought to be the click-speaking Hadza and Sandawe hunter-gatherers of Tanzania, whose populations may have originally extended from Kenya to Somalia and possibly into Ethiopia (*S50, S110*). The first wave of migration is thought to be by Southern Cushitic speakers (ancestral to the Iraqw, Gorowa (Fiome), Burunge, and Mbugu), moving south from Ethiopia into Kenya and then into Tanzania where they currently reside (S*84,S 85, S110, S111*). There are also two linguistically attested movements of Eastern Cushitic-speaking forager populations at ~4,000 and ~2,000 years, originating from a proposed homeland north of Lake Turkana into Kenya and Tanzania (Yaaku, El Molo) and four other movements originating from southern Ethiopia and extending into northern Kenya (Rendille; Gabra, Borana, and Wata; Burji; and Konso) at successive periods from ~2,500 ya through the present (S*62*). Cushitic peoples since the proto-Cushitic period of ~10,000-9,000 kya practiced pastoralism, and at around 7,000 ya began to cultivate grain crops as well (S*67*). Generally accepted archaeological correlations show that the movement of the Southern Nilotes (e.g. Kalenjin, Okiek, Datog) south from the present-day Sudan/Ethiopia border region into western and central Kenya and subsequently into central northern Tanzania took place between ~2,900-2,400 kya (S*62, S84, S111*). These movements took place at approximately the same time as the settlement of the early Mashariki Bantu in the Lake Victoria and Lake Tanganyika Basins (*S83*). The Mashariki Bantu subsequently migrated out of these regions across eastern and southeastern Africa between ~2,300-1,700 kya (S*83*). Eastern Nilotes (e.g. Turkana, Samburu, Maasai) represent a more recent migration from southern Sudan, within the past ~1,500 - 500 years (*S62, S84, S110, S111*). These patterns of migration are expected to result in a highly diverse and complex genetic structure in East Africa compared to other regions, as we observe.

Within eastern Africa, including southern/central Sudan, clustering is primarily associated with language families, including Niger-Kordofanian (orange), Afroasiatic (purple), Nilo-Saharan (red) and two click-speaking hunter-gatherer groups: the Sandawe (brown) and Hadza (yellow) (**Fig. S19-S21**). However, individuals from the Afroasiatic Cushitic Iraqw and Gorowa (Fiome) and the Nilo-Saharan Datog (dark green), who are close geographically, also cluster. Additionally, several hunter-gatherer populations were distinct, including the Okiek (blue), Akie (pink), and Yaaku and El Molo (dark grey). Of particular interest is the common ancestry of the Akie (who have remnants of a Cushitic language) and the Eastern Cushitic El Molo and Yaaku at K = 9, consistent with linguistic data suggesting that these populations originated from southern Ethiopia and migrated into Kenya and Tanzania within the past ~4,000 years (S*62, S91, S110*). Note that possible cryptic relatedness (more distantly related than 3rd degree relatives) among the Hadza, whose census size is only 1,000, as well as genetic drift, could contribute to their genetic distinctiveness in the STRUCTURE and PCA analyses.

Our data also shed light on the history of particular eastern African populations. The Mbugu, who live in the Usambara mountain range in eastern Tanzania, speak a unique "mixed language" that contains Bantu syntax and Cushitic vocabulary (*S50, S92*). According to oral tradition, the original Mbugu homeland was in Lukupuya*,* which some scholars suggest might have been the Laikipia Plateau in Kenya, where many Maasai currently reside (*S112*). Their oral history further suggests that they have experienced conflict in the sixteenth and seventeenth centuries with neighboring Maasai, who allegedly stole their cattle, and with the neighboring Pare and Sambaa, with whom they eventually reached a peaceful accommodation (*S112*). It has been speculated that they maintained their indigenous language (which they call Ma'a) in order to maintain their cultural identity and distinction from neighboring societies (*S112, S113*). In our STRUCTURE analyses of Africa and East Africa (**Figs. 5B, 5C, S15, S16, S19**) the Mbugu form a distinct AAC (dark purple), but with moderate levels of the Niger-Kordofanian AAC, consistent with gene flow with neighboring Bantu populations. Additionally, we observe evidence for low levels of gene flow from the Mbugu into neighboring Bantu-speaking Pare and Sambaa populations (with whom they share language similarities (*S50, S92*). In the NJ tree of genetic distance between AACs from the Africa STRUCTURE analysis (**Fig. S18**), the Mbugu AAC clusters together with the Fulani, Cushitic, and Saharan AACs, consistent with their proposed Cushitic origins. However, in the phylogenetic analyses, the Mbugu cluster near the Nilo-Saharan speaking populations (**Figs. 1, S7, S8**), perhaps reflecting historic admixture.

The genetic history of the Sandawe is described in the main text. Here we note that we see evidence of gene flow from the click-speaking Sandawe into the neighboring Cushitic speaking Burunge, and Bantu-speaking Turu (**Figs. 3, 4, 5B, 5C, S15, S16, S19**), consistent with language and technology exchange between these groups (S50, S91, S110).

History of East African hunter-gatherer populations

In addition to the history of the Khoesan-speakers and Pygmies, our data also shed light on the population history of several other hunter-gatherer populations in Africa, whose subsistence is from hunting small game, foraging for roots and plants, and collecting honey. The term Dorobo (considered derogatory) is commonly used by Maasai populations to refer to hunter-gatherer populations who have "lost their cattle". Here, we refer to these populations by their self-identified ethnicity, except in the case of a group of Dorobo who live near the Maasai in the Ngorongoro district of Tanzania and do not have a distinct ethnic affiliation. The Dorobo (Tanzania), Okiek (Kenya), and Akie (Tanzania) are foragers who live near the Maasai and speak the Maasai language, although the Okiek also speak several different Kalenjin Southern Nilotic languages of their own (*S114*) and the Akie, too, speak a Kalenjin language that contains loanwords from an extinct Rift Southern Cushitic language related to Burunge and Iraqw *(S110)*. Some anthropologists have hypothesized that these populations are genetically Maasai who no longer raise cattle, while others argue that they are descendants of a more ancient group of East African foragers (*S111, S115*). Our data indicate that the Dorobo cluster near the Maasai in the phylogenetic analyses (**Figs. 1, S7 and S8**) and are not distinguishable from the Maasai on the basis of STRUCTURE analysis (**Figs. S15 and S19**). The Okiek cluster near the Maasai and other Nilo-Saharan-speaking populations in the phylogenetic trees (**Figs. 1, S7 and S8**) and are included in the Nilo-Saharan AAC from K = 2-7 in the STRUCTURE analysis of East Africa (**Fig. S19**) but form a distinct AAC at K = 8 and higher, suggesting that they share common ancestry with Nilo-Saharans but have more recently become differentiated. Interestingly, they cluster with the Mbugu at K = 8 (**Fig. S19**) perhaps, indicating Cushitic ancestry as well. Therefore, our data and analyses support the conclusion that the Dorobo and Okiek are genetically related to Nilo-Saharan speaking populations, and that they may have adopted a foraging subsistence pattern. However, additional analyses will be required to determine the time of divergence of the Okiek from the other Nilo-Saharan speaking (and possibly Cushitic) populations. Indeed, linguistic and archaeological evidence suggests that the Okiek have a significant degree of cultural continuity back to the Eburran hunter-gatherer populations of 12,000-2,000 kya in Kenya (*S83, S84, S110, S111*). It is possible that the Okiek differentiated from the other Nilo-Saharan speaking populations prior to the origins of cattle domestication within the past ~5,500 years in East Africa (*S116*).

In contrast, the Akie, who have a unique AAC at K = 10 in the East Africa STRUCTURE analysis (**Fig. S19**), appear to share ancestry with neighboring Tanzanians at K = 2 – 8 (**Fig. S19**), and share a distinct AAC at K = 9 with the Eastern Cushitic El

Molo, who are fishermen on the eastern shore of Lake Turkana in Northern Kenya, and with the Eastern Cushitic Yaaku from southern Kenya, who historically were foragers for honey, plants, and small game (*S115, S117*). The El Molo and Yaaku form a distinct AAC at K = 10 and we observe considerable asymmetric gene flow from the Yaaku into the neighboring Maasai (Maasai Mumonyot and Maasai Il'Ngwesi) and Samburu populations. Additionally, the Akie cluster close to Cushitic speaking populations, including the Yaaku, in the phylogenetic trees (**Figs. 1, S7, S8**), and interestingly, they cluster close to the Okiek in the D^2 phylogenetic tree (**Fig. 1**). These results are consistent with linguistic and archeological data suggesting that the Akie, Yaaku, and El Molo were part of an early wave of Cushitic speaking populations into Kenya and Tanzania within the past 5,000 years (*S57*) who likely practiced a foraging subsistence pattern, and in the case of the Akie and Yaaku, more recently adopted a Nilo-Saharan language and/or culture (*S62, S84, S110, S111, S115, S118*), and admixed with Nilo-Saharans. The Yaaku in fact adopted the Maasai language only in the past 80 years (*S115, S117*). The Wata (also known as Boni) hunter-gathers from northern Kenya, are indistinguishable from neighboring Cushitic agro-pastoralist groups in the STRUCTURE analyses (**Figs S15, S19**), and cluster near the Cushitic speaking populations from northern Kenya and southern Ethiopia in the phylogenetic trees (**Figs. 1, S7, S8**), suggesting that they have recently adopted a hunting-gathering subsistence pattern, consistent with oral tradition. We do not find evidence of shared ancestry of any of the East African foraging groups with the indigenous Hadza and Sandawe populations of Tanzania, suggesting that other East African foraging groups were part of more recent migrations into the region.

Common origin of Pygmies and Khoesan-speakers

Our observation that the Pygmies appear to share common ancestry with several Khoesan-speaking populations raises the possibility that the indigenous Pygmy language may have contained click consonants. A recent examination of the skeletal evidence for the development of the human vocal tract indicates that full human language capacity evolved before 50 kya but after 100 kya (*S119*). Considering that the normative directions of phonological evolution are from greater to lesser markedness, and that clicks are among the most marked of all sounds, the fact that click consonants exist at all in present-day languages favors their existence back to the earliest human languages of 100-50 kya. Some scholars have proposed that the language families of the world can be divided into two primary branches (*S120*). In this view a single extant language family, Khoesan, which preserves clicks, is the last representative of one primary branch; all the rest of the language families of the world descend from the second branch, the defining phonological development of which was the dropping of clicks from the consonant inventory (*S120*). Did the extinct language family or families of the Pygmies belong to the same primary branch as Khoesan? The apparent shared ancestry of Pygmies with several Khoesan speaking populations suggests that this is a possibility. It should be noted that some linguists have identified words associated with forest terms or with hunting that may be shared amongst diverse Pygmy populations and could potentially stem from an ancestral indigenous Pygmy language (*S121-123*). These words do not have click consonants. Future studies may be informative for reconstructing a proto-Pygmy language and for examining possible connections with modern Khoesan languages. However, because of the limited time depth for resolving language phylogenies, it may not be possible to reconstruct such ancient linguistic affiliations.

Our observation of possible shared common ancestry amongst the SAK, Pygmies, and Hadza, who are all short statured, also raises the possibility that short stature may have been an ancestral trait rather than a recent adaptation in Pygmies to a tropical environment. Indeed, it is possible that the ancestors of these populations lived in a savannah environment, and that Pygmies migrated more recently into the tropical forest. However, it is also possible that short stature arose through convergent adaptation in these different populations. Future genome-wide genetic and phenotypic analyses will be informative for distinguishing the effects of local adaptation and genetic drift in these geographically diverse hunter-gatherer populations.

Gene-language and gene-geography associations

Any consideration of the relationship of linguistic and genetic diversity must start with a major caveat. The spread of languages into new areas does indeed require the migration/movement of at least some speakers of the languages. But this number may range from a major movement of many individuals into a new region to only a few (*S124*). In addition, it is possible that language shift could occur without genetic exchange. One particular linguistic tool, the study of word borrowing patterns between two languages over a sustained period of time, may have particular utility in generating predictions about the relative proportions of the populations involved in histories of language shift (*S94*). The predictive capacities of this type of evidence have already received mention above with respect to the Khoekhoe demic components in South African Bantu-speaking societies (*S93*).

In most cases, we observed a strong correlation between genetic clustering and language classification, consistent with prior studies (*S124-S127*). However, we found several exceptions. In these cases, the word-borrowing histories show that the former language of the majority of the population typically did fit with the genetic clustering. For example, in the phylogenetic trees (**Figs. 1, S7, S8**), Afroasiatic speakers cluster into four primary groups, which correlate with geography; one group consists of Semitic-speaking and Berber-speaking populations in the Sahara, a second consists of Cushitic-speaking populations from southern Ethiopia/Northern Kenya, a third consists of Southern Cushitic-speaking populations from Kenya and Tanzania, and a fourth consists of Chadic-speaking populations from northern Cameroon and Chad. The Afroasiatic Chadic speaking populations from Northern Cameroon cluster close to the Nilo-Saharan-speaking populations from southern Sudan in the $(\delta\mu)^2$ phylogenetic tree (**Fig. S7**) and close to the Central Sudanic speaking populations in the D^2 and R_{ST} trees (**Figs. 1 and S8**) rather than with East African Afroasiatic speakers, consistent with STRUCTURE results (**Figs. 5B, 5C, S15**). These results are consistent with the linguistic evidence of notable Nilo-Saharan loanword sets in the Chadic languages of the types that imply a shift of former Nilo-Saharan-speaking populations to Chadic languages (*S99*). The Hausa, who speak a Chadic Afroasiatic language cluster with the Niger-Kordofanian speaking populations in the phylogenetic trees (**Figs. 1, S7, S8**), consistent with the high levels of the Niger-Kordofanian AACs observed in this population on the basis of STRUCTURE analysis (**Figs. 3, 4, 5B, 5C, S15, S24**) and in keeping with long-recognized linguistic evidence of Niger-Kordofanian influence on the Hausa, including several salient loanwords, such as the word for meat (*S128*). The Dogon from Mali, who speak a Niger-Kordofanian language, cluster near the Saharan populations in the phylogenetic trees (**Figs. 1, S7, S8**), consistent with the results from STRUCTURE analysis, showing considerable Saharan (blue) ancestry, and consistent with oral history of a northern African origin (although it should be noted that the sample size for this population, 9 individuals, is very small and many markers did not amplify well) (**Figs. 5B and 5C; Table S9**). The linguistic evidence remains to be studied in this case.

Regression analyses of linguistic, geographic, and genetic distance are given in the main text. Here we note that co-linearity between geographic and linguistic distance could confound interpretation of results. Specifically, for the Cushitic only analyses we found that both linguistic distance and geographic distance explain a significant portion of the genetic variance (0.27 and 0.29, respectively; $p < 0.0001$ for both). However, in the analysis that adjusts geographic distance for language the p value increases to 0.67, but language distance remains a significant factor after adjusting for geography. This result is consistent with the co-linearity of geography and language found in this group. In contrast, we did not find evidence for high levels of co-linearity in the other language families.

Genetic ancestry of African Americans

Ancestry from multiple global populations was detected in both the African American and CMA populations. In contrast to prior studies of African American (*S129*) ancestry that focused on uniparentally inherited mtDNA or Y chromosome markers (*S129-131*), or on nuclear markers genotyped in a small subset of Africans (*S129, S132, S133*), the current study infers African American ancestry across the nuclear genome by comparison with 121 geographically and ethnically diverse African populations and an extensive sample of 60 non-African populations. In African American populations, from Chicago, Baltimore, Pittsburgh, and North Carolina, ancestry was predominantly from the African Niger-Kordofanian AAC (means 0.69-0.74), which is most common in western Africa, and the European/Middle Eastern AAC (means 0.11-0.15) **(Fig. 6 and Table S6)**, consistent with prior studies (*S129-133*). This result is also consistent with the history of the slave trade, indicating that most slaves reached North America, often via the Carribean, ultimately from the western coasts of Africa (*S134*). Low levels of ancestry from several additional populations were also detected (**Table S6**): Fulani (means 0.0 - 0.03, individual range 0.00-0.14), Cushitic East African (means 0.02, individual range 0.05 - 0.10), Sandawe East African (means 0.01- 0.03, individual range 0.00 - 0.12), East Asian (means 0.01 – 0.02, individual range 0.0 - 0.08), and Indian (means 0.04 – 0.06, individual range 0.01 -0.17). The Fulani are present across West Africa and, therefore, would be expected to have contributed to the slave trade, and the Cushitic and Sandawe ancestry could represent slave trade originating from the east coast of Africa (*S134*). It should be noted that the levels of Indian ancestry in African Americans may be slightly overestimated, and the levels of European ancestry slightly underestimated, due to moderate levels of the Indian AAC in European/Middle Eastern individuals (**Figs. 3 and 4**). We did not observe significant levels of Native American ancestry. However, other regions of the U.S., may reveal Native American Ancestry, as previously reported (*S133*). Finally, European and African ancestry levels varied considerably among individuals (**Fig. 6**).

Origins of the Cape Mixed Ancestry population

Based on unsupervised STRUCTURE analysis, this population shows nearly equal high levels of southern African Khoesan (mean 0.25, individual range 0.01-0.48), Niger-Kordofanian (mean 0.19, individual range 0.01 - 0.71), Indian (mean 0.20, individual range 0.0 - 0.69), and European (mean 0.19, individual range 0.0 - 0.86) ancestry (**Fig. 6; Table S6**). The CMA population also has low levels of East Asian (mean 0.08, individual range 0.0 – 0.21) and Cushitic (mean 0.03, individual range 0.0 – 0.40) ancestry. These results are consistent with the history of the CMA population, which is thought to have descended from the indigenous Khoekhoe (Khoesan-speaking herders), and admixed initially with European Dutch "Afrikaaner" colonialists from 1652 up to the present as well as with Bantu-speaking slaves from West Africa and Mozambique, and with Austronesian-speaking slaves from Madagascar and Indonesia during the 18th and early 19th centuries (*S135, S136*). Additionally, there were many Indian and a few Chinese in the Capetown area, from the 1860s onward, who are thought to have contributed to CMA ancestry (*S135*). These results are consistent with the supervised STRUCTURE analysis, using the same set of training populations as was used for determining African American ancestry (**Table S7**). As expected, the proportion of Bantu Niger-Kordofanian ancestry (mean 0.10) is higher than the proportion of non-Bantu Niger-Kordofanian ancestry (mean 0.04) in the CMA.

The genetic, linguistic, and geographic landscape of Africa

A number of factors could have contributed to isolation and differentiation of populations in Africa, including climatic fluctuations such as an extreme dry spell from ~60,000 – 30,000 ya (*S137*) , the cold and dry last glacial maximum in Africa, ~21K – 15 kya, followed at ~13,700 – 12,300 ya by another period of cooler and drier conditions (*S93*). Indeed, the period of diversification of the major African language families is thought to date back to ~15K – 11 kya. Differentiation and expansion of ancestral African populations across Africa from regions of refugia, such as the Ethiopian highlands, may have occurred during the interludes of these cold and dry periods (*S93*). Additionally, based on the archeological record, expansion and differentiation of African populations during the later stages of the African Middle Stone Age (MSA), from ~75,000 – 55,000 years ago, may have coincided with the origin and spread of radically new technological developments, including new patterns of blade technology, tools used for skin working and for cutting bone and wood, and stone barbs using for hunting and fishing as well long distance trade of shell ornamentation and red ochre engraved with unique decorative patterns interpreted as "abstract art" (*S137-139*). Analyses of nucleotide and haplotype variation in Africans will be informative for estimating the age of the inferred ancestral clusters and for distinguishing genetic similarity due to common ancestry or gene flow.

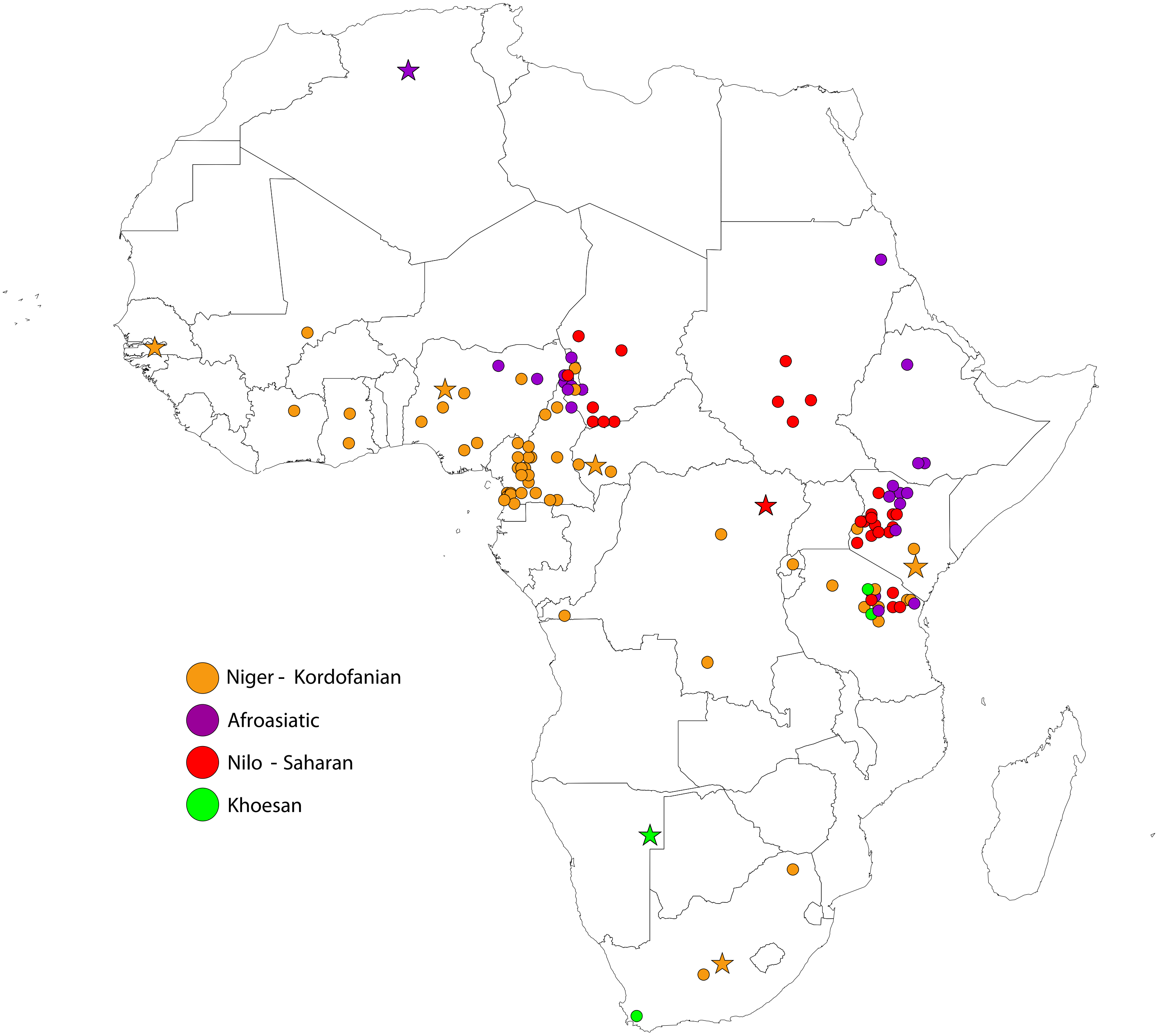


Figure S1 : Geographic distribution of 121 African populations included in the current study. Populations are color-coded according to language classification. Stars indicate samples from the Centre d'Etude du Polymorphisme Humain (CEPH) human genome diversity panel (HGDP) (S4-6, S30).

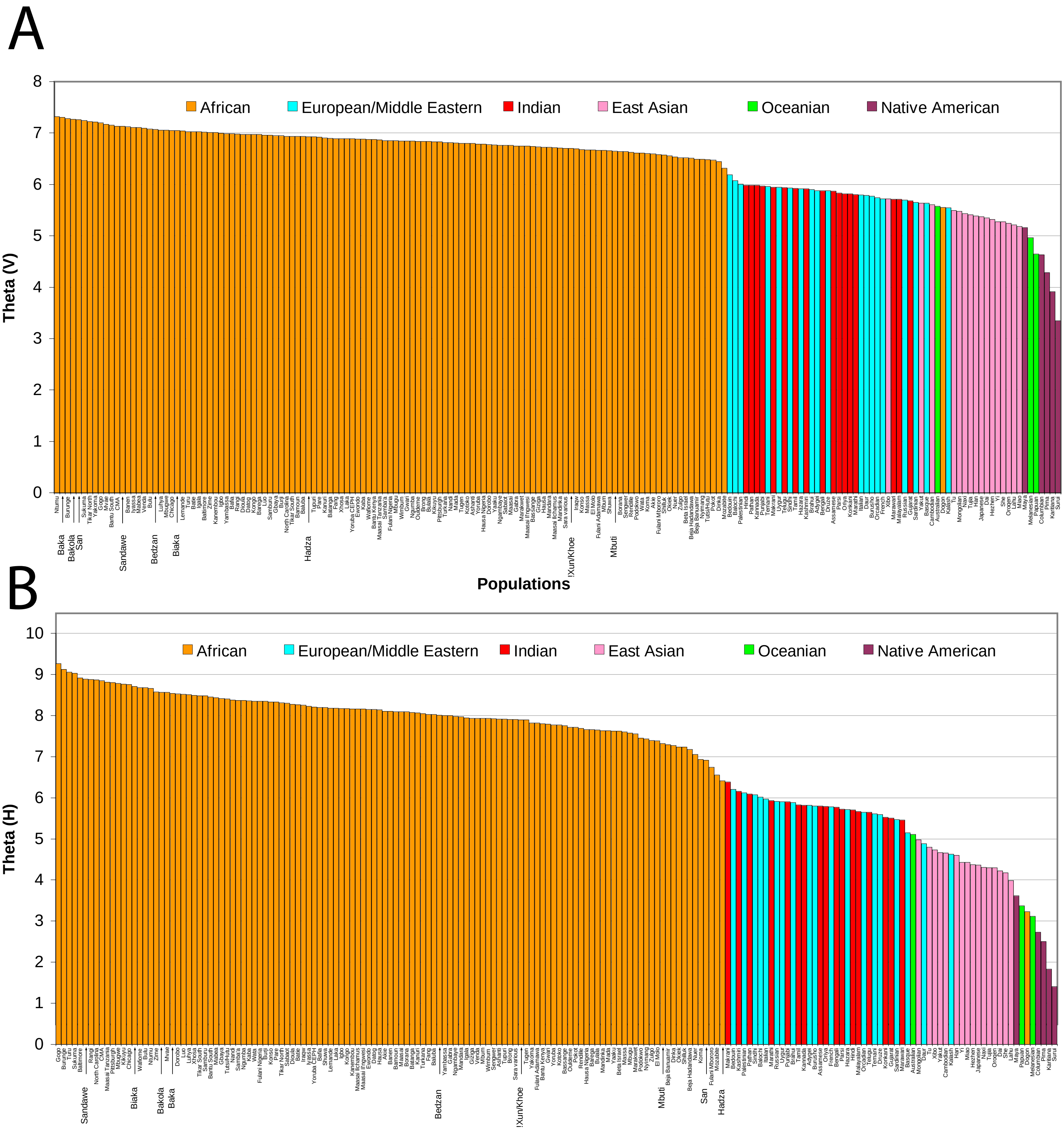

A
Theta (V)
0
1
2
3
4
5
6
7
8
African
European/Middle Eastern
Indian
East Asian
Oceanian
Native American
Baka
Bakola
San
Sandawe
Bedzan
Biaka
Hadza
!Xun/Khoe
Mbuti
Populations
B
Theta (H)
0
1
2
3
4
5
6
7
8
9
10
African
European/Middle Eastern
Indian
East Asian
Oceanian
Native American
Sandawe
Biaka
Bakola
Baka
Bedzan
!Xun/Khoe
Mbuti
San
Hadza
Populations

Figure S2

C

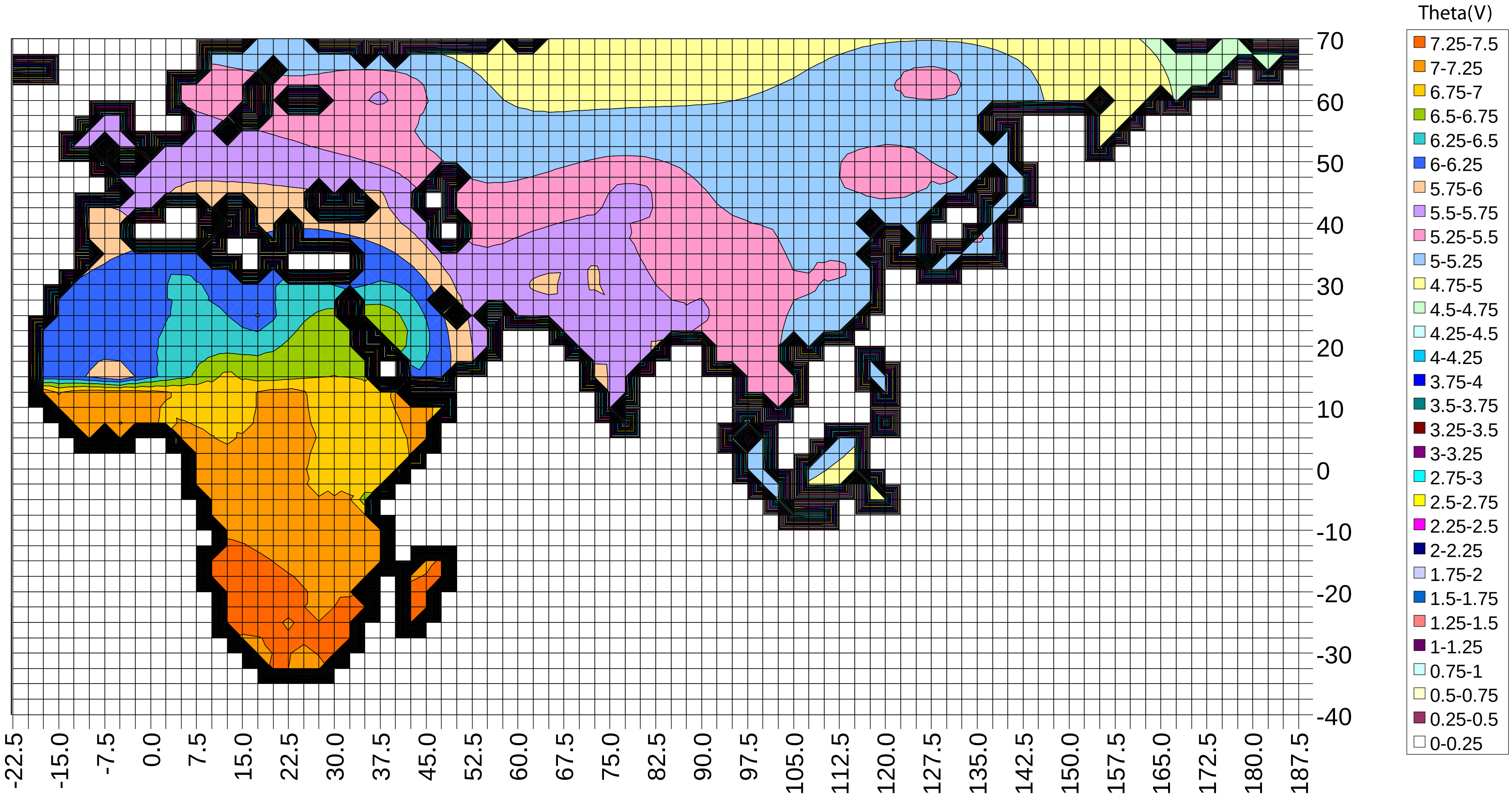


Figure S2 : Comparison of genetic diversity *(θ)* **A)** from microsatellite allele size variance, **B)** from heterozygosity. Pygmy, San, Hadza, and Sandawe hunter gatherer populations are highlighted in larger font, and **C**) Estimated levels of ***θ*** by using a linear regression weighted by geographic distance away from each point of evaluation.

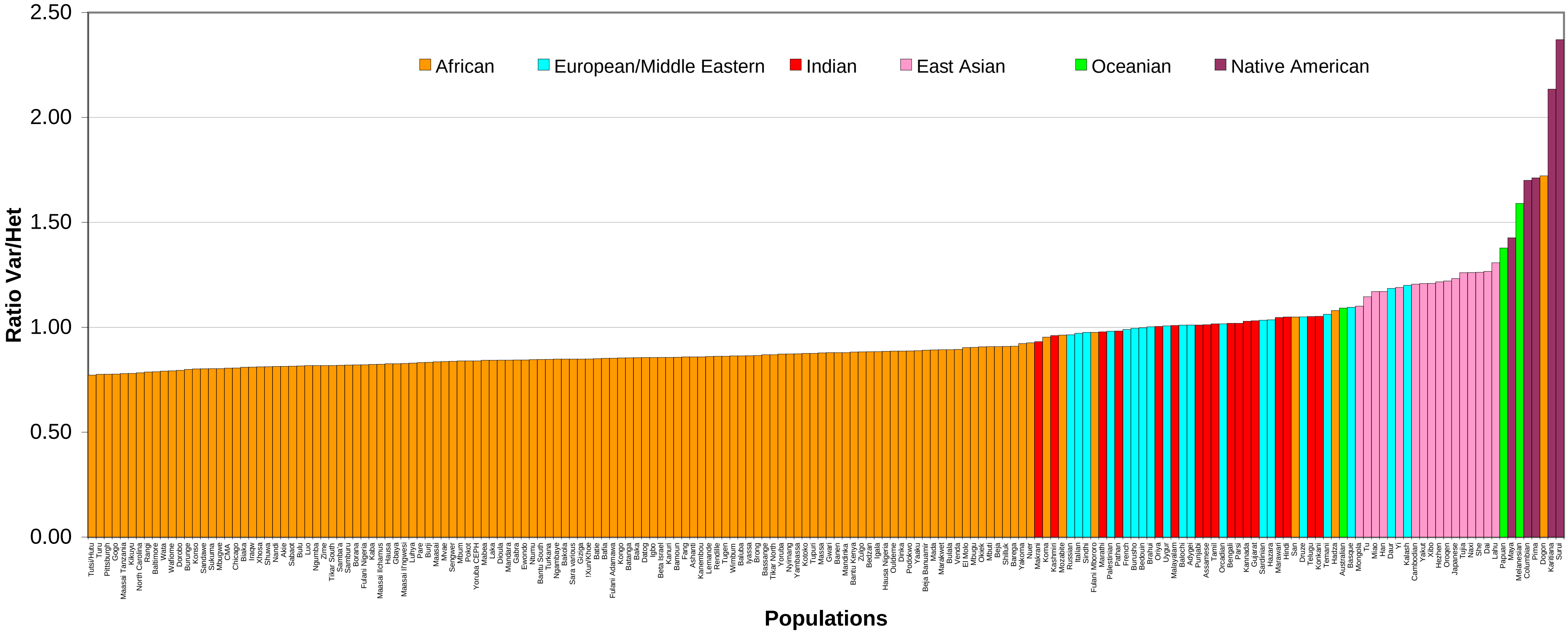


Figure S3: Ratio of genetic diversity *(θ)* inferred from the sample variance in repeat length of the microsatellite alleles to that from heterozygosity (Het)

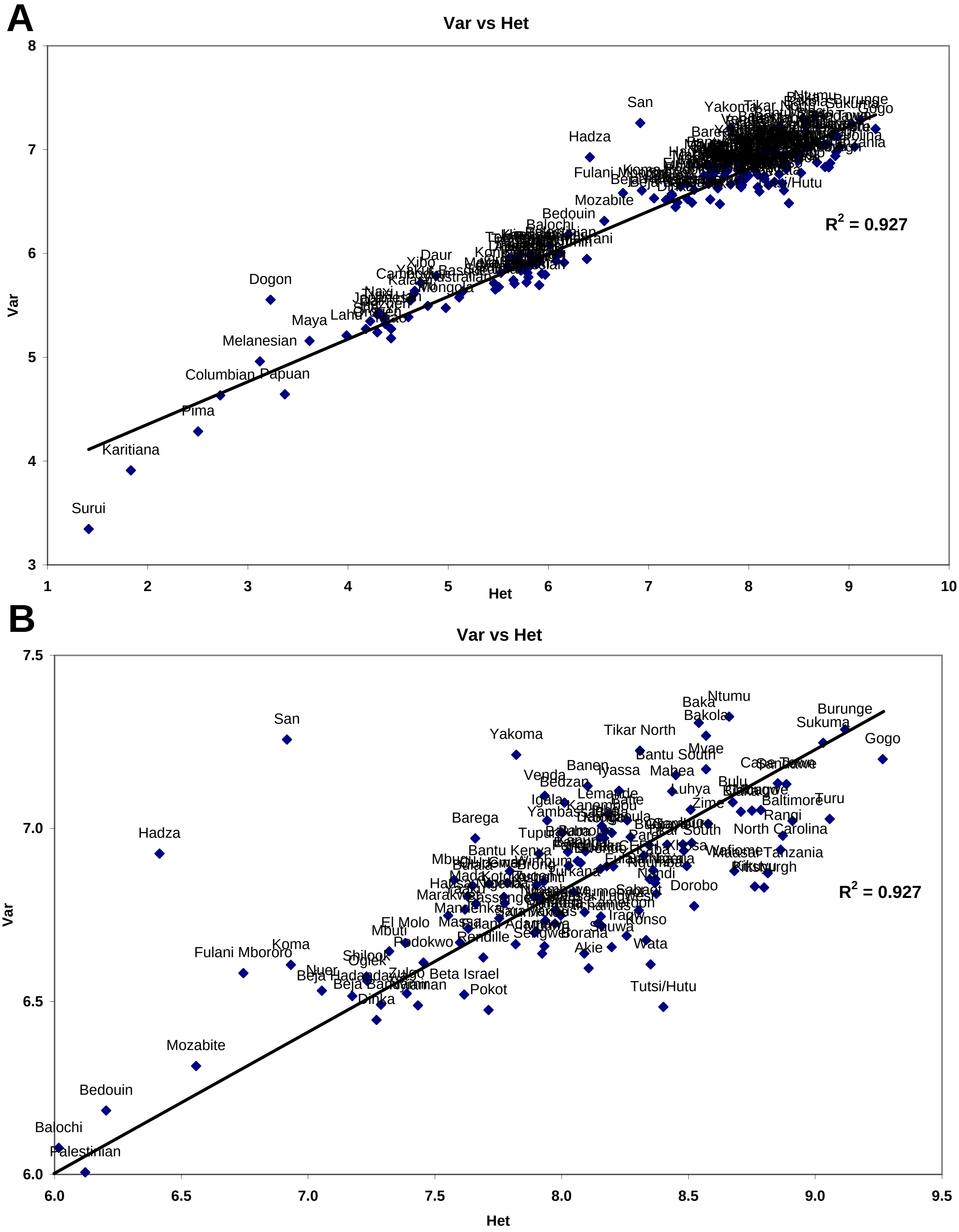


Figure S4: Correlations between genetic diversity *(θ)* inferred from the sample variance in repeat length of the microsatellite alleles (Var) and from heterozygosity (Het). **A)** Comparisons of the two measures of genetic diversity for all studied populations, and **B)** A blowup of populations with higher levels of diversity from the upper right hand corner of panel A.

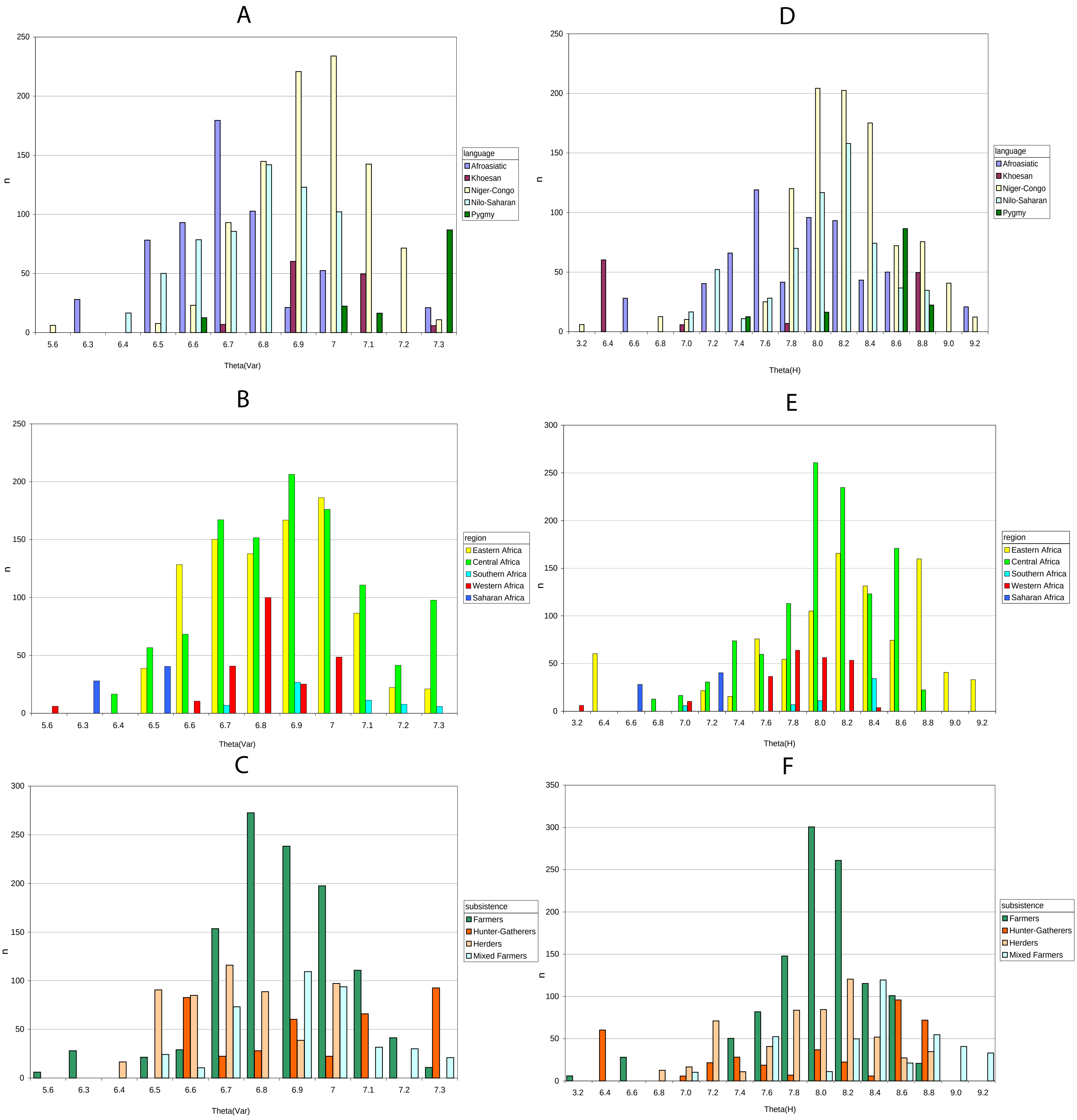


Figure S5: Genetic diversity *(θ)* from the sample variance in repeat length of the microsatellite alleles (A – C) and heterozygosity (D – F) for populations sorted by language classification (A, D), geographic location (B, E), and subsistence patterns (C,F).

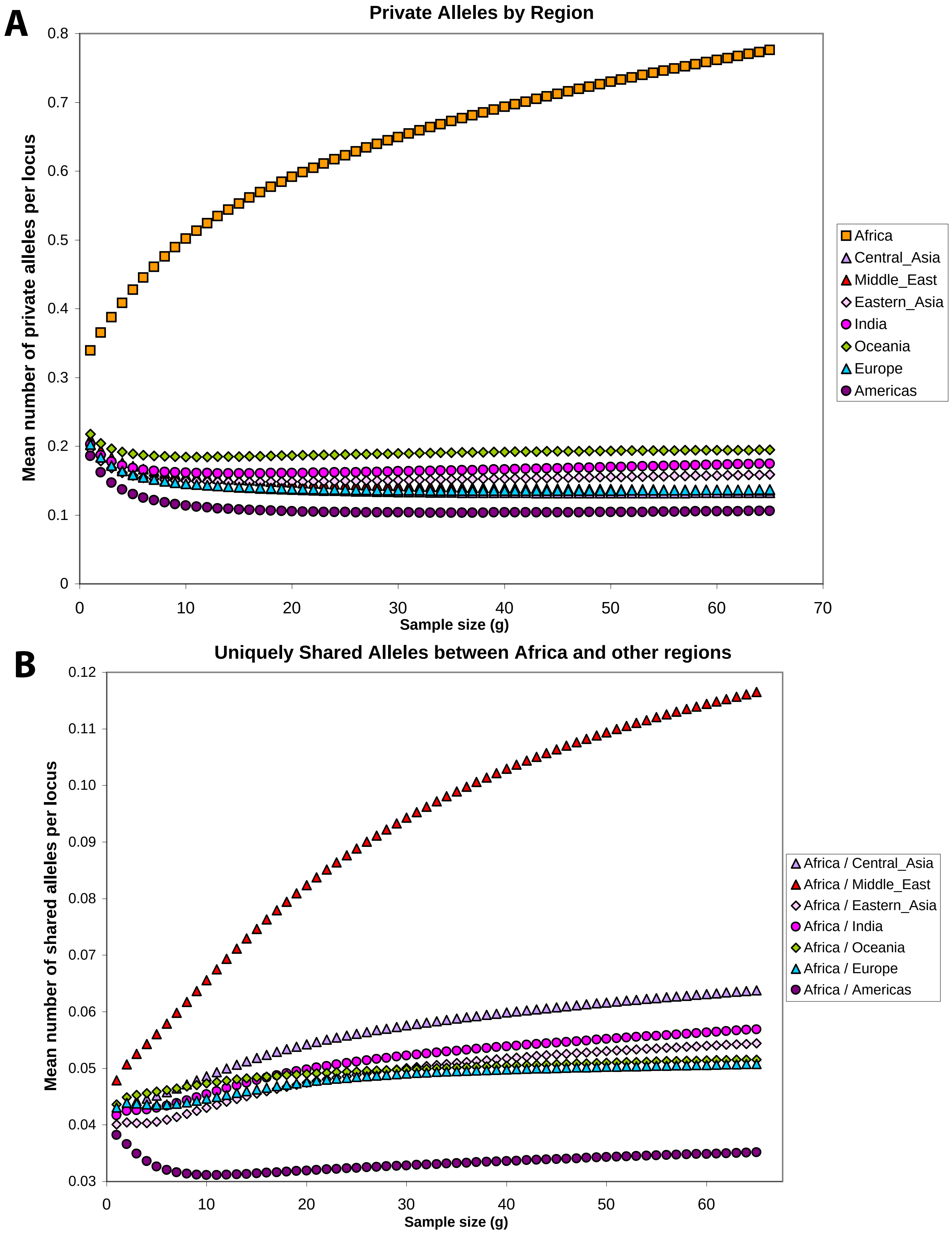

A
Private Alleles by Region
Mean number of private alleles per locus
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
Sample size (g)
Africa
Central_Asia
Middle_East
Eastern_Asia
India
Oceania
Europe
Americas
B
Uniquely Shared Alleles between Africa and other regions
Mean number of shared alleles per locus
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0
10
20
30
40
50
60
Sample size (g)
Africa / Central_Asia
Africa / Middle_East
Africa / Eastern_Asia
Africa / India
Africa / Oceania
Africa / Europe
Africa / Americas

Figure S6

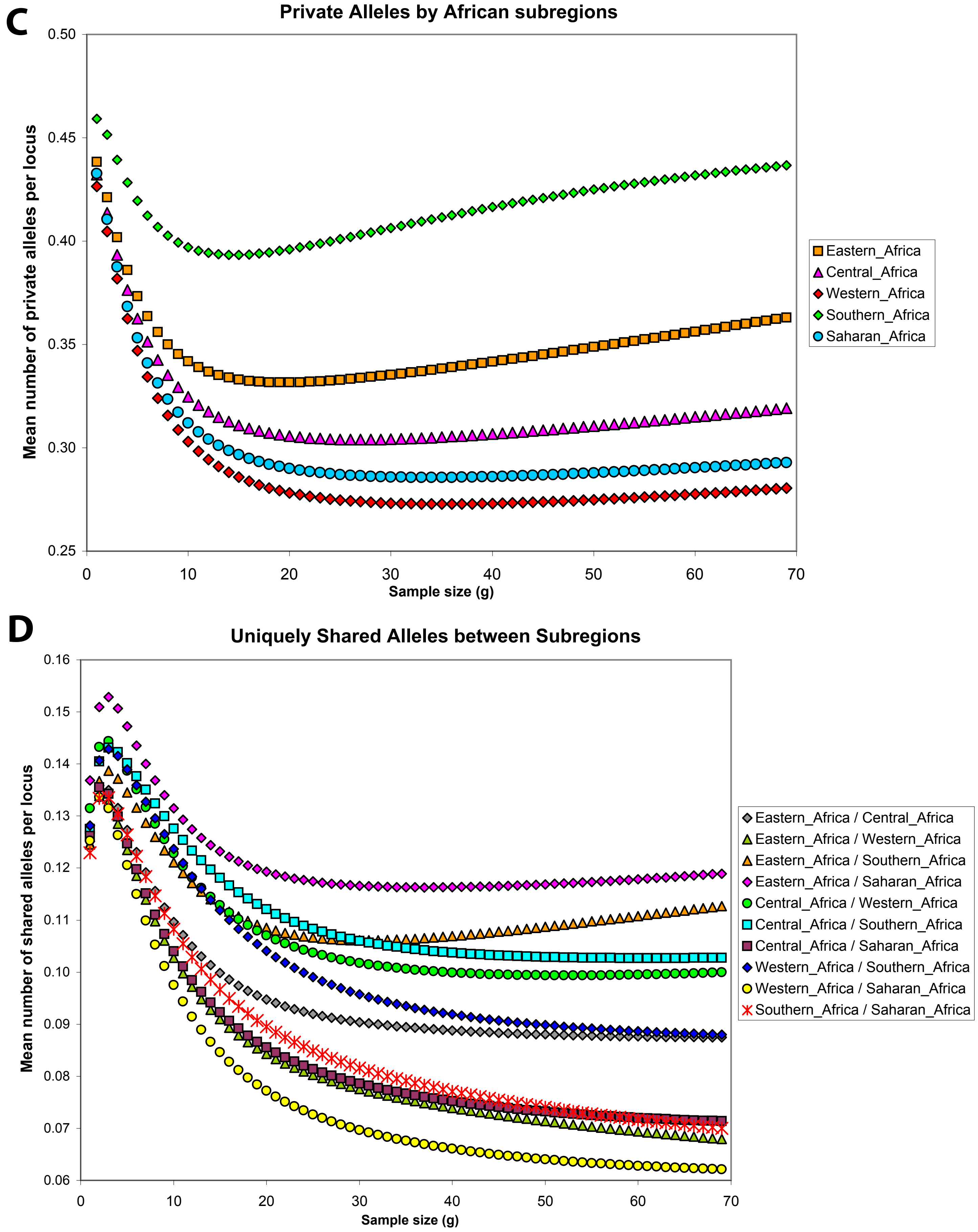


Figure S6: Inference of number of private and shared alleles across regions with the ADZE program (*S27*). This method estimates the number of alleles found in each of a set of populations but absent in all remaining populations, considering equal-sized sub-samples from each population. A) Number of private alleles by major geographic regions B). Uniquely shared alleles between Africa and other regions. C) Number of private alleles within African sub-regions. D) Uniquely shared alleles between African sub-regions.

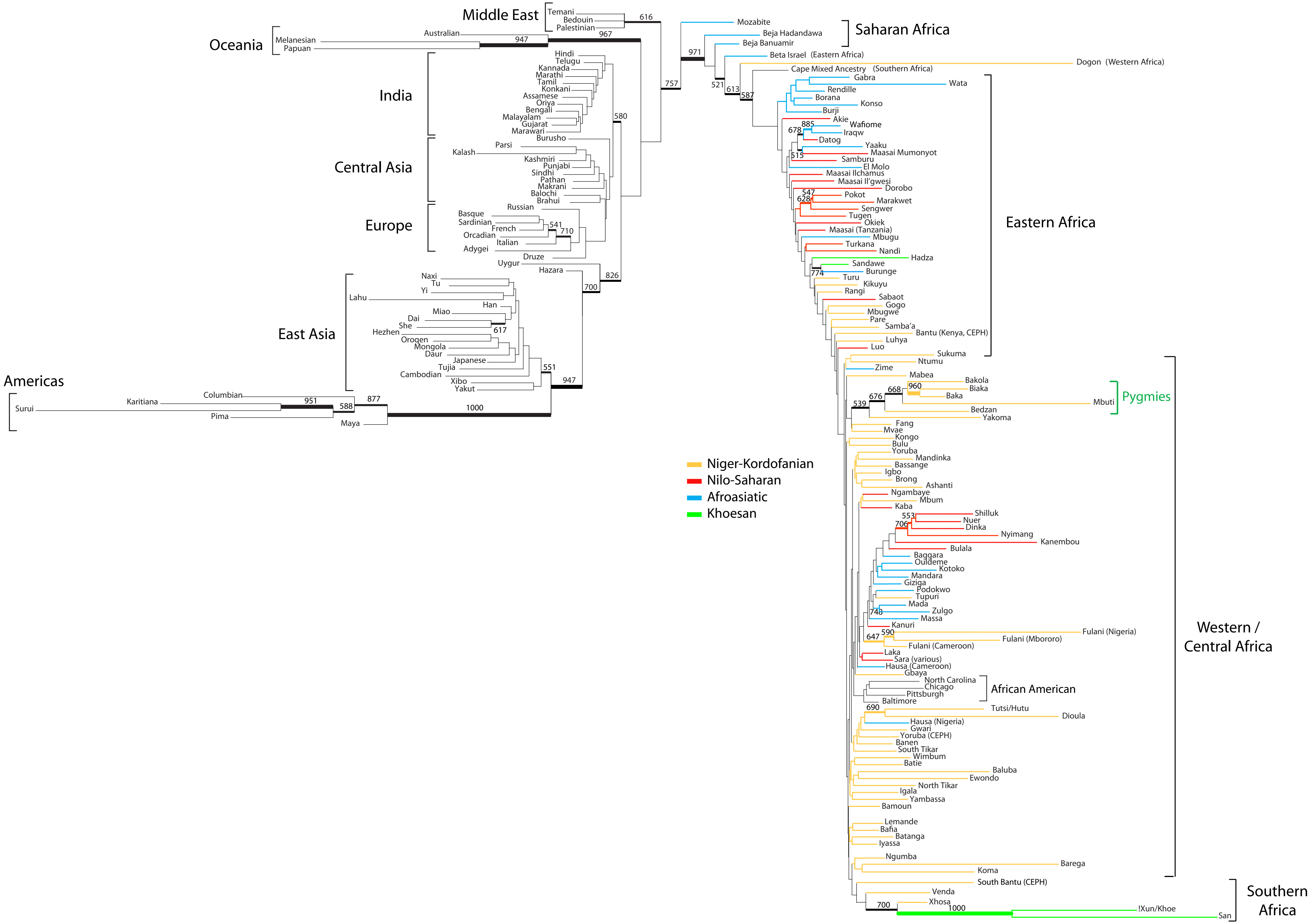


Figure S7: Neighbor-joining tree from pairwise $(\delta\mu)^2$ microsatellite genetic distances between populations (*S16*). African population branches (right) are color-coded according to language family classification; non-Africans are shown on the left. Population clusters by major geographic region are noted. Nodes supported by bootstrap values of at least 50% are labeled.

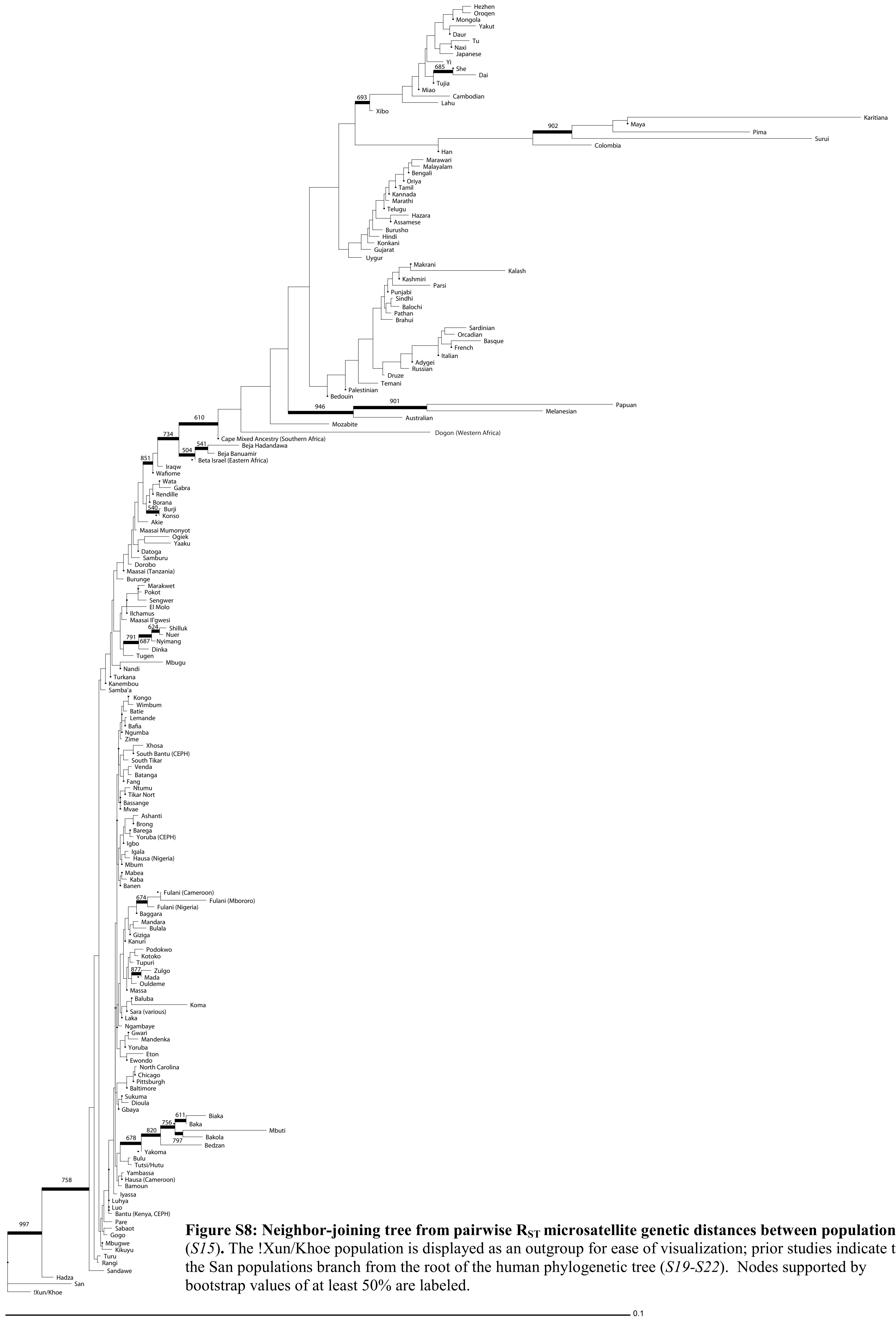


Figure S8: Neighbor-joining tree from pairwise R_{ST} microsatellite genetic distances between populations (*S15*). The !Xun/Khoe population is displayed as an outgroup for ease of visualization; prior studies indicate that the San populations branch from the root of the human phylogenetic tree (*S19*-*S22*). Nodes supported by bootstrap values of at least 50% are labeled.

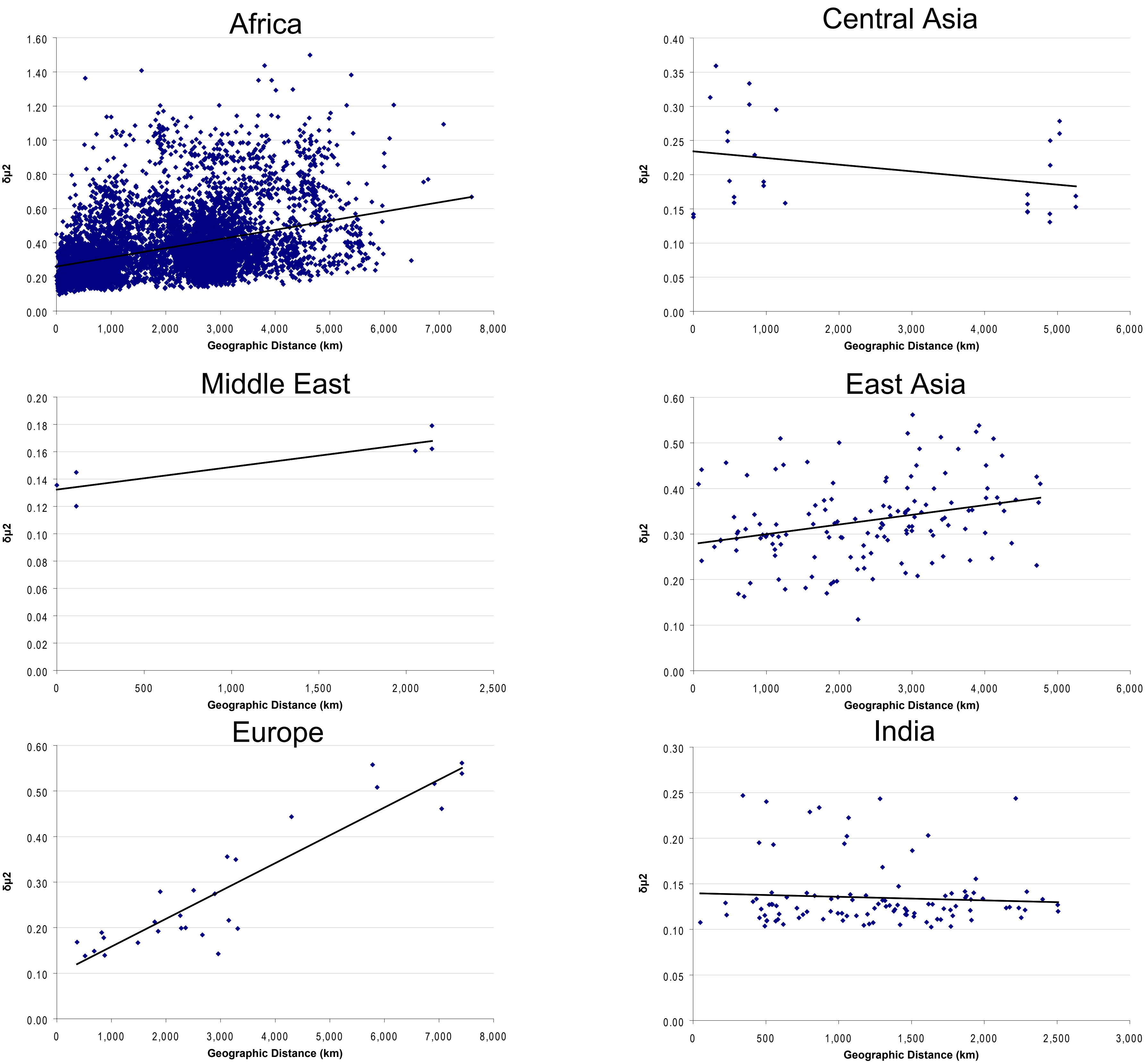


Figure S9: Geographic distance between populations (km) versus pairwise $(\delta\mu)^2$ genetic distances between populations for each major geographic region. We observe a linear correlation between geographic and genetic distance as expected under an isolation by distance model. The largest correlations (Spearman ρ) were observed in Europe (ρ= 0.83), the Middle East (ρ= 0.88) and in Africa (ρ= 0.40) (**Table S5**).

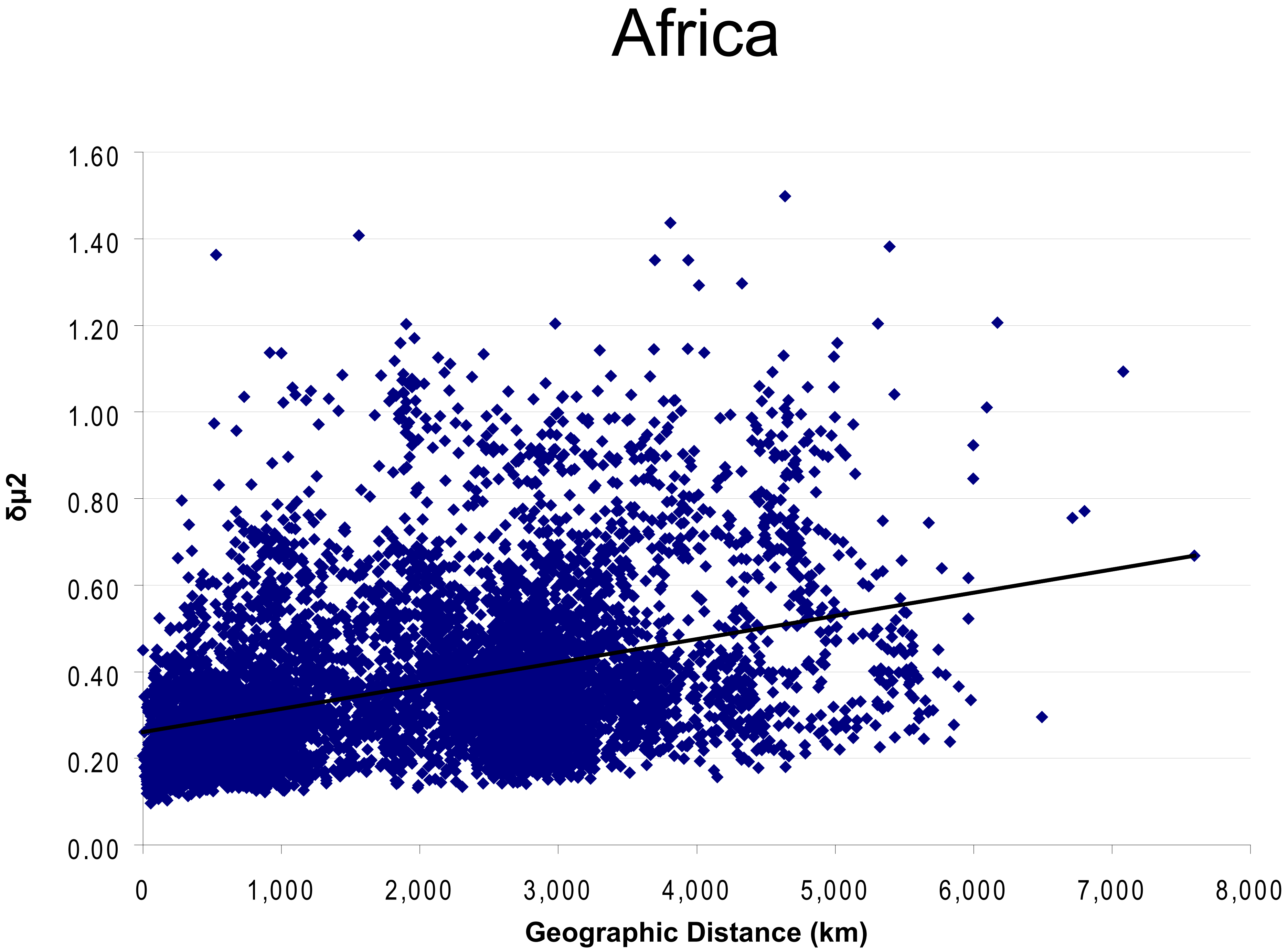


Figure S10: Great circle geographic distance (km) versus pairwise genetic distances $(\delta\mu)^2$ between populations within Africa. Generally speaking, we observe a linear correlation between geographic and genetic distance as expected under an isolation by distance model

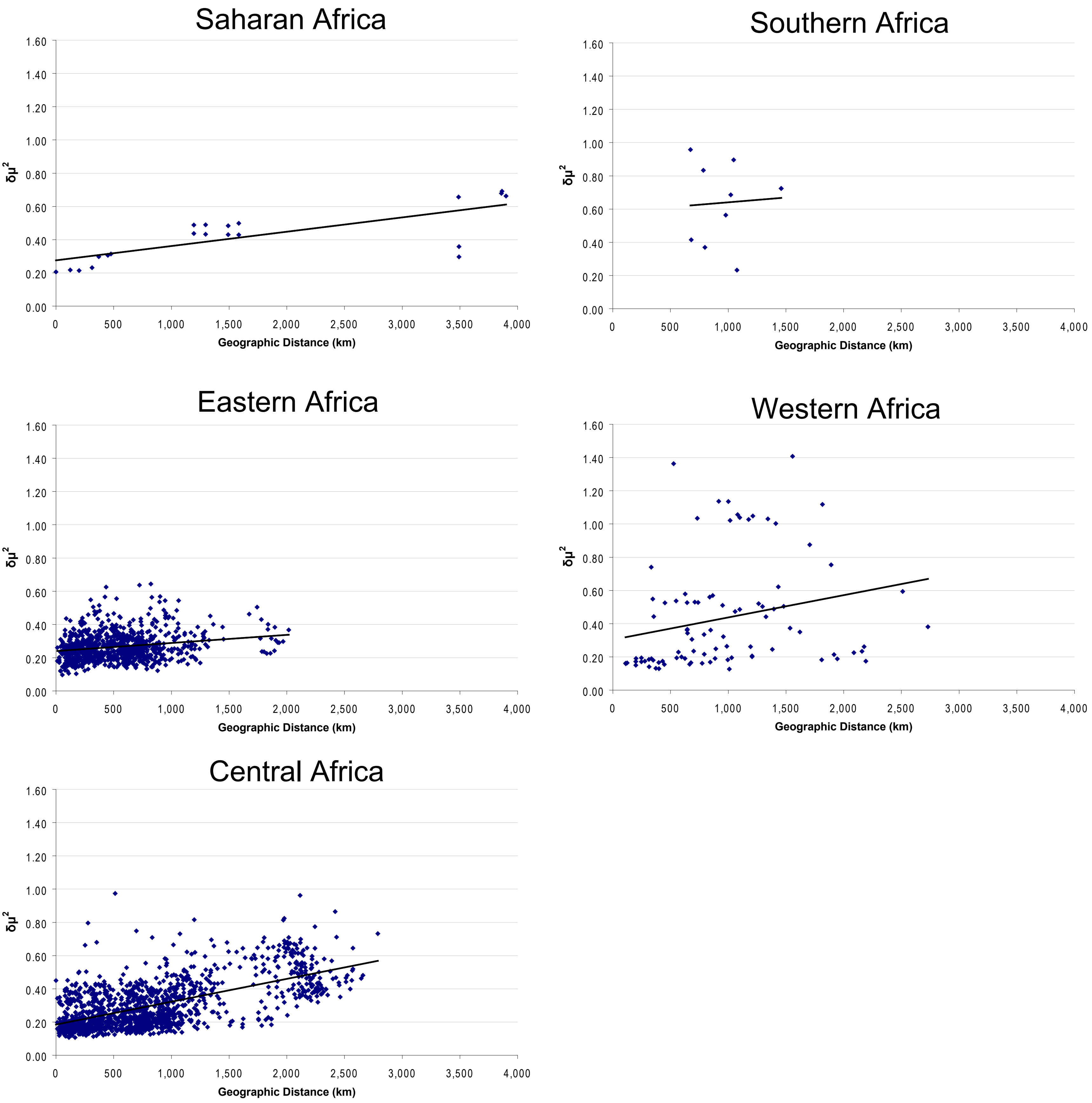


Figure S11: Great circle geographic distance (km) versus pairwise genetic distances $(\delta\mu)^2$ between populations within each major geographic region of Africa. The largest correlations (Spearman ρ) were observed in Saharan Africa (ρ= 0.76) and the smallest in East Africa (ρ= 0.19) (**Table S4**).

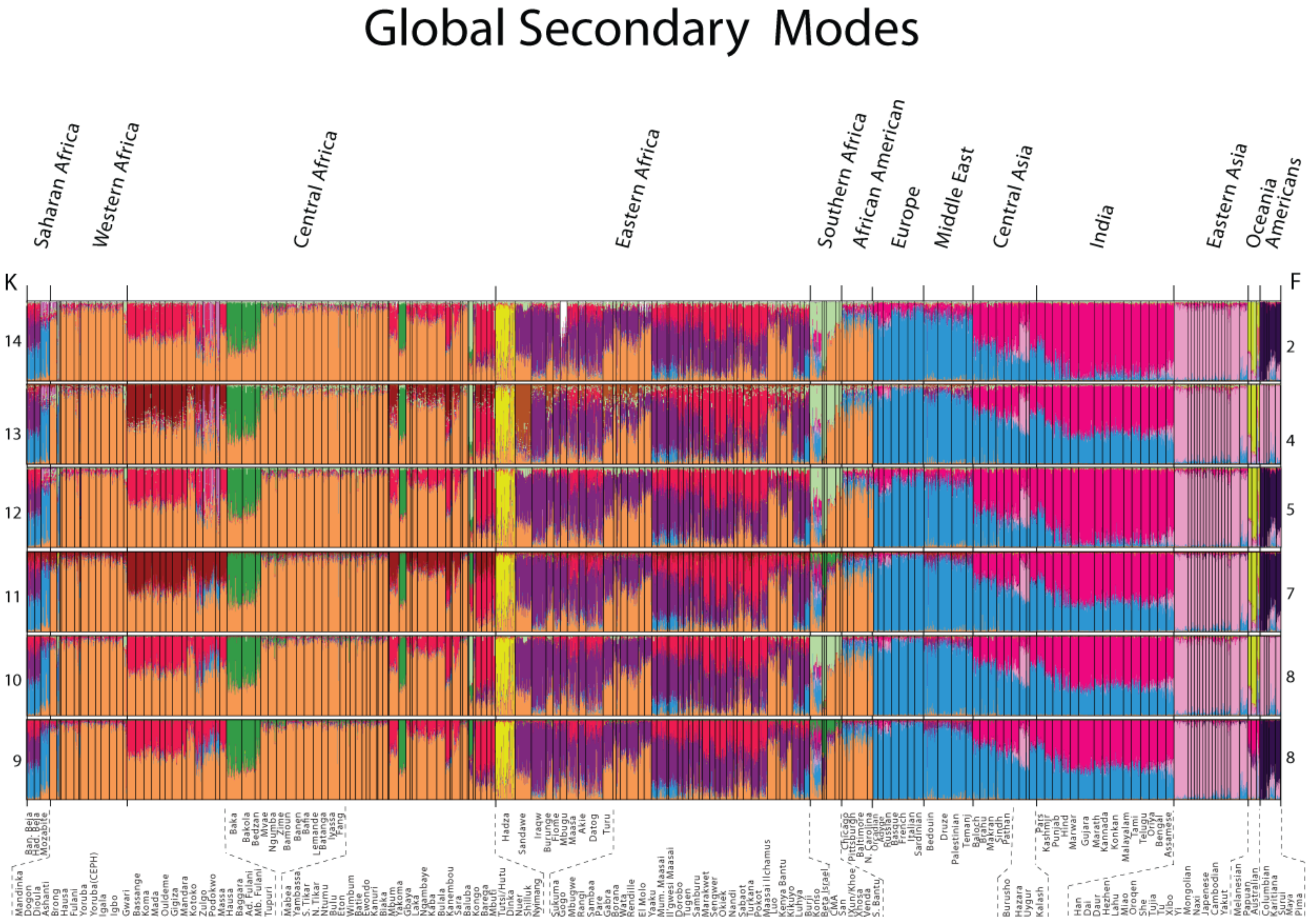


Figure S12: Secondary Modes for STRUCTURE analysis (*29*) for K = 9 – 14 for the global population dataset.

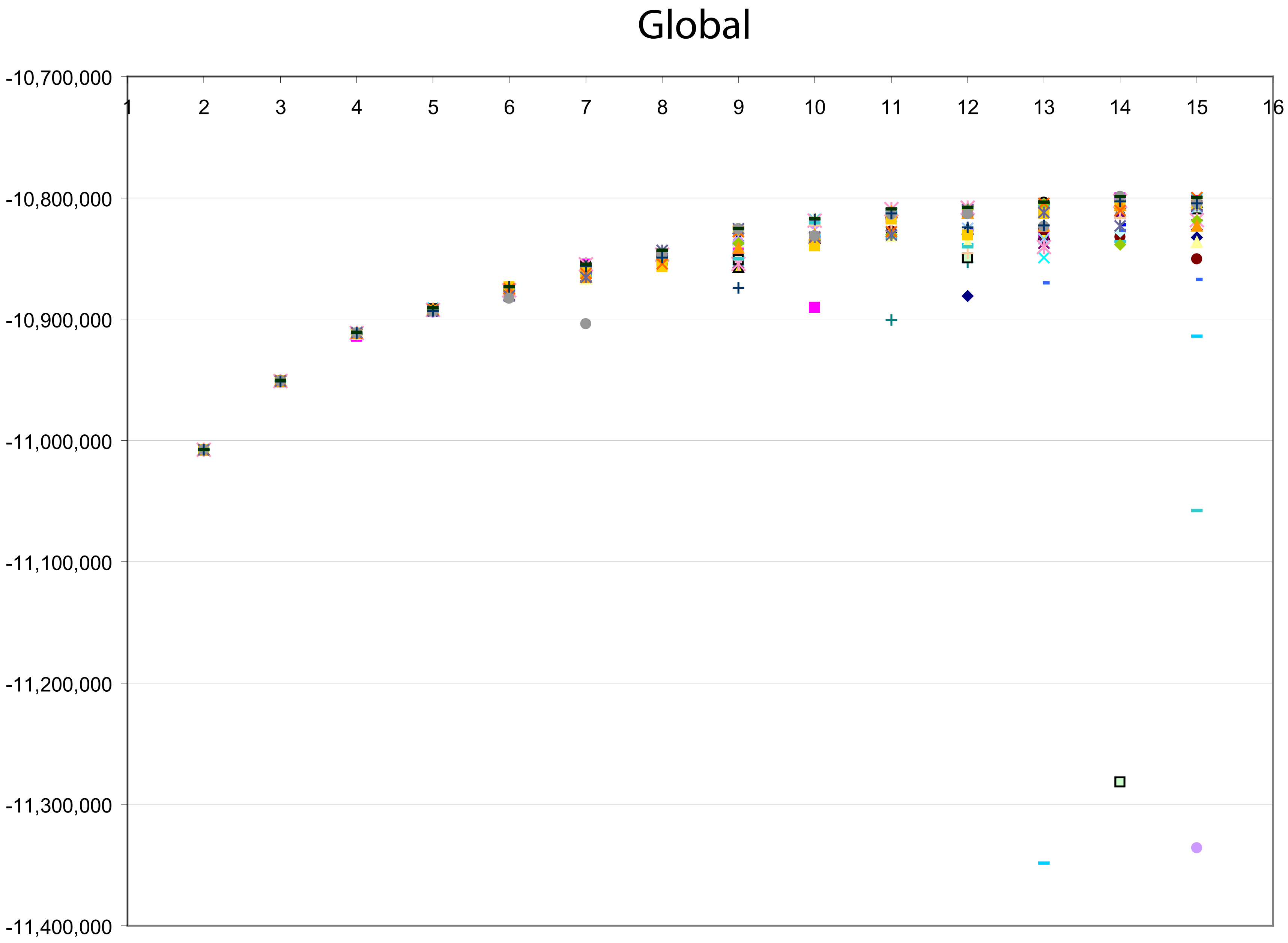


Figure S13: Likelihood scores for 25 STRUCTURE runs for the global set of samples for K = 2 - 15

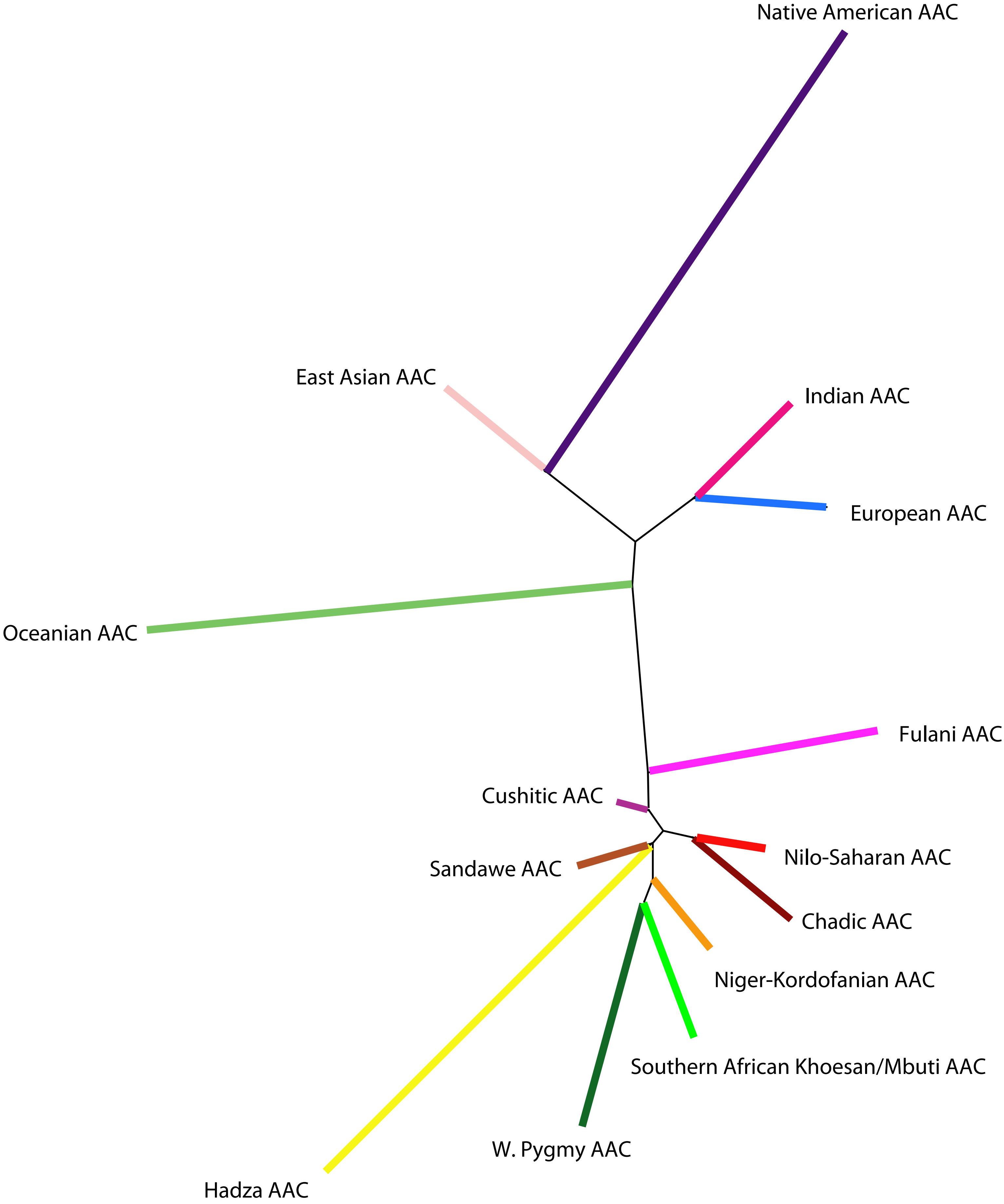


Figure S14: Un-rooted neighbor-joining tree based on pairwise nucleotide genetic distances using inferred ancestral allele frequencies from the global STRUCTURE analysis at K = 14.

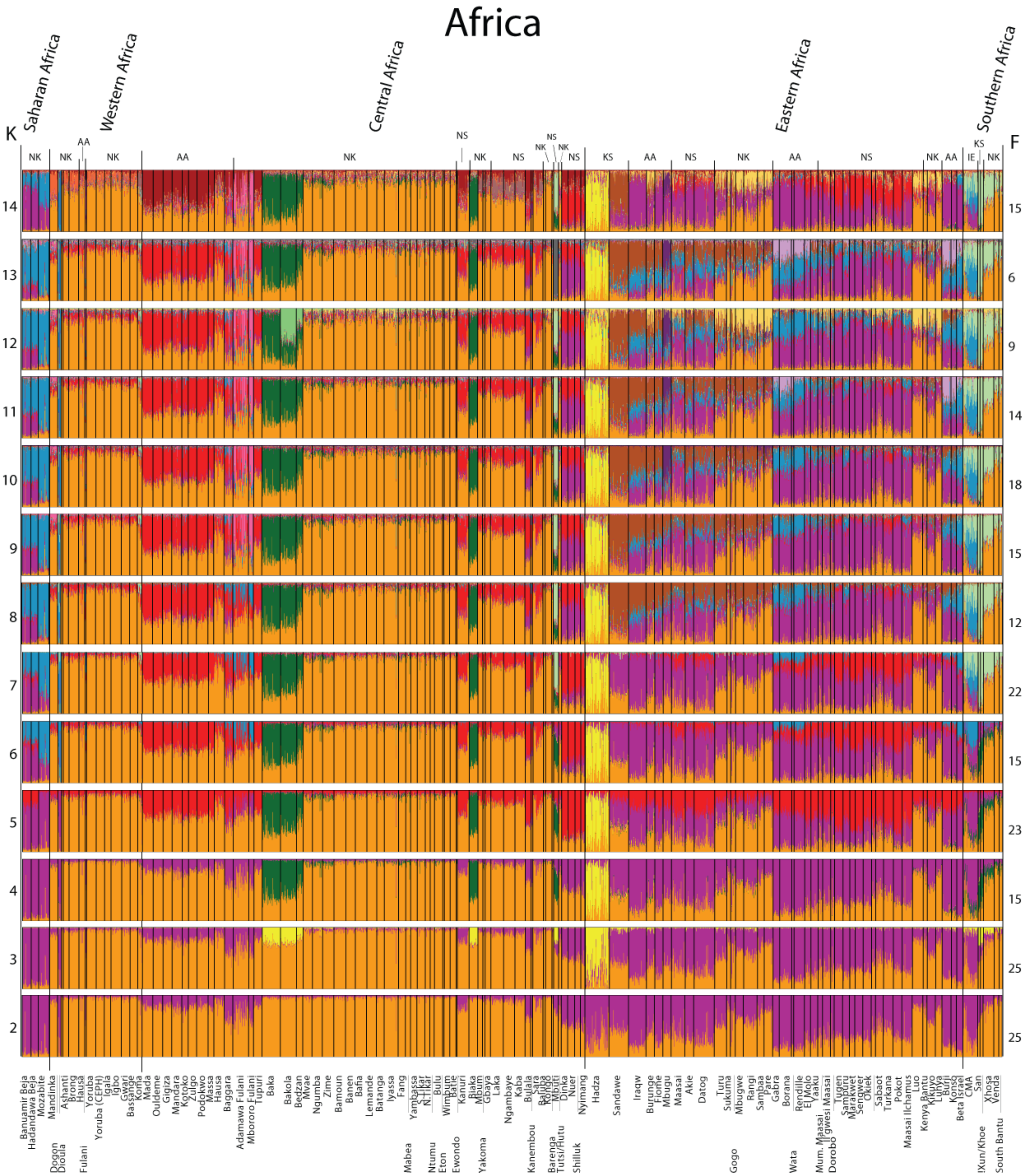


Figure S15: STRUCTURE analysis of the African dataset only (121 populations) with all genetic data (848 microsatellites, 476 indels and 3 SNPs) from K = 2 - 14. Each vertical line represents an individual. Individuals are clustered by self-identified ethnic group (shown at bottom) and ethnic groups are clustered by major geographic region. The colors represent the proportion of inferred ancestry from K ancestral populations. Values for K are shown on the left and the number of similar runs (F) for the primary mode for each set of 25 STRUCTURE runs at each K value is shown on the right.

Africa Secondary Modes

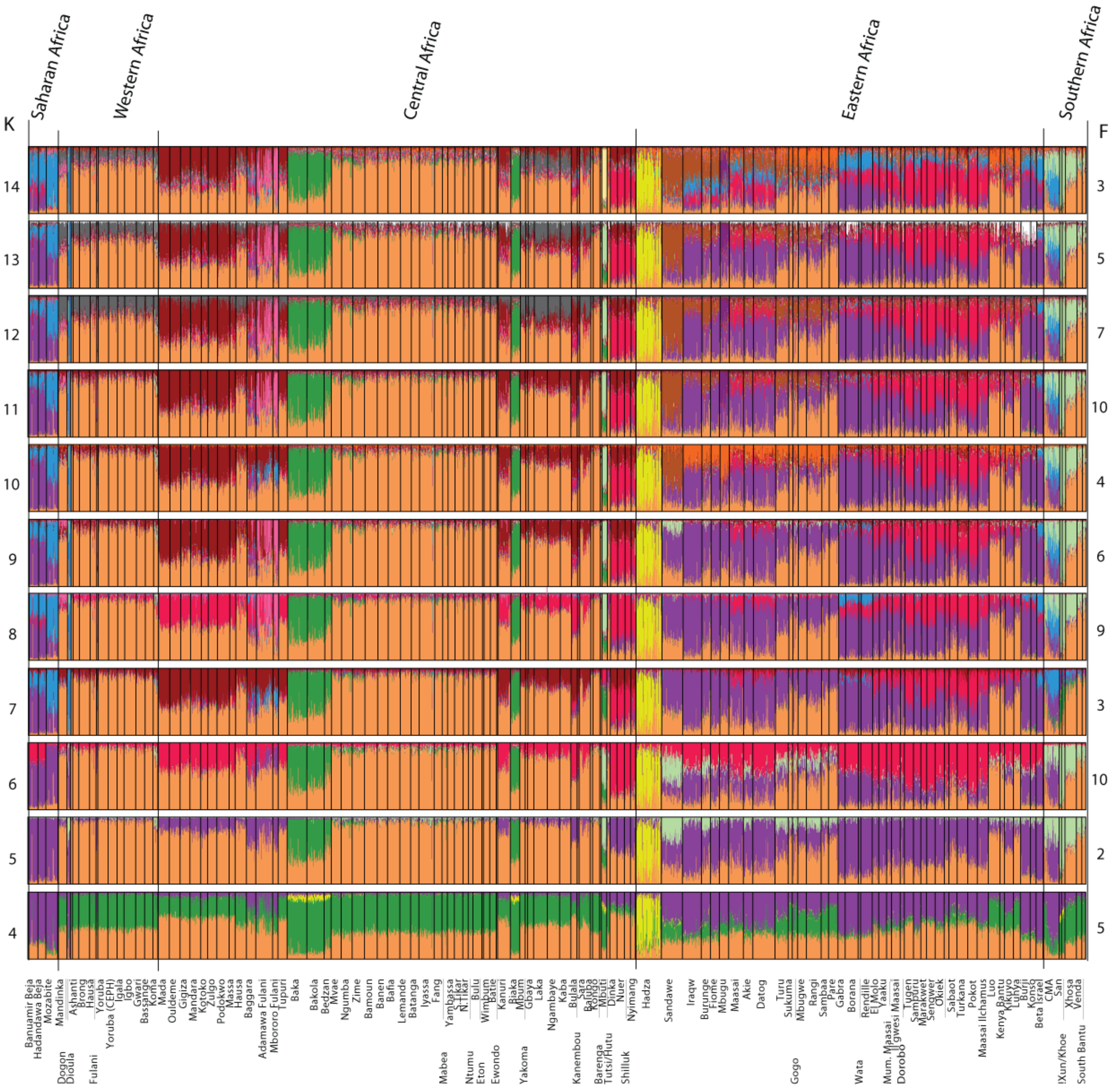


Figure S16: Secondary Modes for STRUCTURE for K = 4 – 14 for the African population dataset. Values for K are shown on the left and the number of similar runs (F) for the primary mode for each set of 25 STRUCTURE runs at each K value is shown on the right.

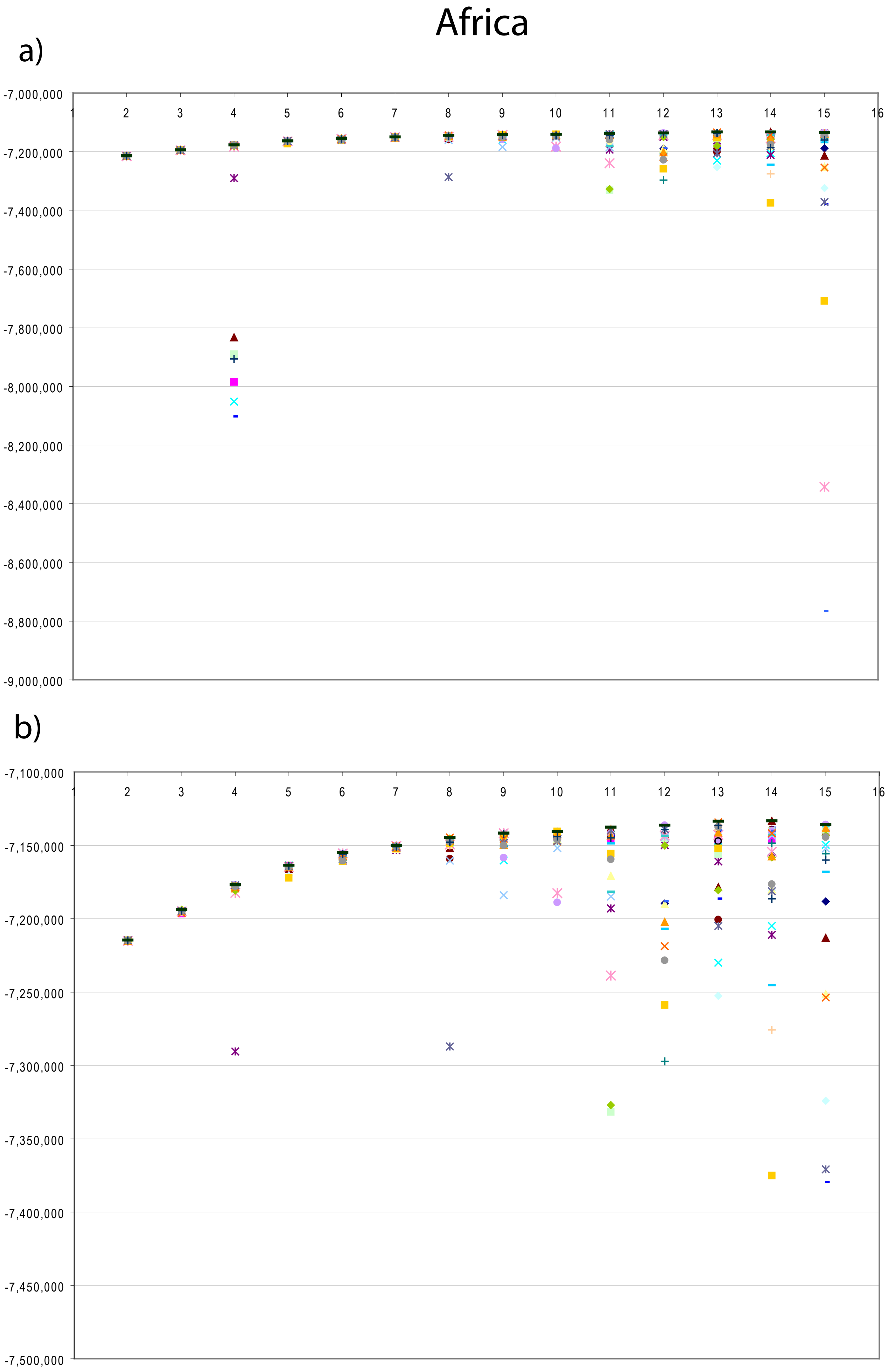


Figure S17: Likelihood scores for 25 STRUCTURE runs for African populations for K = 2 - 15. (a) results for all runs (b) results after removing outliers and rescaling the Y axis.

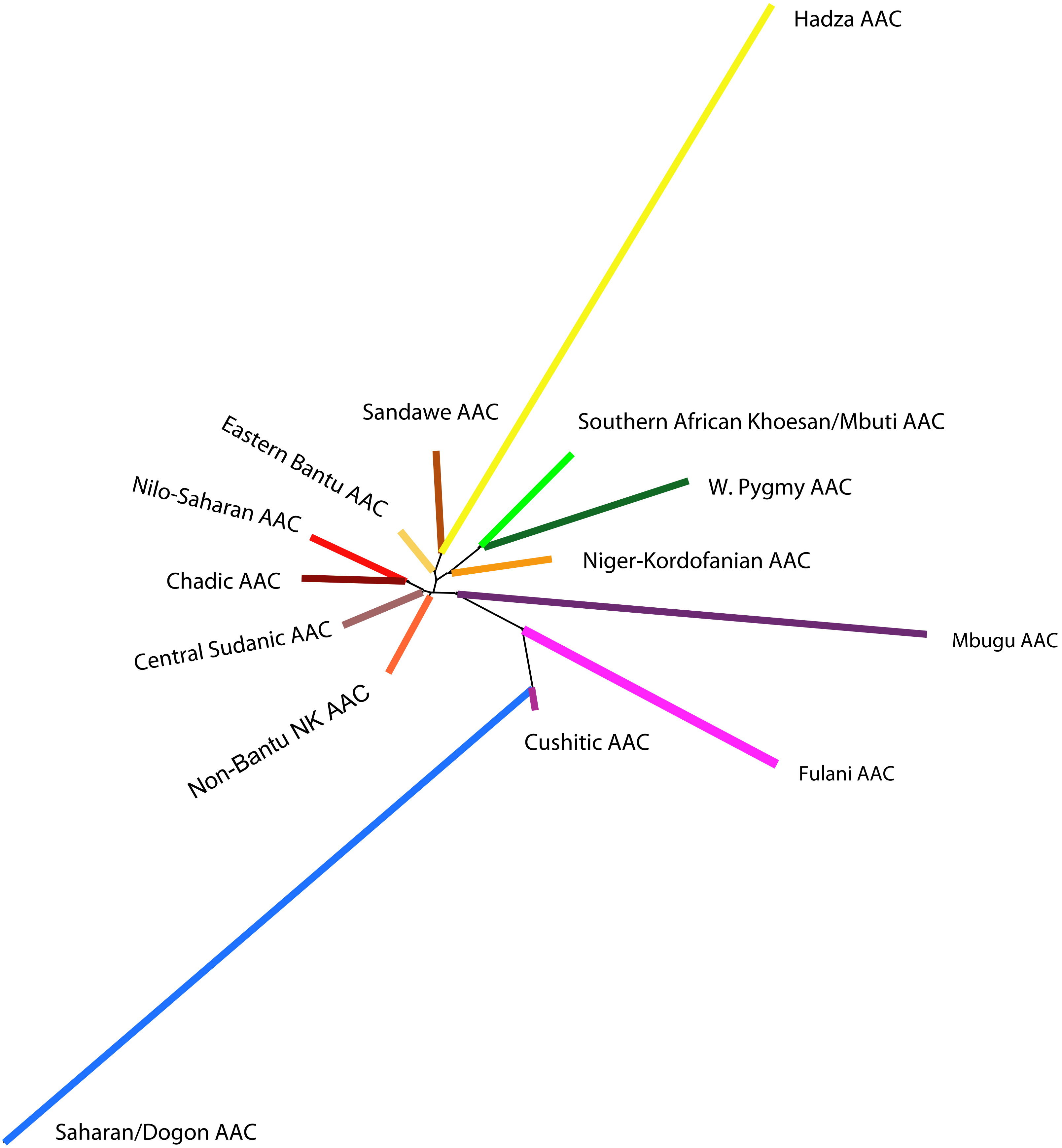


Figure S18: Un-rooted neighbor-joining tree from pairwise net nucleotide genetic distances calculated from the inferred ancestral allele frequencies at K = 14 from STRUCTURE analysis of the African dataset. Major clades observed in the tree include the North African/Dogon, Fulani, Cushitic, and Mbugu AACs, the SAK-Mbuti Pygmy and W. Pygmy AACs, and the Hadza and Sandawe AACs. The Nilo-Saharan and Chadic-Saharan AAC form a sister group and cluster with the Central Sudanic AAC. The Bantu AACs radiate from the center of the tree

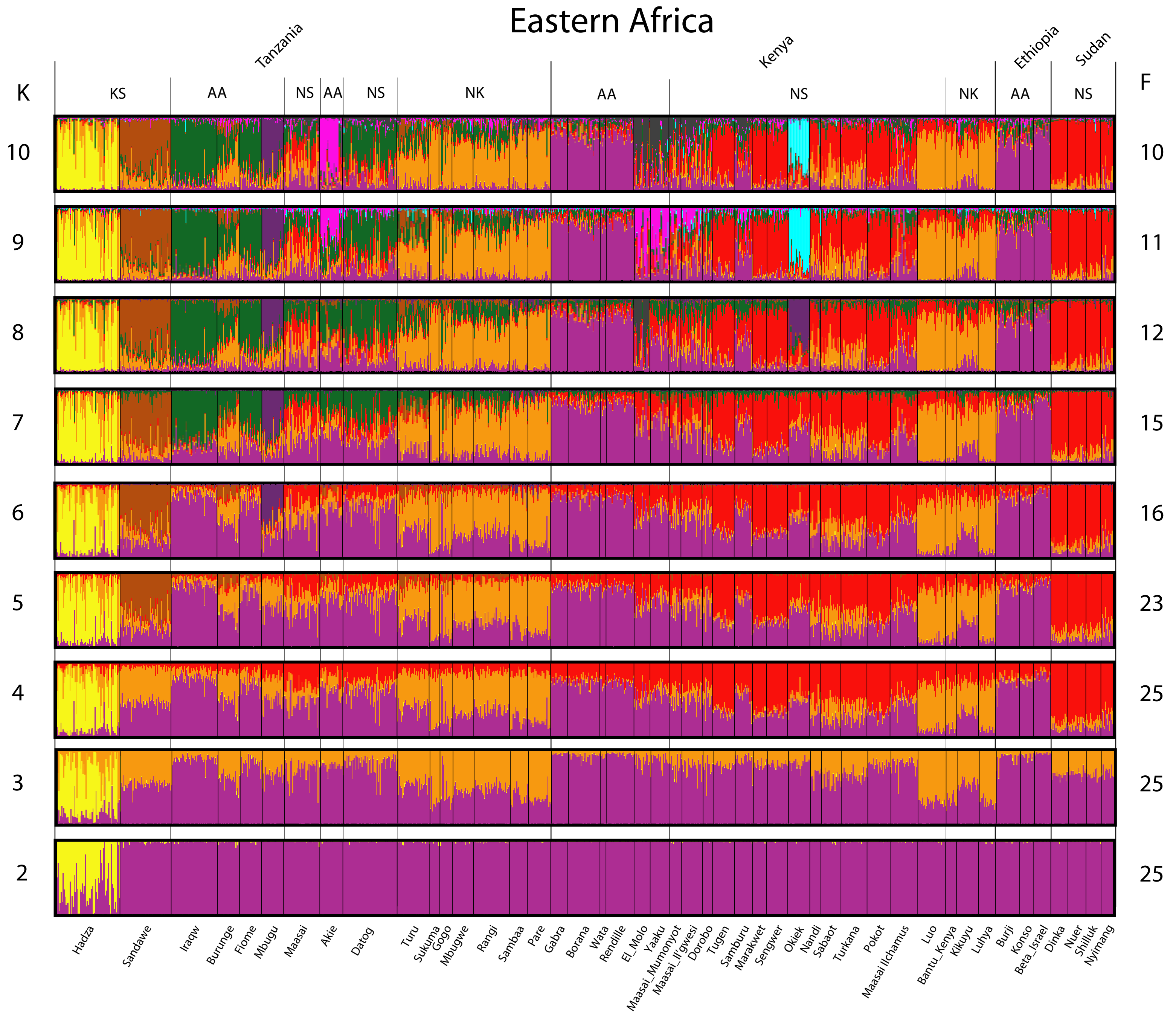


Figure S19: Results of STRUCTURE analysis for Eastern Africa at K = 2 -10. Values for K are shown on the left and the number of similar runs (F) for the primary mode for each set of 25 STRUCTURE runs at each K value is shown on the right.

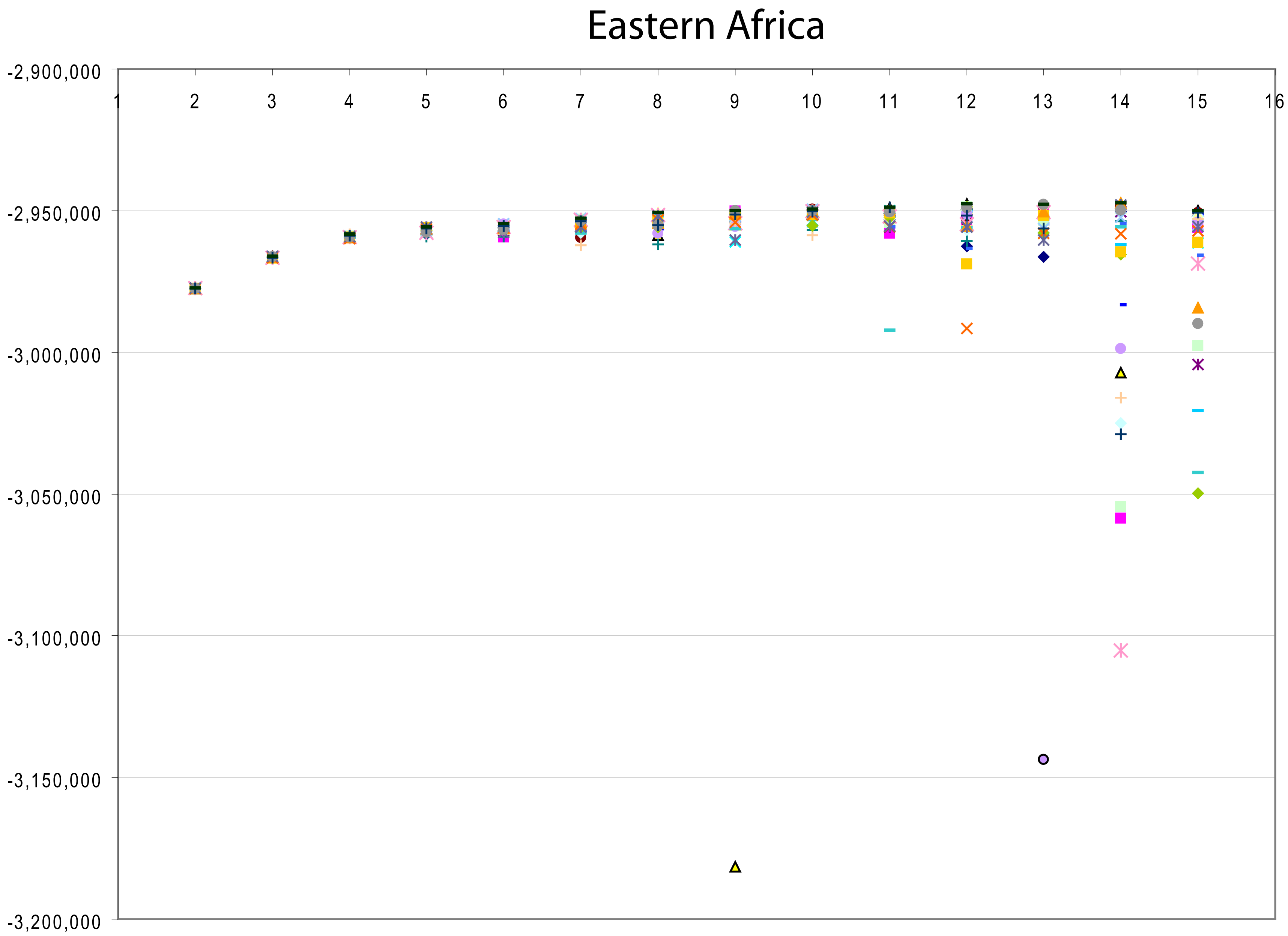


Figure S20: Likelihood scores for 25 STRUCTURE runs for Eastern African populations for K = 2 - 15.

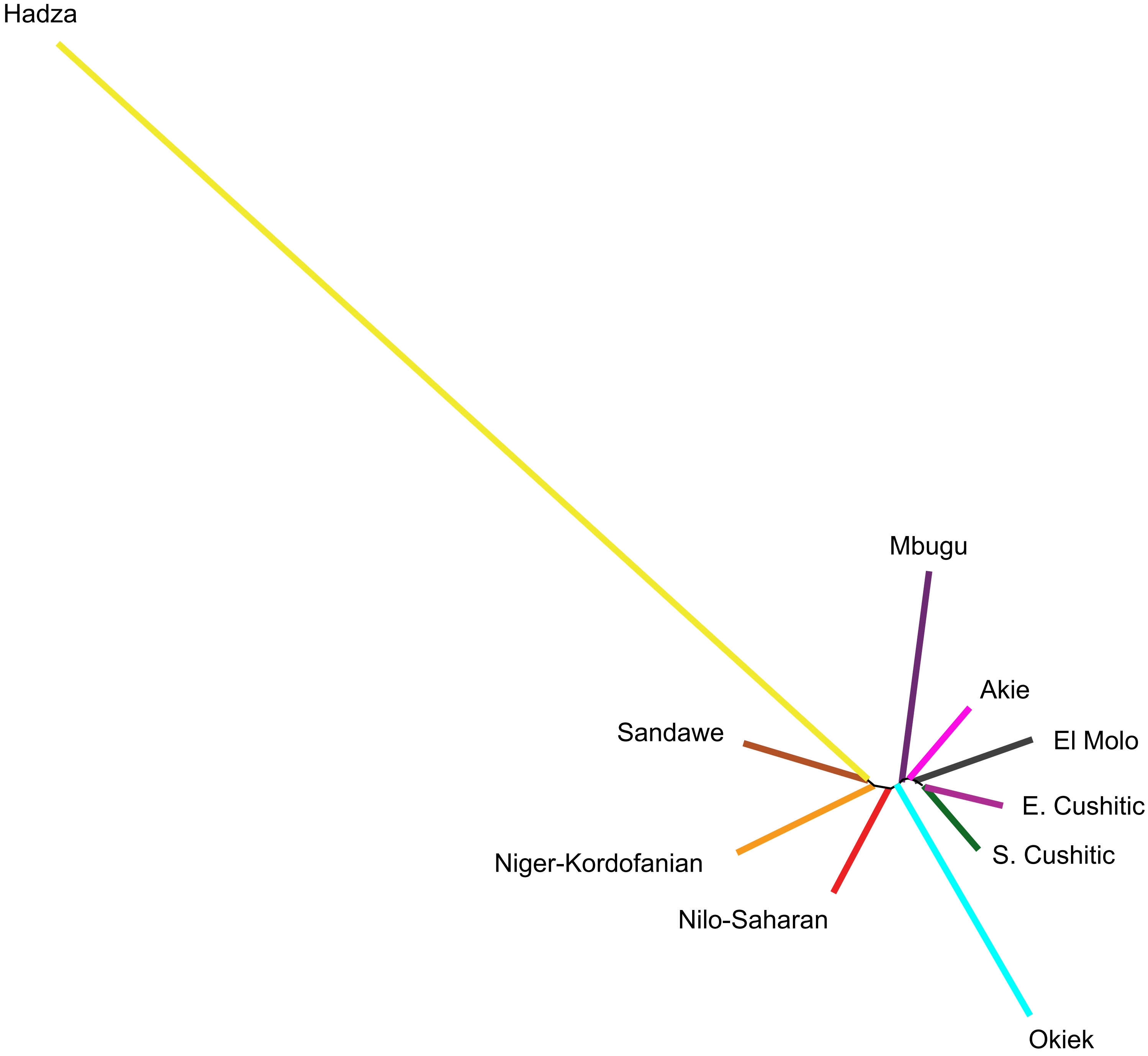


Figure S21: Un-rooted neighbor-joining tree from pairwise net nucleotide genetic distances calculated from the inferred ancestral allele frequencies for clusters at K = 10 with STRUCTURE analysis of the Eastern African dataset

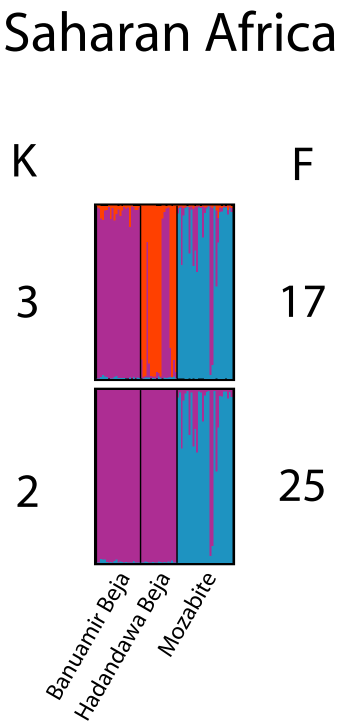


Figure S22: Results of STRUCTURE analysis for Saharan Africa at K = 2 -3. Values for K are shown on the left and the number of similar runs (F) for the primary mode for each set of 25 STRUCTURE runs at each K value is shown on the right.

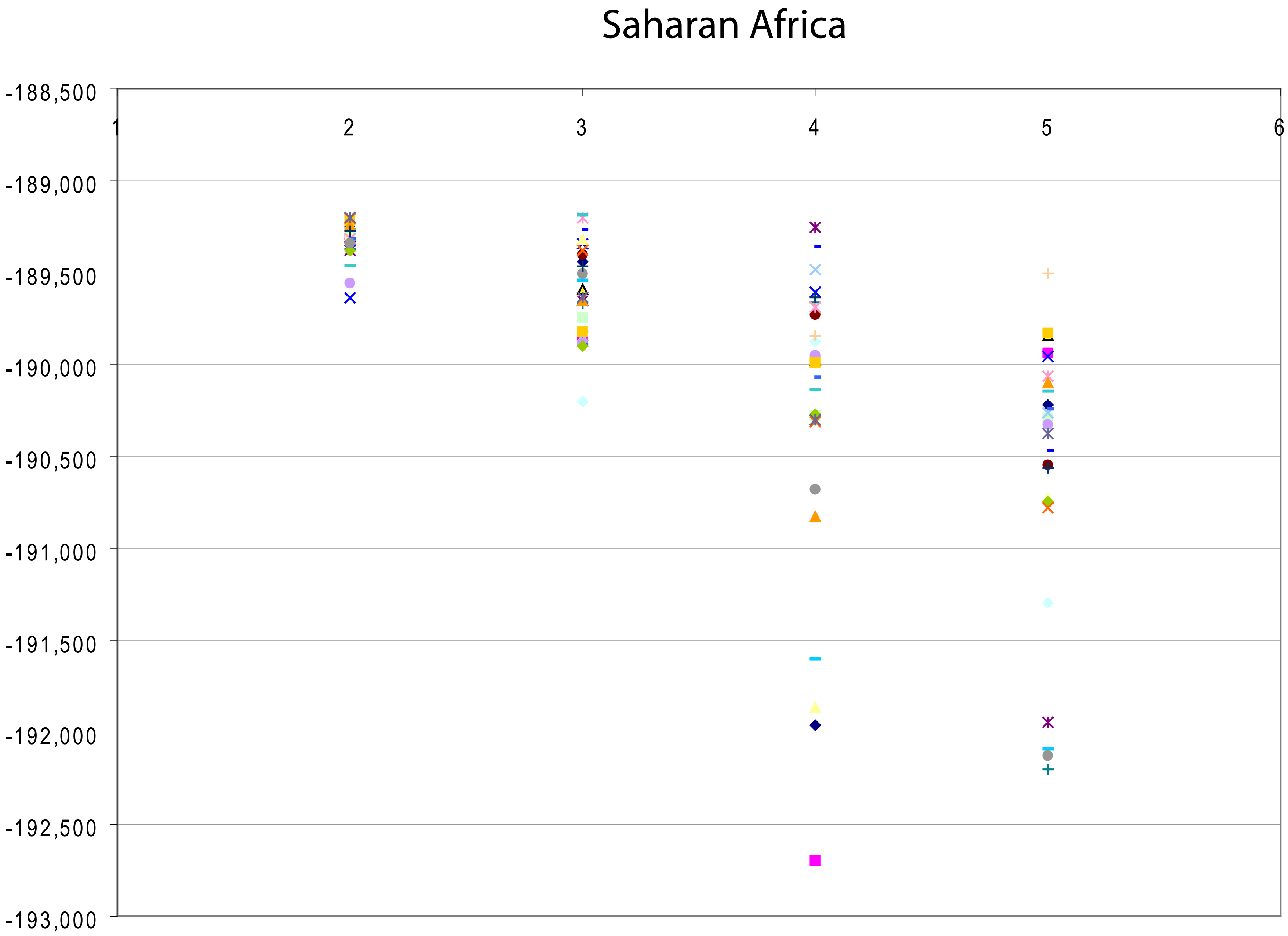


Figure S23: Likelihood scores for 25 STRUCTURE runs for Saharan African populations for K = 2 - 5.

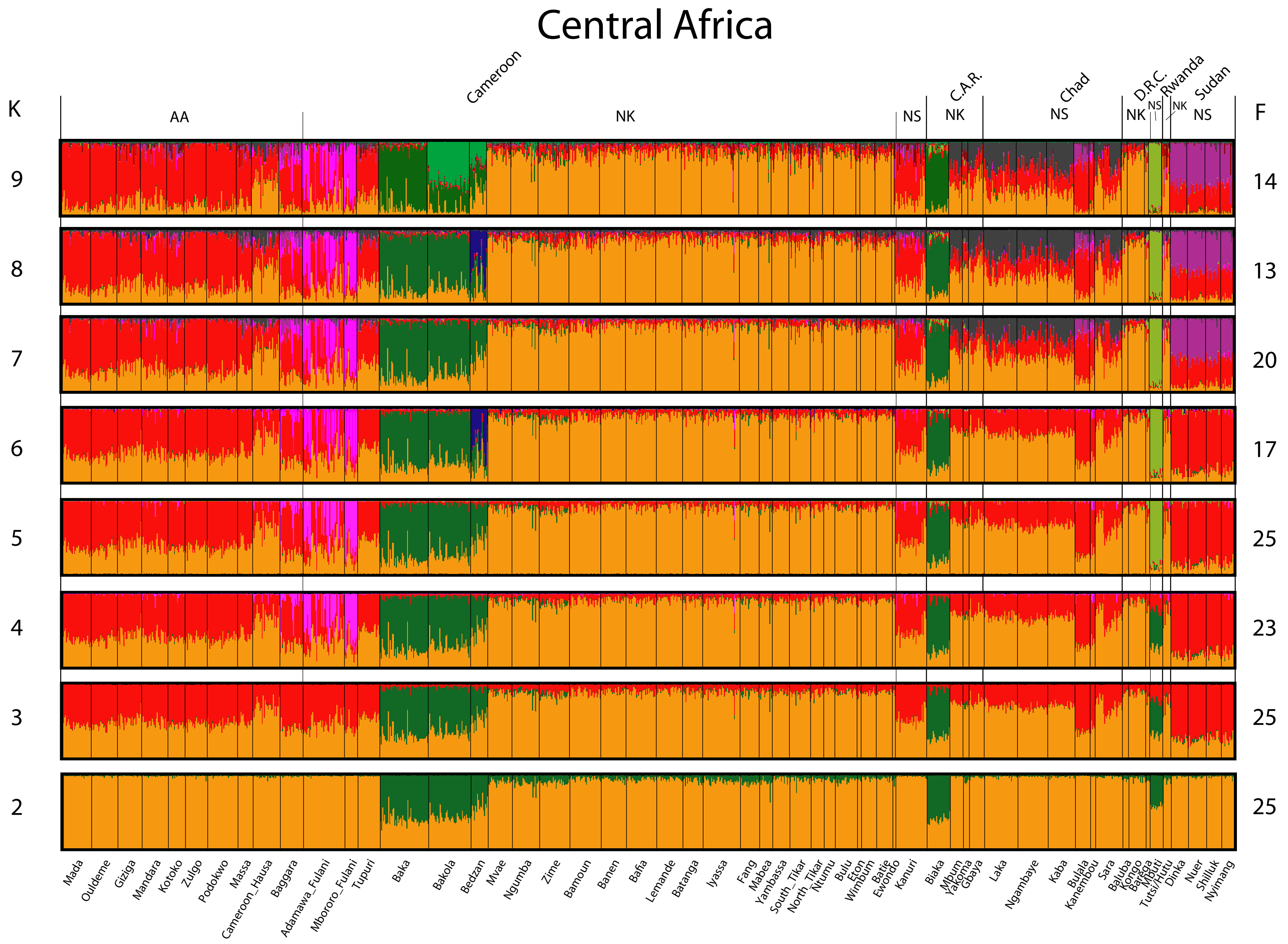


Figure S24: Results of STRUCTURE analysis for Central Africa at K = 2 -9. Note that the southern Sudanese are included in these analyses because of shared ancestral clusters as inferred on the basis of analyses of the global and African datasets. Values for K are shown on the left and the number of similar runs (F) for the primary mode for each set of 25 STRUCTURE runs at each K value is shown on the right.

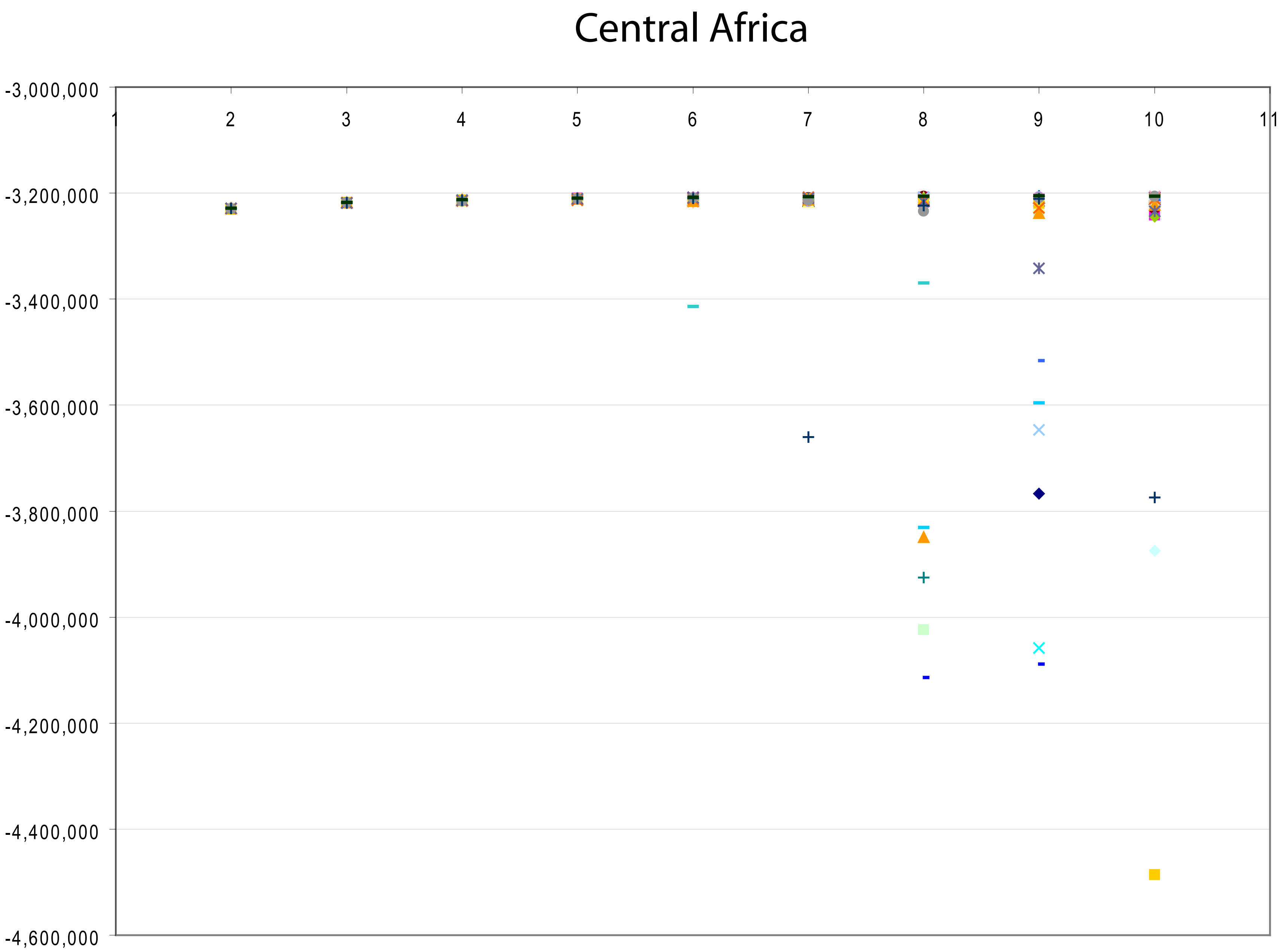


Figure S25: Likelihood scores for 25 STRUCTURE runs for Central African populations for K = 2 - 10.

Western Africa

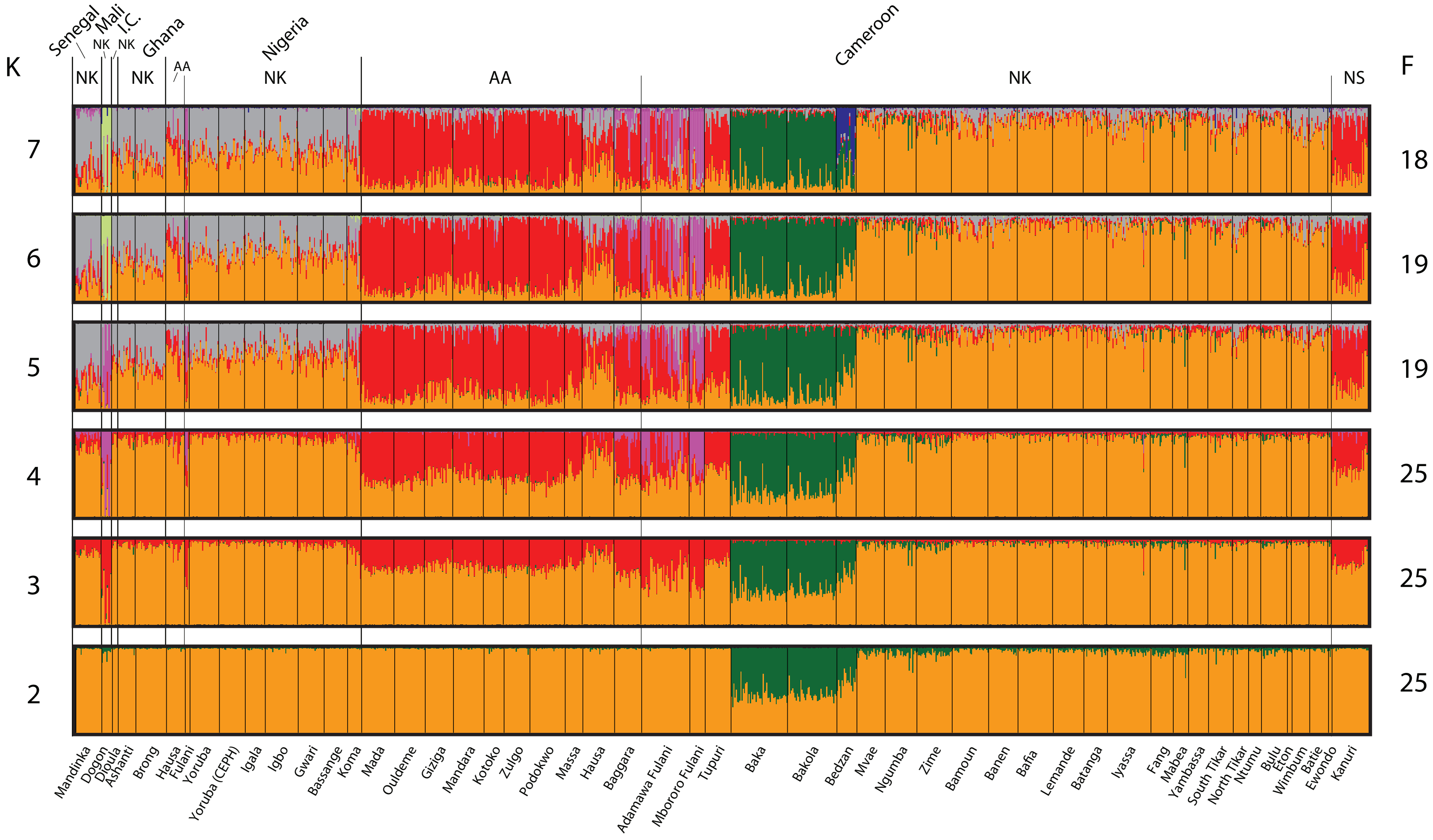


Figure S26: Results of STRUCTURE analysis for Western Africa at K = 2 -7. Note that the southern Cameroon populations are included because of shared ancestral clusters as inferred the basis of analyses of the global and African datasets and to distinguish subtle substructure between Bantu and non-Bantu Niger-Kordofanian speakers. Values for K are shown on the left and the number of similar runs (F) for the primary mode for each set of 25 STRUCTURE runs at each K value is shown on the right.

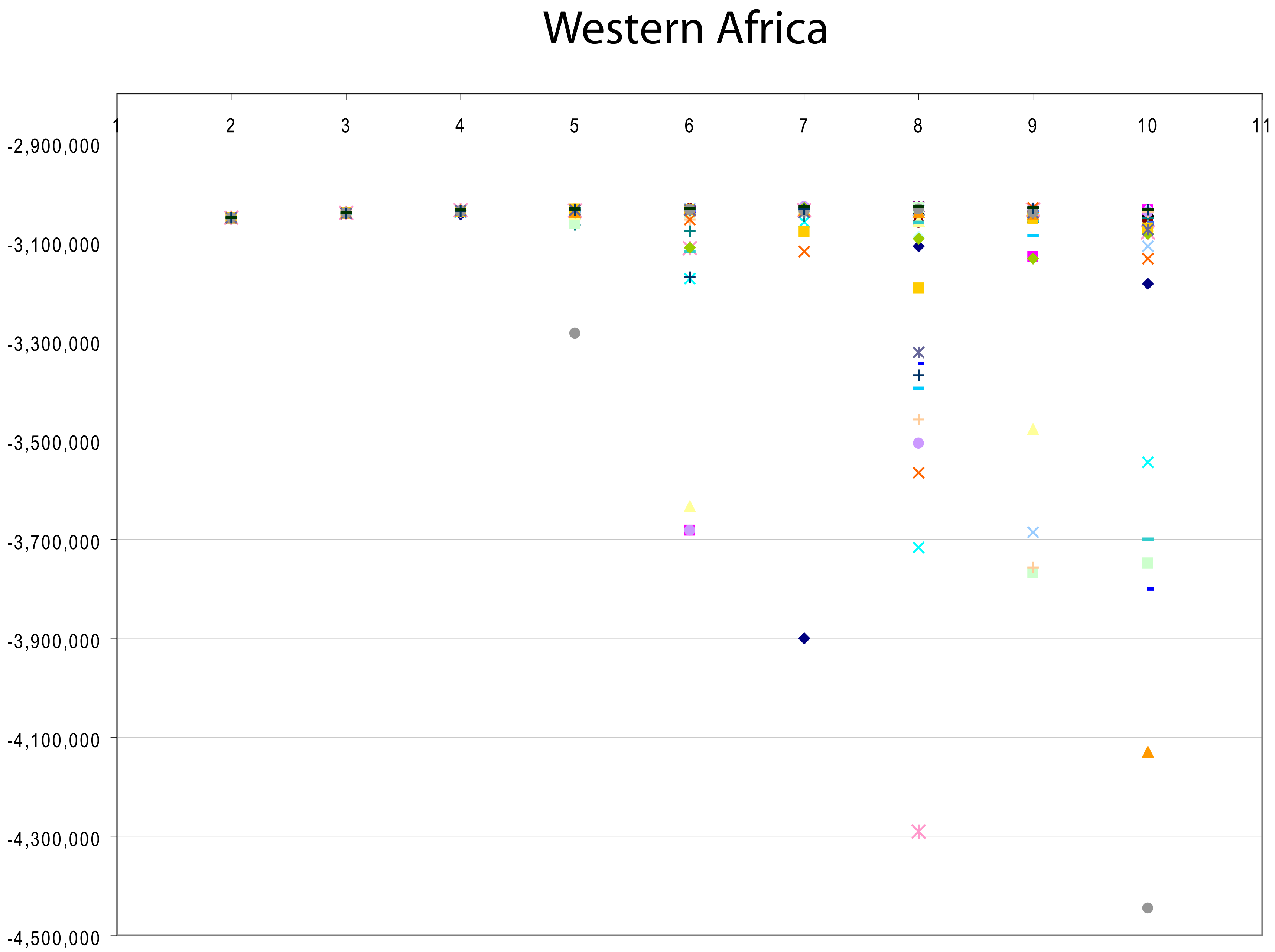


Figure S27: Likelihood scores for 25 STRUCTURE runs for Western African populations for K = 2 - 10.

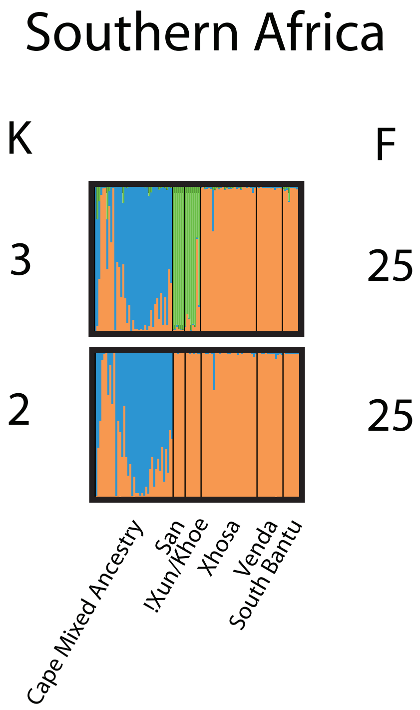


Figure S28: Results of STRUCTURE analysis for Southern Africa at K = 2 -3. Values for K are shown on the left and the number of similar runs (F) for the primary mode for each set of 25 STRUCTURE runs at each K value is shown on the right.

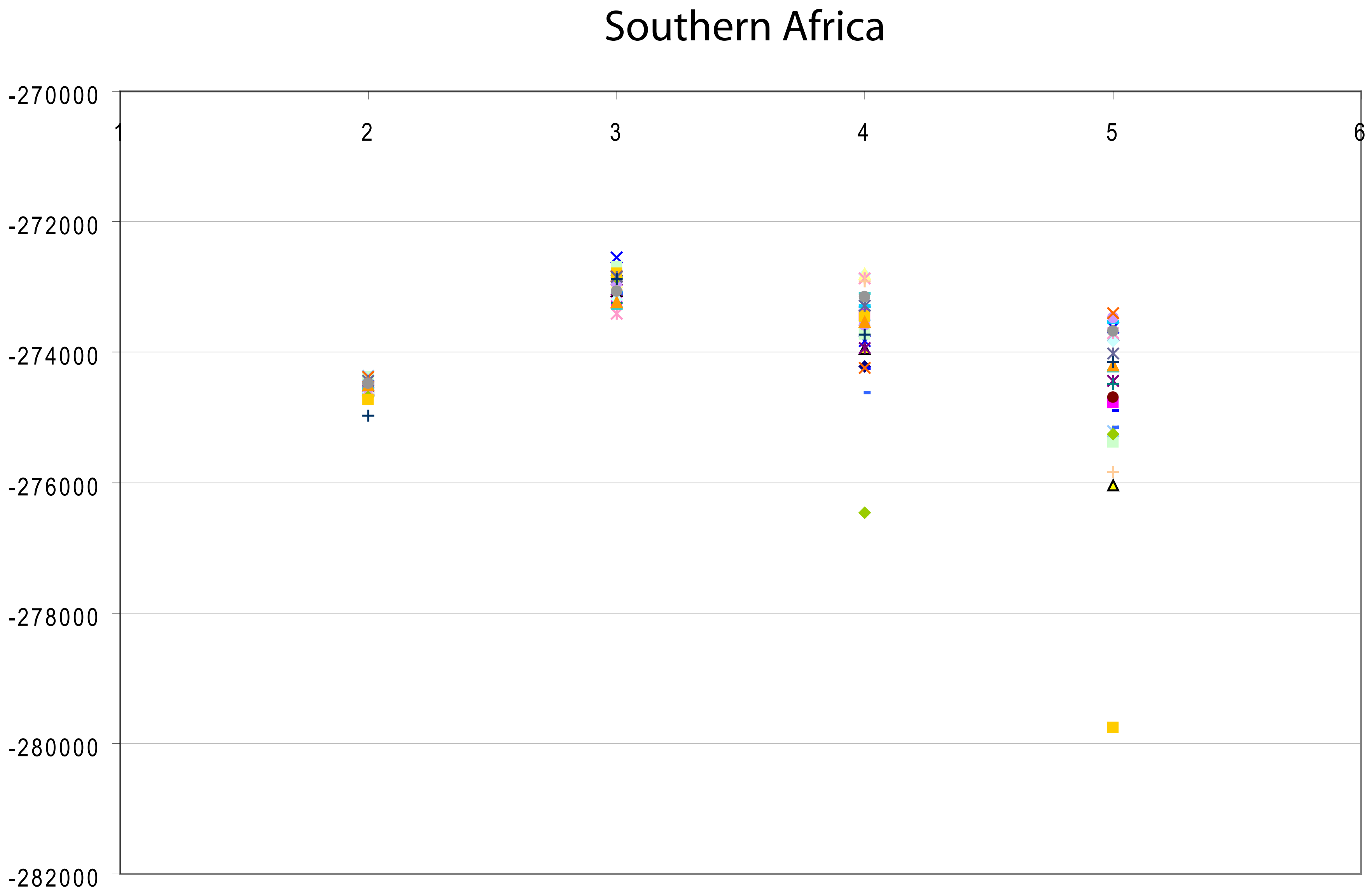


Figure S29: Likelihood scores for 25 STRUCTURE runs for Southern African populations for K = 2 - 5.

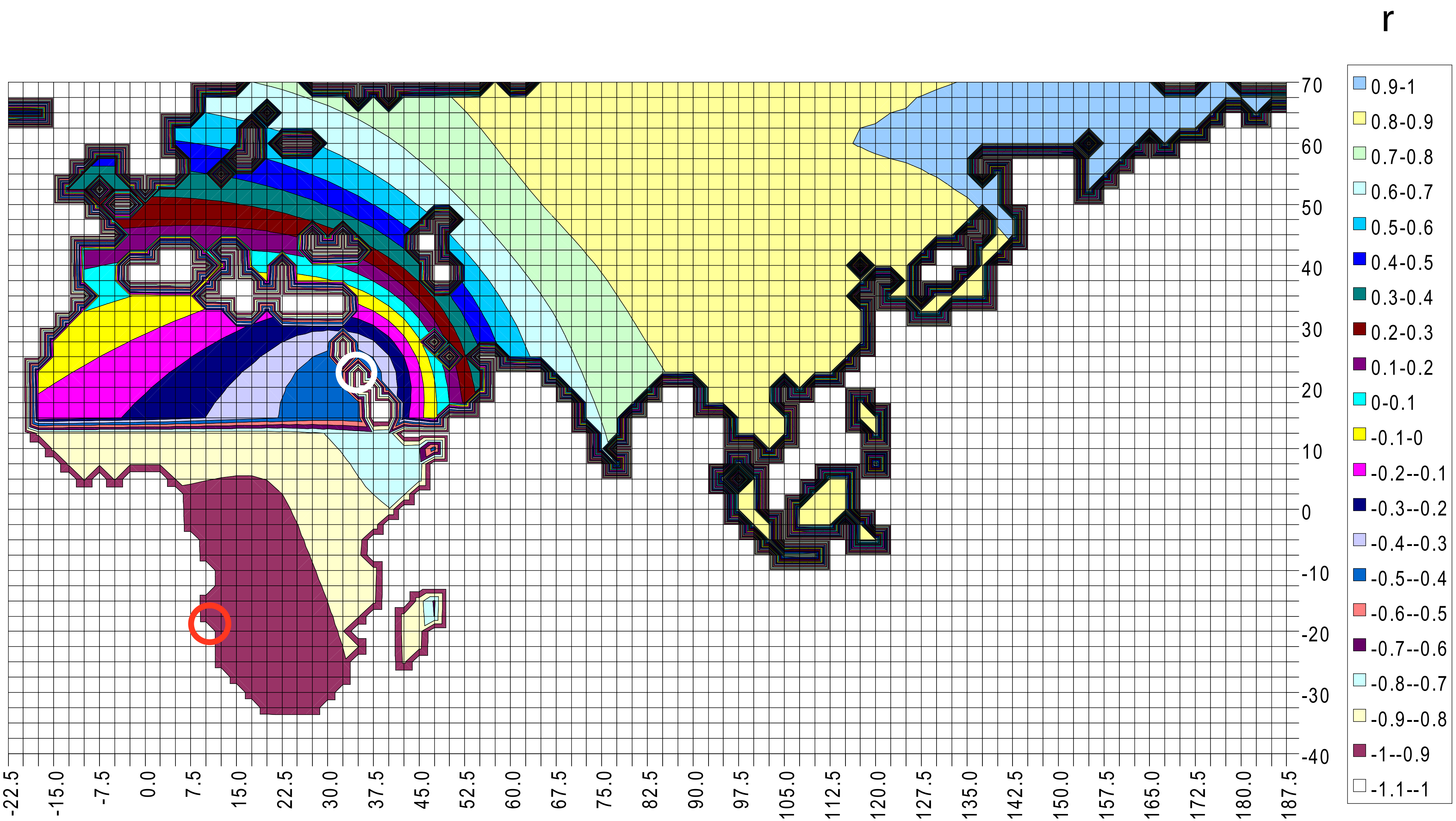


Figure S30: The geographic origin of modern humans and the route of migration out of Africa estimated under a serial founder effect model (*S31*) . This inference was based on the geographic position of sampled populations and the correlation between estimated levels of genetic diversity for each population and estimated distance from the origin (r). The red circle indicates the inferred origin of migration within Africa and white circle indicates the inferred waypoint out of Africa

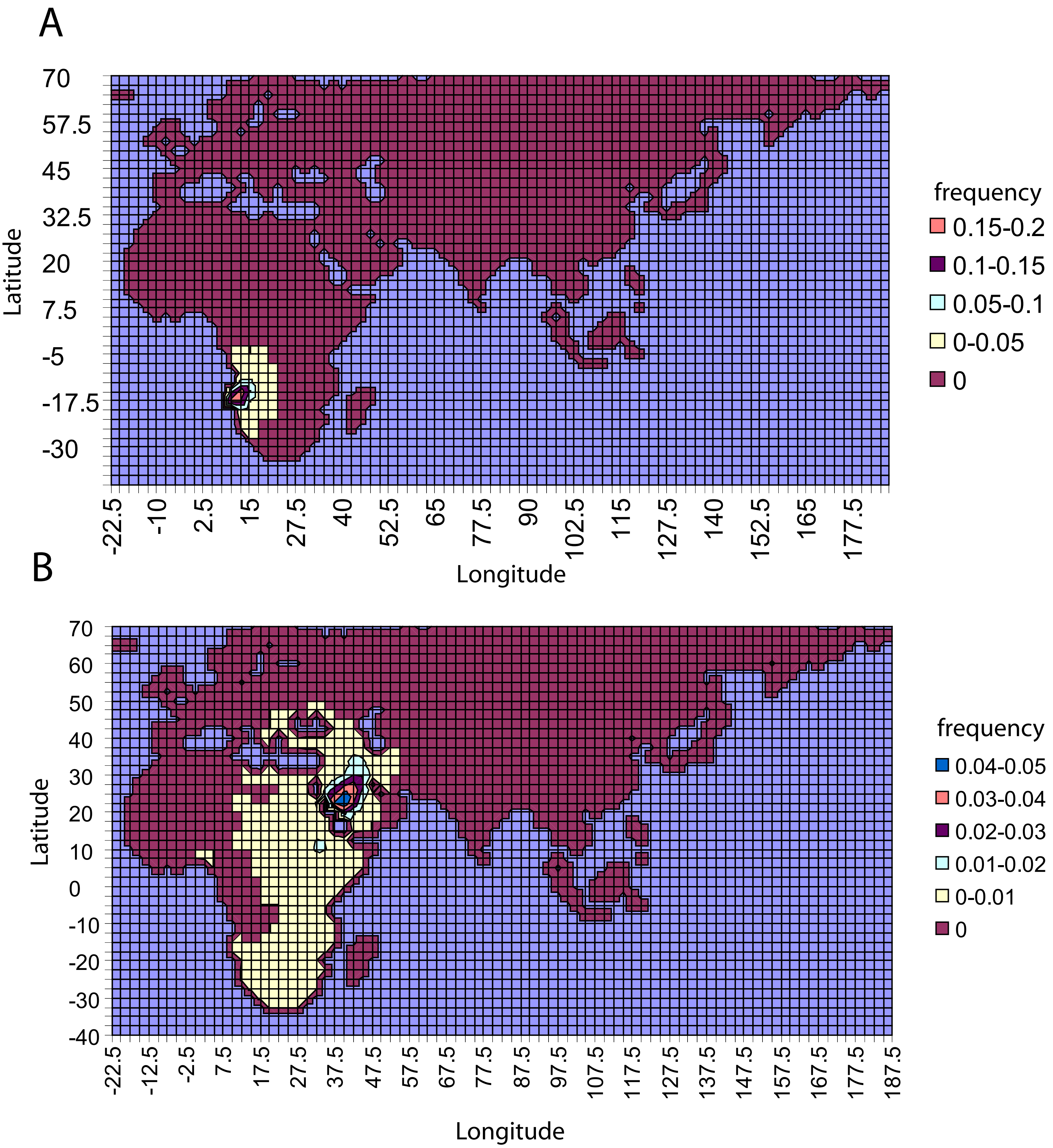


Figure S31: Results of Bootstrap analyses used to infer origin and waypoint of migration of modern humans within and out of Africa. For each grid point the frequency of replicates that fall upon that point is indicated. A) Bootstrap results for origin analysis; B) Bootstrap results for waypoint analysis.

a)

DEEP BRANCHING OF KHOISAN, UNDIFFERENTIATED

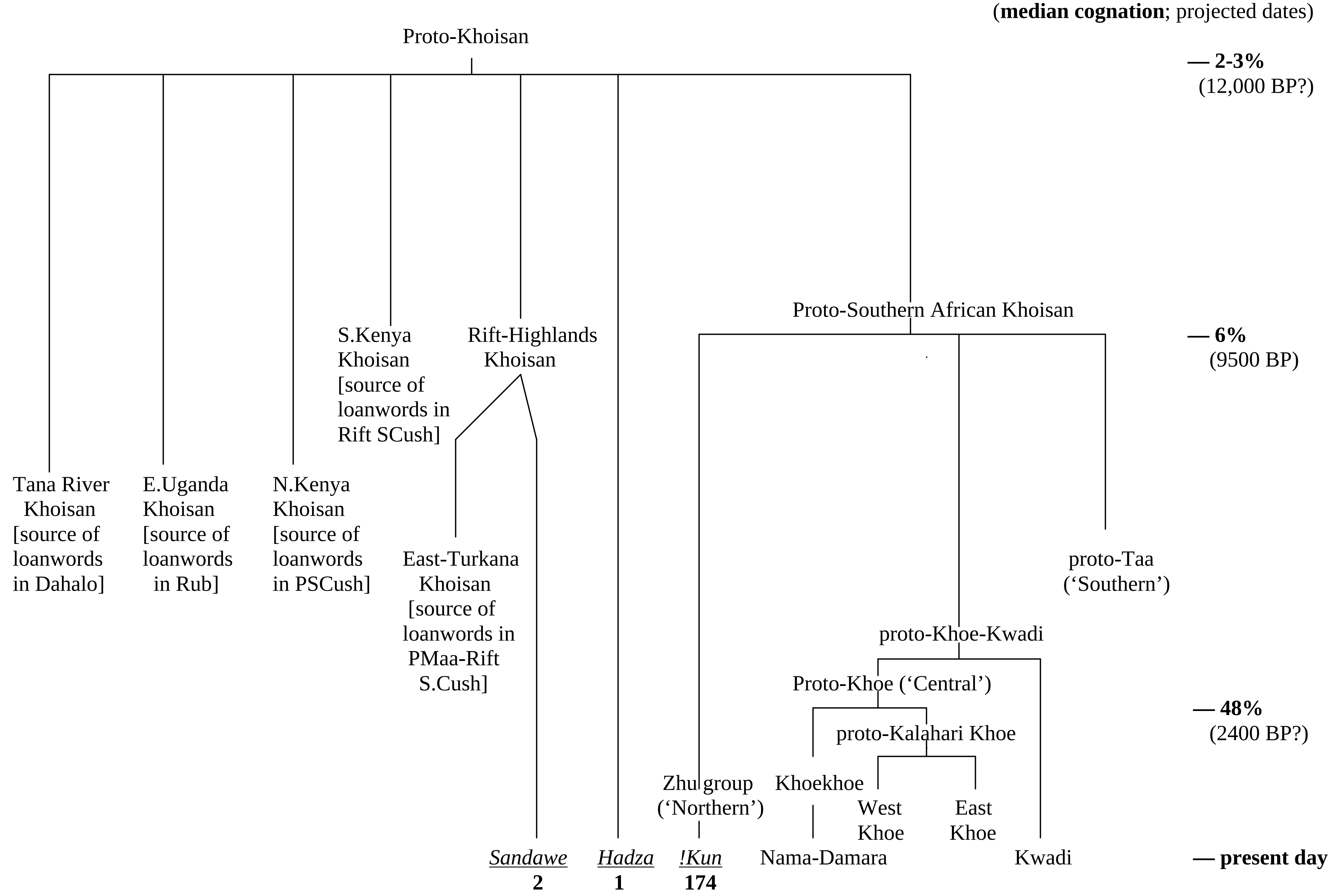


Figure S32a

b)

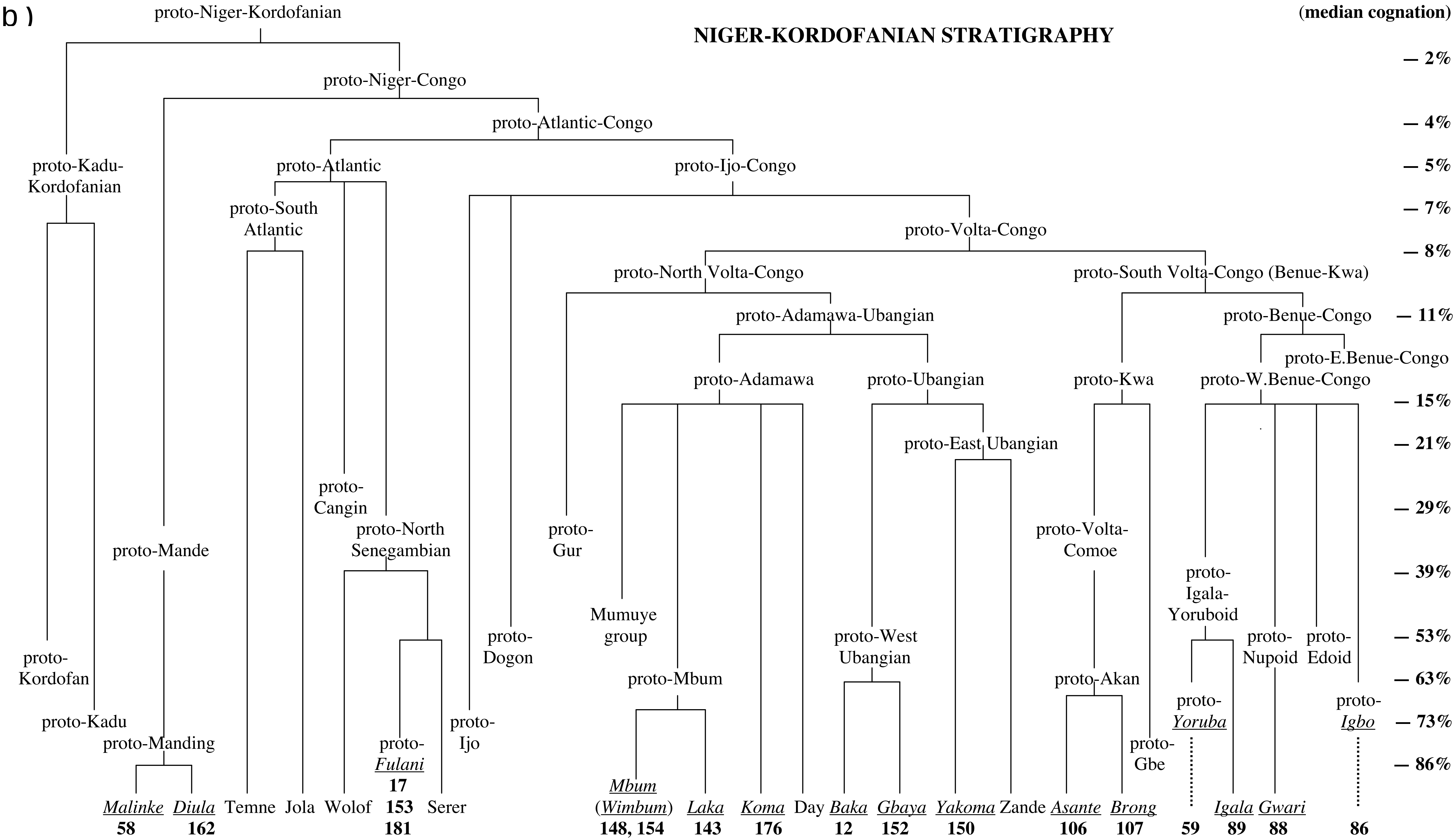


Figure S32b

c)

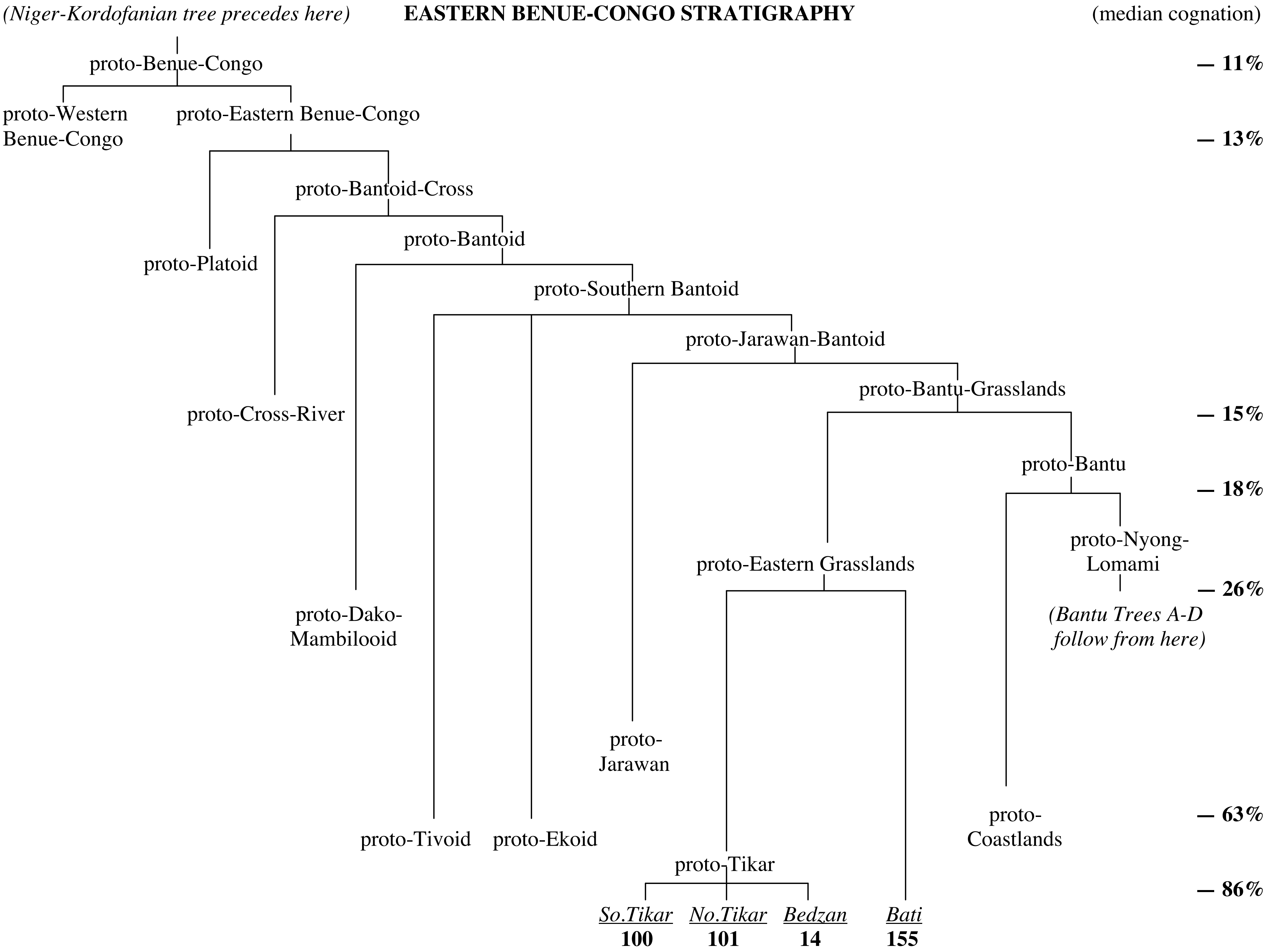


Figure S32c

d)

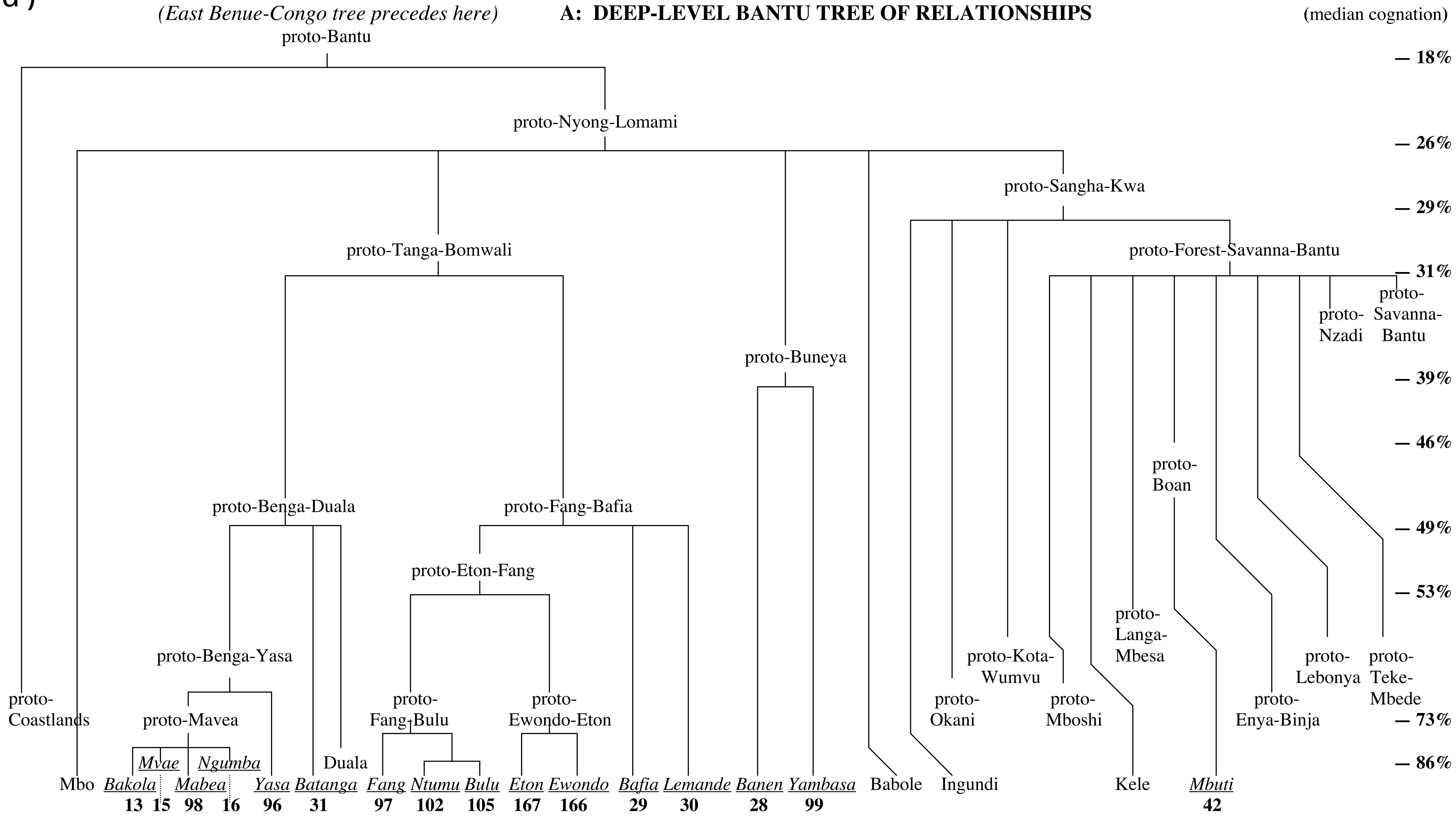


Figure S32d

e)

B: SECOND-LEVEL BANTU TREE OF RELATIONSHIPS: NZADI GROUP

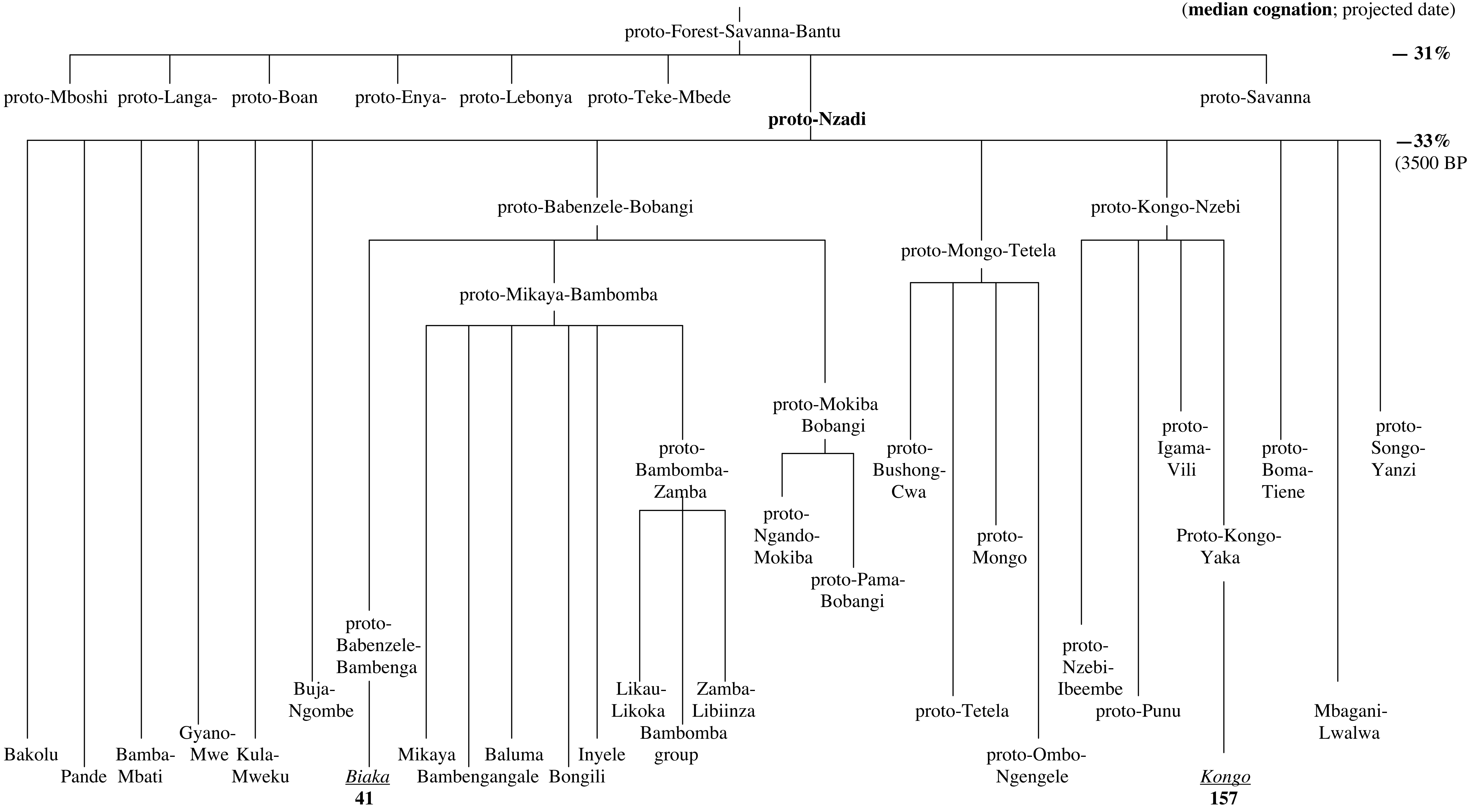


Figure S32e

†)

C: SECOND-LEVEL BANTU TREE OF RELATIONSHIPS: SAVANNA-BANTU

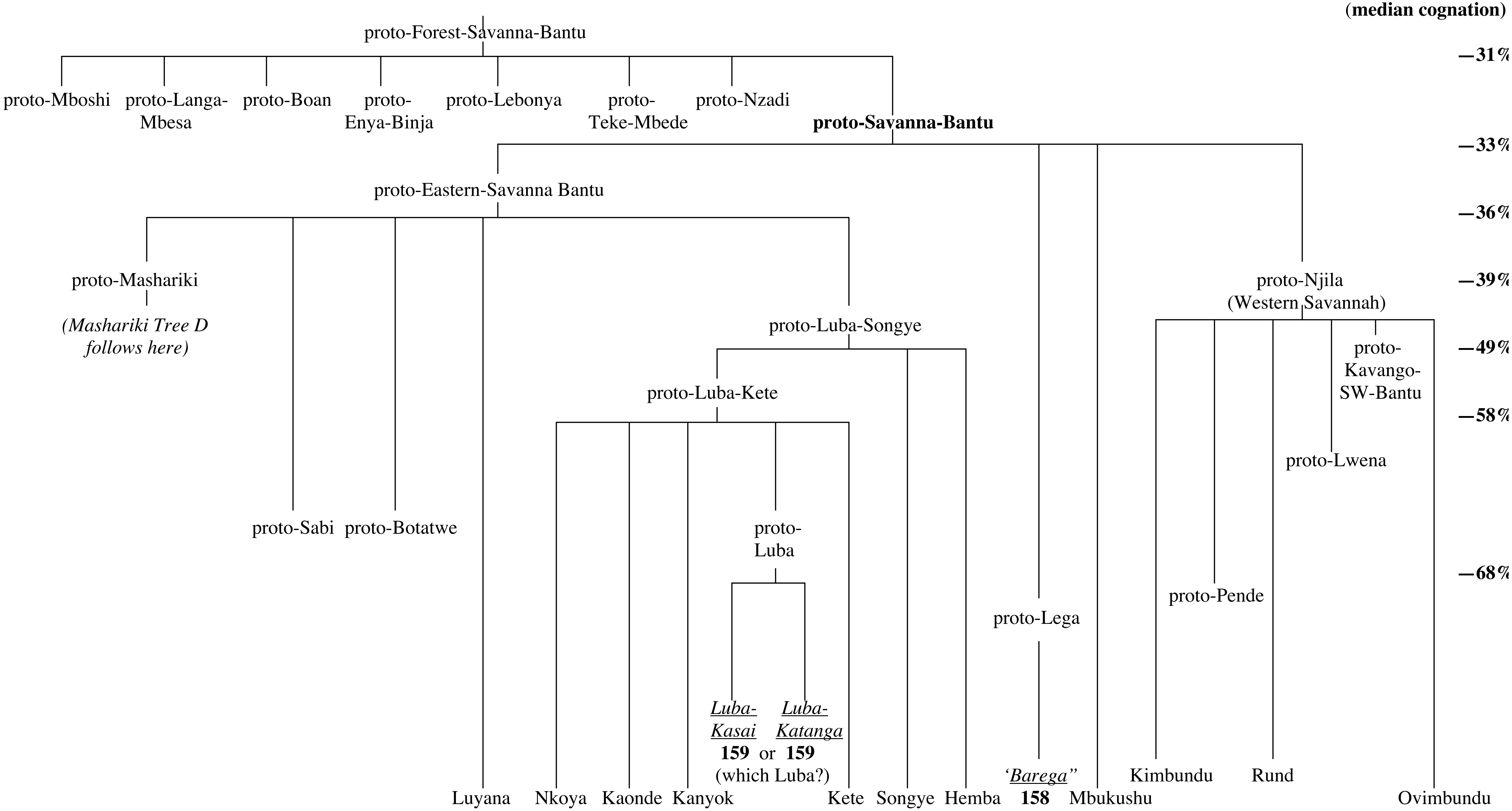


Figure S32†

g)

D: THIRD-LEVEL BANTU TREE OF RELATIONSHIPS: MASHARIKI

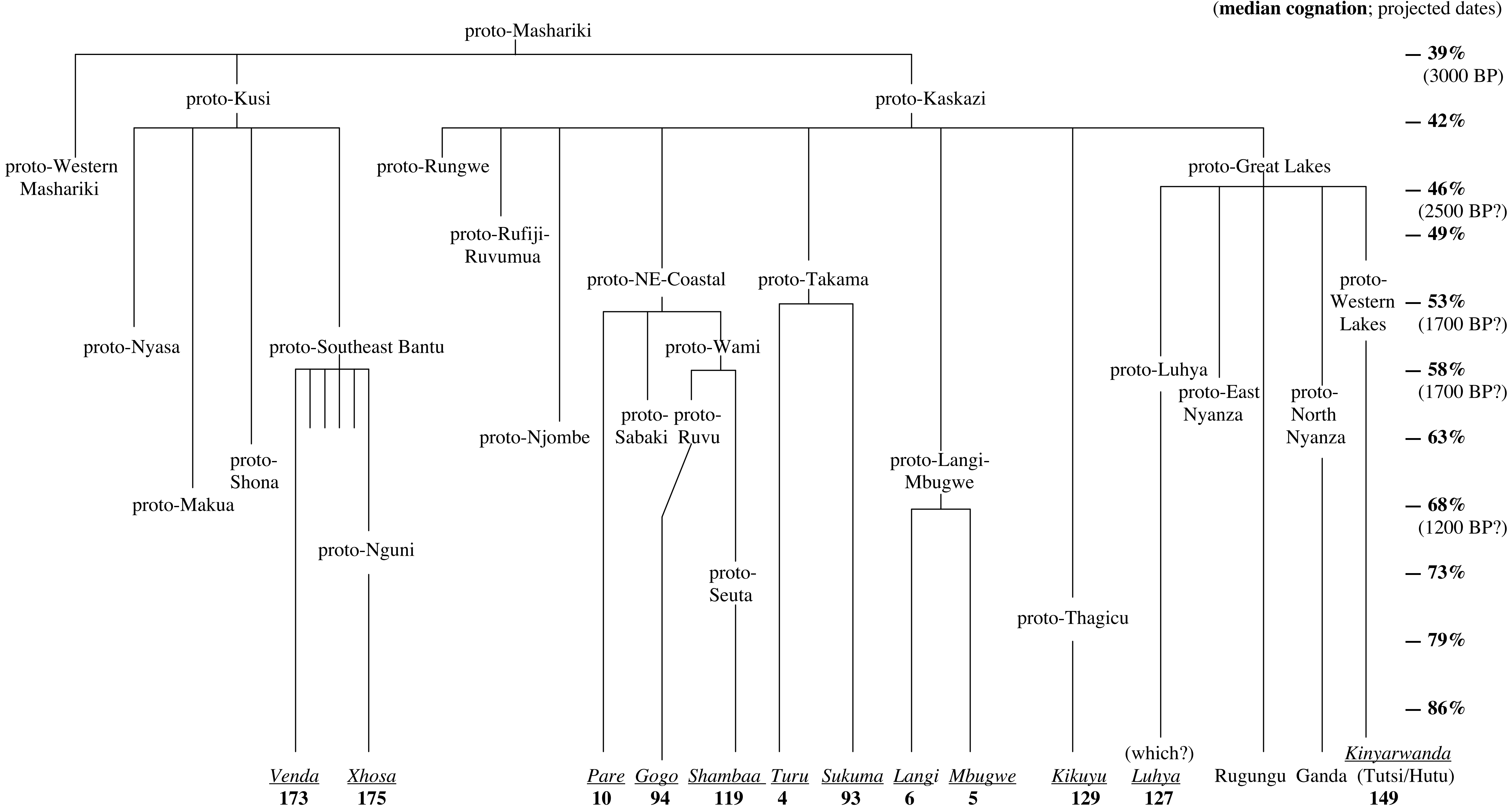


Figure S32g

h)

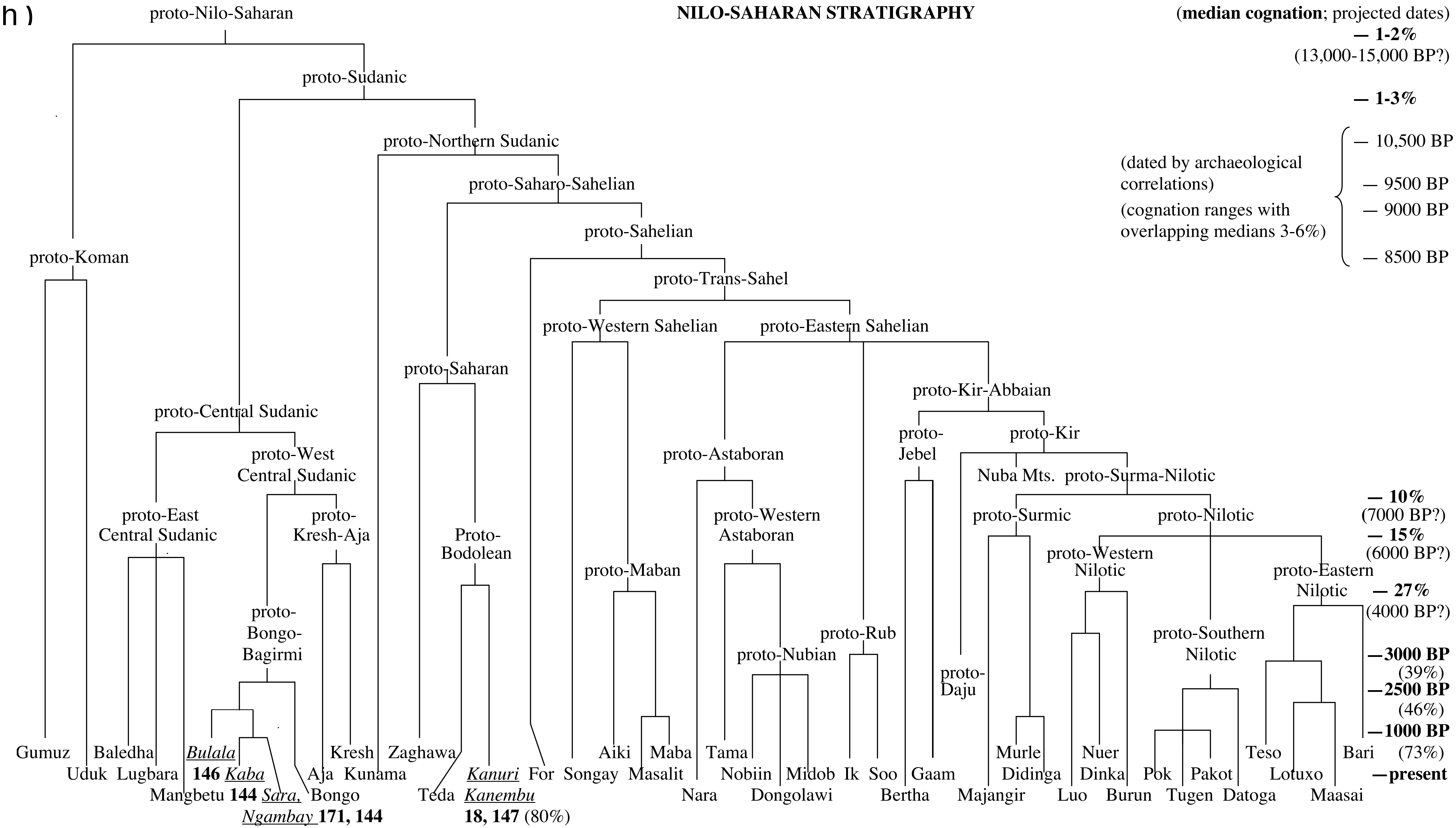


Figure S32h

l)

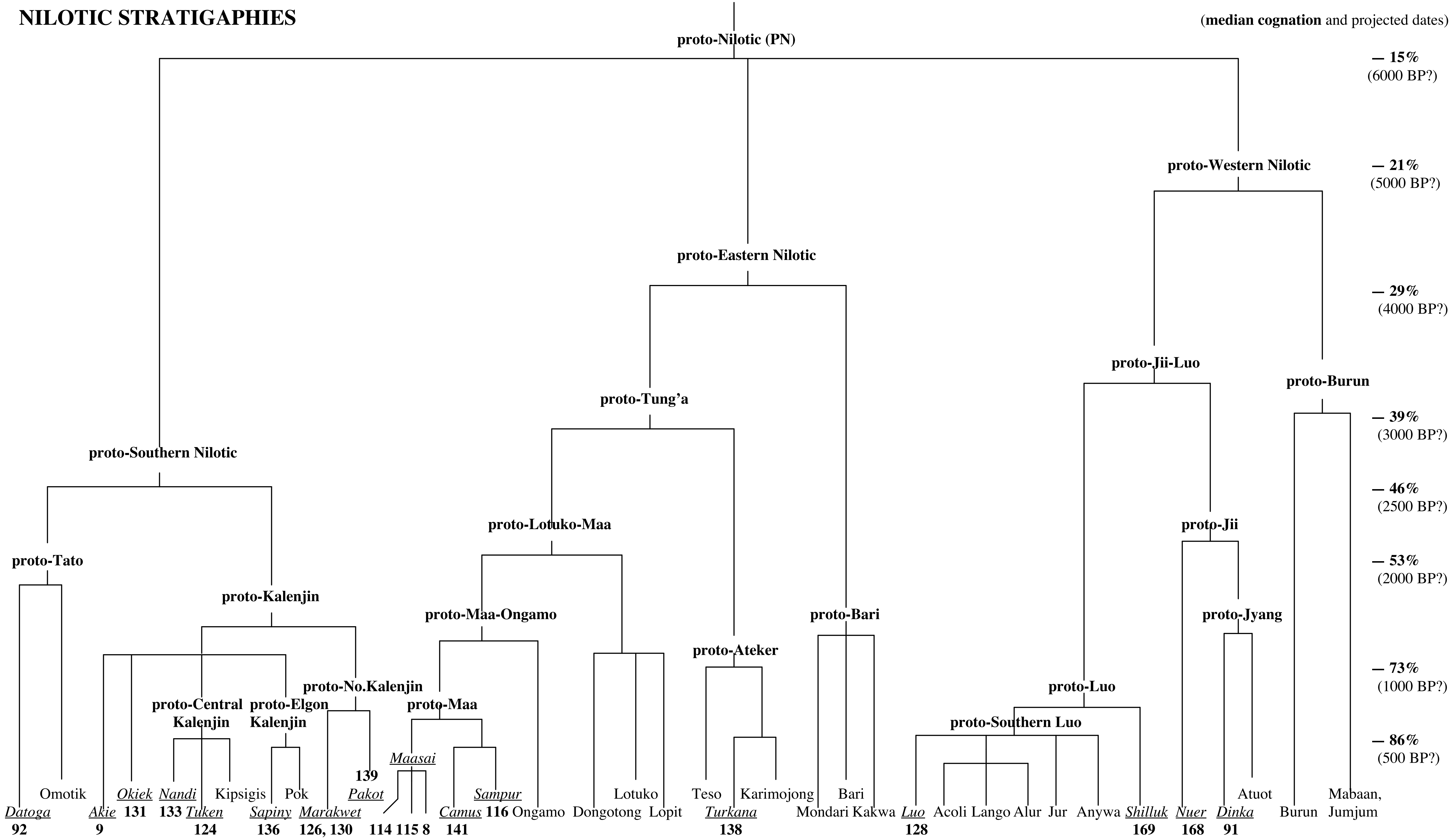


Figure S32l

j)

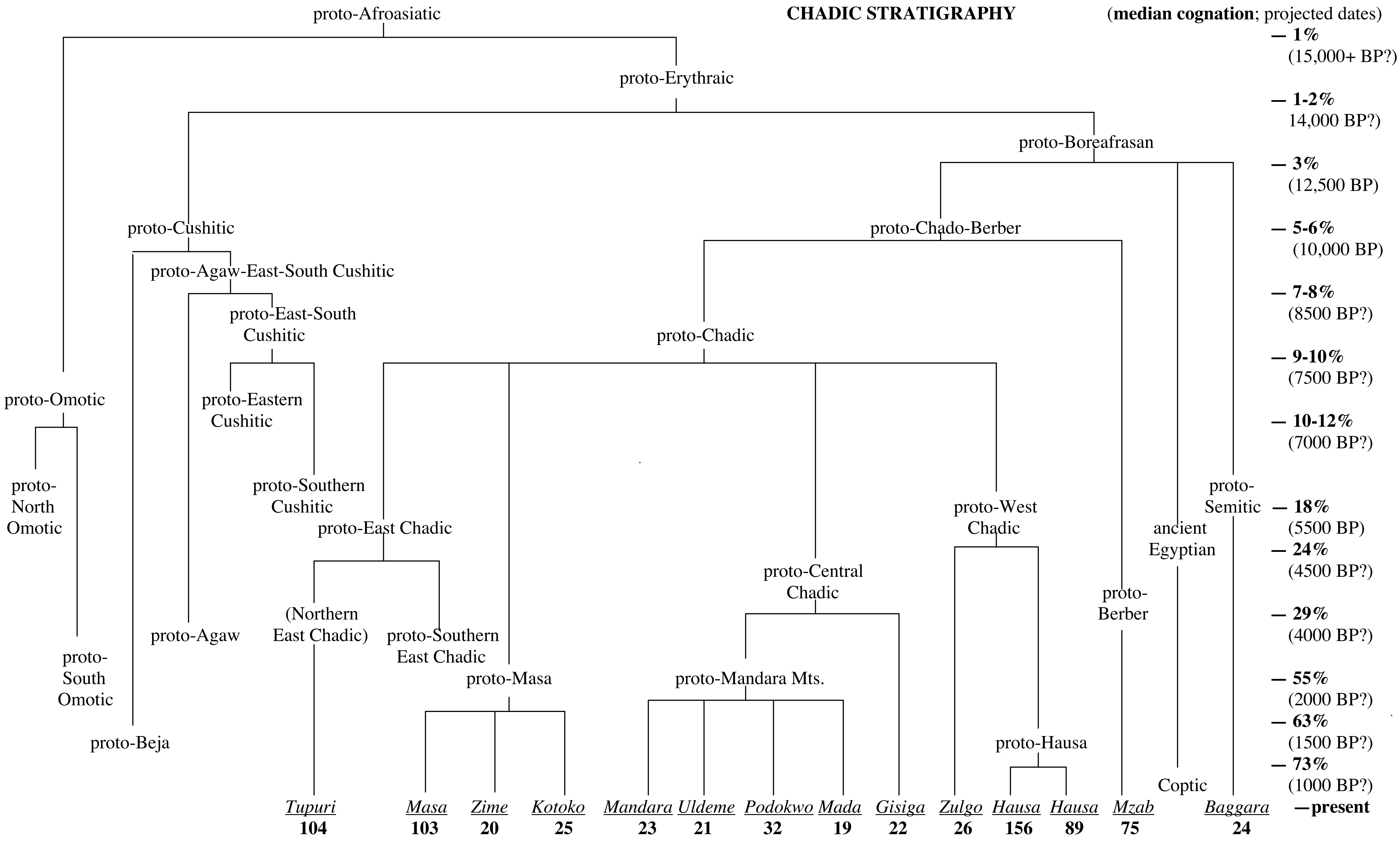


Figure S32j

k)

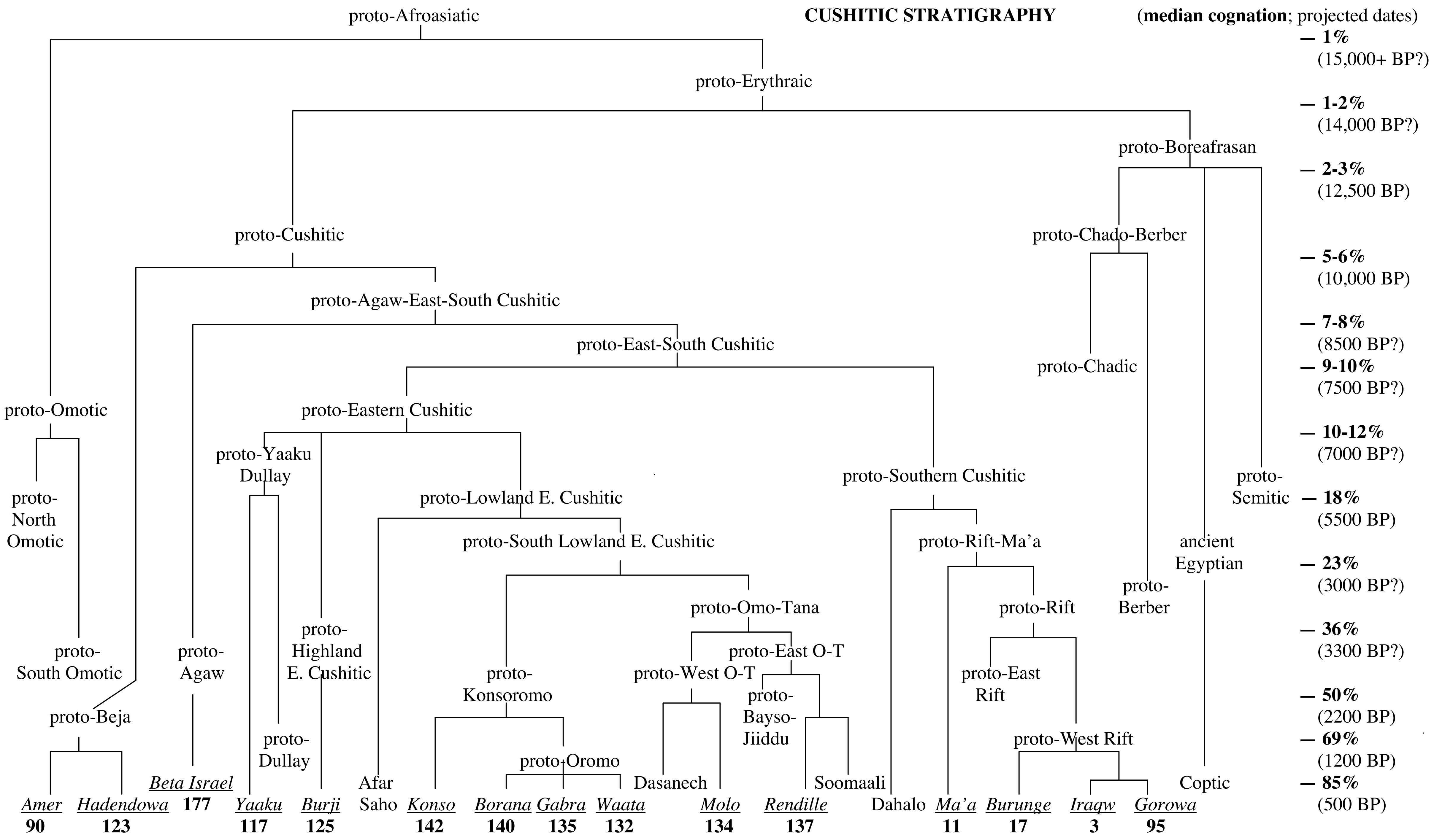


Figure S32k

Figure S32: Linguistic relationships among populations and trees of language divergence constructed with several sources of linguistic, archeological, and ethnographic data (*S50, S51*). Divergence times between related languages were estimated with archeological dates and glottochronological methods (*S54*). Percent similarity is from the number of shared cognates between languages. Linguistic phylogenies for
(a) the Khoesan language family
(b) Niger-Kordofanian language family
(c) Eastern Bantu-congo languages
(d) Deep-level Bantu relationships
(e) Second-level Bantu relationships
(f) Second-level Bantu relationships for Savanna-bantu
(g) Third-level bantu relationships for Mashariki Bantu languages
(h) Nilo-Saharan languages
(i) Nilotic languages
(j) Chadic languages
(k) Cushitic languages

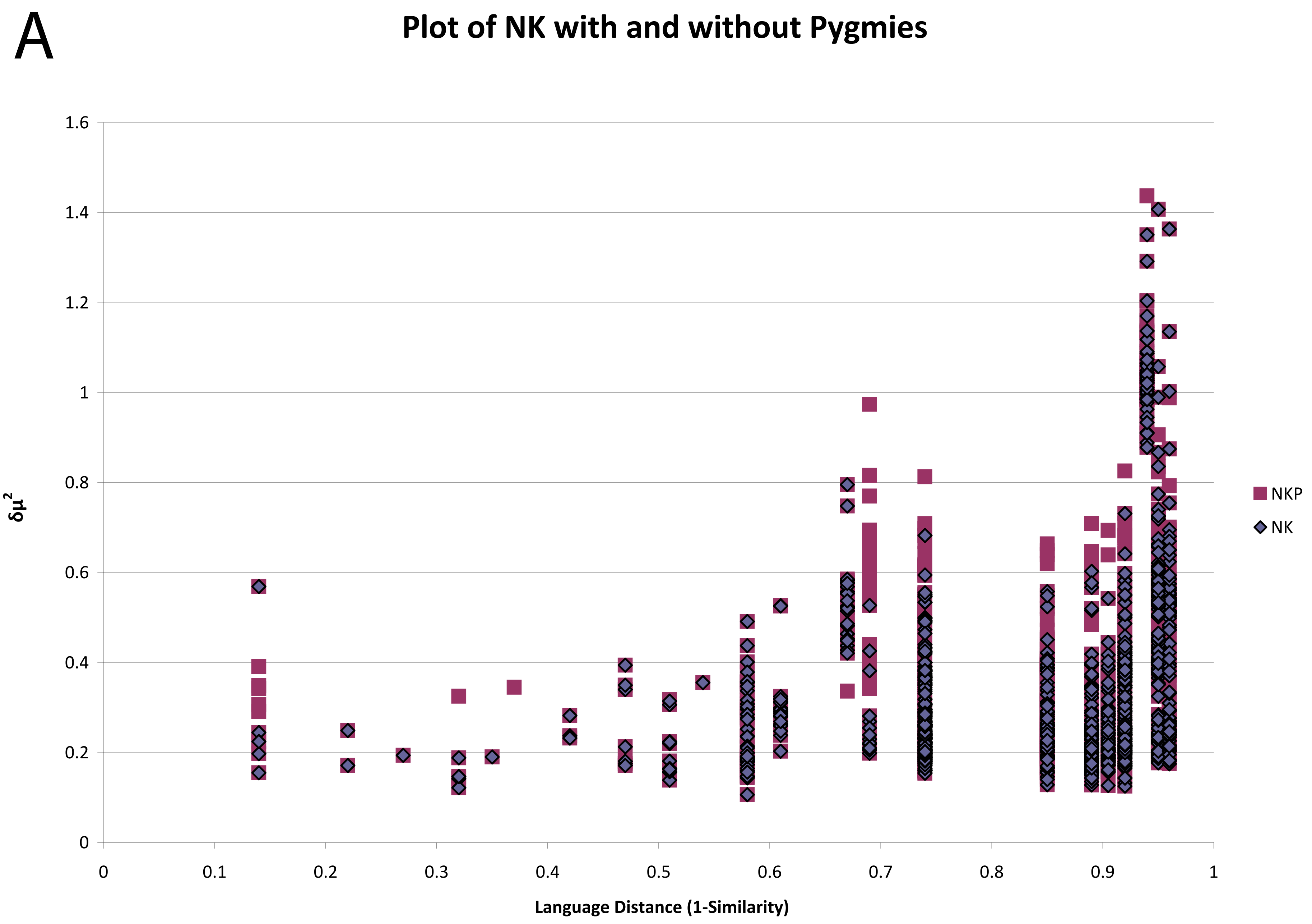

A
Plot of NK with and without Pygmies
δμ²
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Language Distance (1-Similarity)
NKP
NK

Figure S33

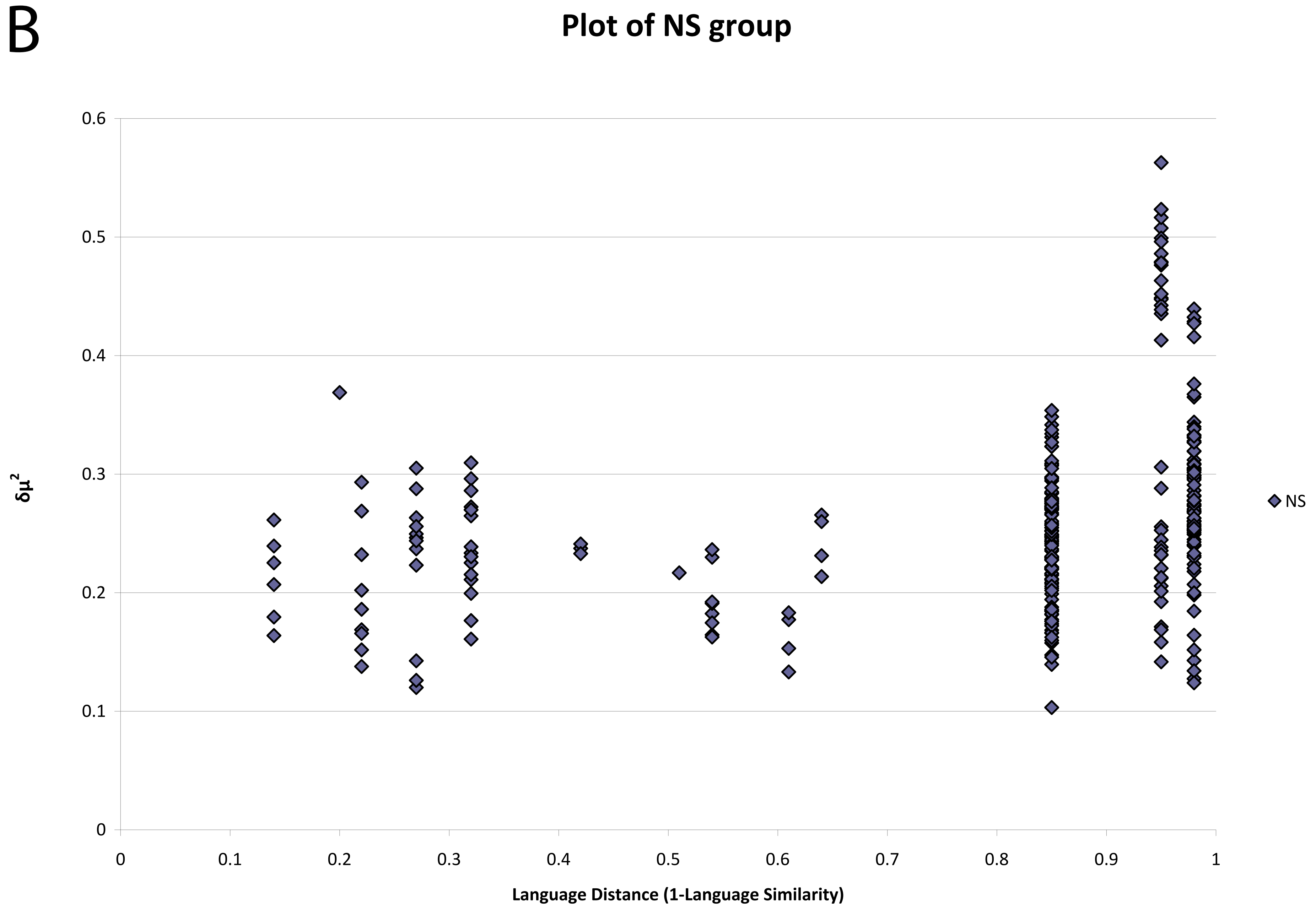

B
Plot of NS group
0.6
0.5
0.4
0.3
0.2
0.1
0
$\delta\mu^2$
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Language Distance (1-Language Similarity)
◆ NS

Figure S33

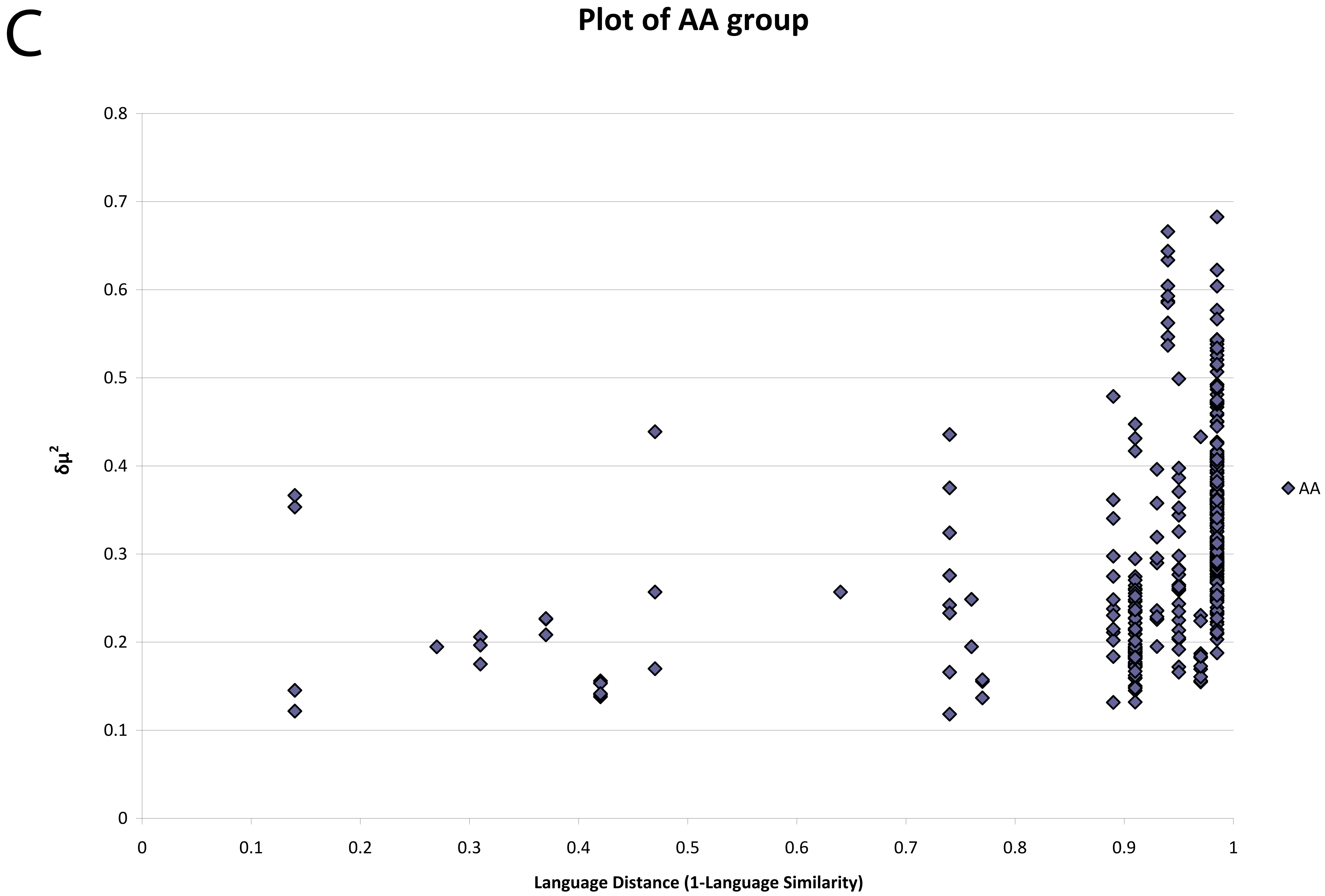


Figure S33: Plot of genetic distance $(\delta\mu)^2$ vs. linguistic distance on the basis of number of shared cognates between pairs of populations within language families. A) NK includes populations speaking a Niger-Kordofanian language, excluding the Pygmies. NKP includes only the pygmies compared to the NK speaking populations and to each other. B) NS includes Nilo-Saharan speaking populations and C) AA includes Afroasiatic speaking populations. The analyses presented in **Table S5** for NK with Pygmies includes the NK and NKP comparisons shown in this figure.

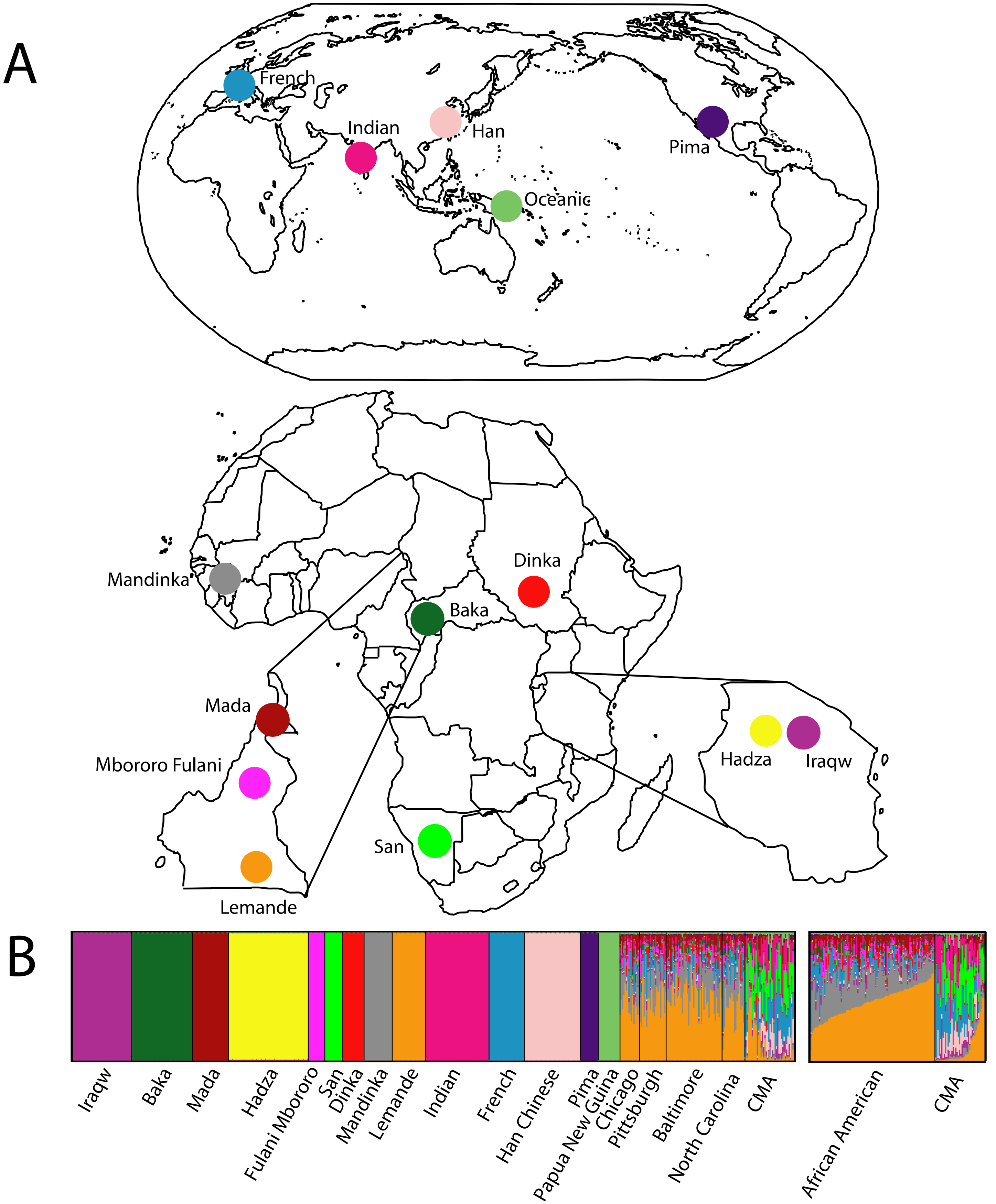


Figure S34: Analyses of Cape Mixed Ancestry (CMA) and African American populations. A) The global distribution of training populations used in supervised STRUCTURE analysis to infer CMA and African American ancestry; B) Inferred ancestry in African American and CMA populations using supervised STRUCTURE analysis. Results for individuals from African American sub-populations and the CMA population are shown on the left, and ancestry by proportion of the Mandinka AAC in the pooled African American population and in the CMA population are shown on the right.

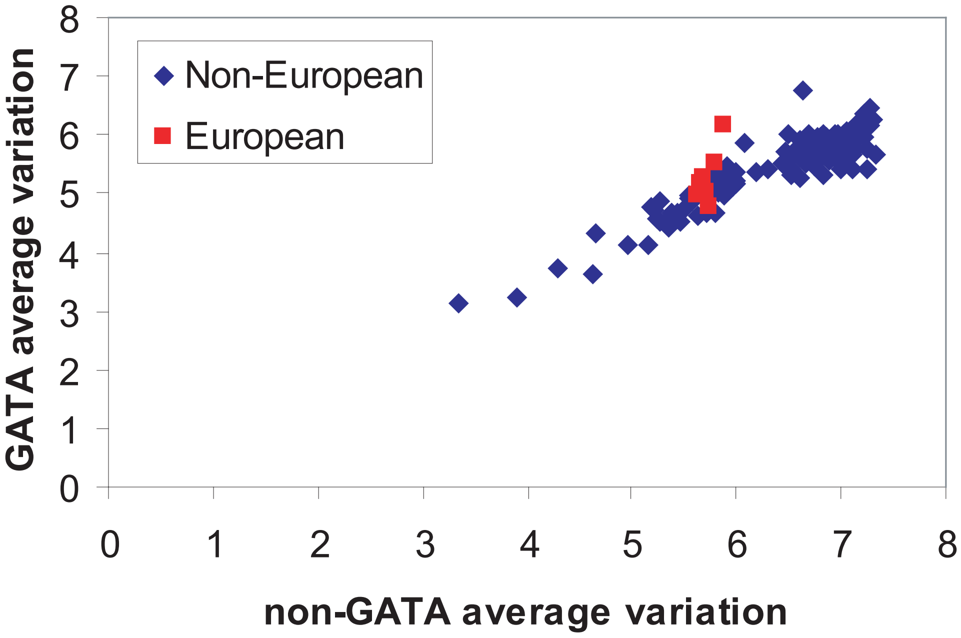


Figure S35: Plot of GATA vs non-GATA microsatellite marker variation. These data indicate a significant difference in deviation between European and non-European samples in the ratio of GATA to non-GATA population variances as described in methods.

Supplementary Tables

Table S1: Populations used in the current study. Also given are approximate geographic coordinates for location of populations and classification on the basis of language and traditional subsistence method.

Table S2: Allele size adjustments made for integrating new African datasets with previously genotyped data from the CEPH diversity panel, India, and Australia.

Table S3: Analysis of Molecular Variance (AMOVA) with 848 microsatellites in the global population and among African populations classified by geographic origin, languages spoken, and subsistence.

Table S4. Genetic by geographic distance analyses a) Spearman’s Rho correlation between genetic and geographic distance by region and within Africa b) heterogeneity of Spearman’s Rho across regions

Table S5: Regressions of genetic distance on linguistic distance, subdivided by major language families and into sub families, and of genetic distance on geographic distance

Table S6: Proportion of AACs for the African American and Cape Mixed Ancestry populations with unsupervised STRUCTURE analysis with the global dataset.

Table S7: Proportion of AACs for the African American and Cape Mixed Ancestry populations with a supervised STRUCTURE analysis. Populations with high frequencies of AACs were used as training populations for this analysis. Because none of the African populations show 100% ancestry from a single AAC, these should not be considered as population admixture estimates.

Table S8: Proportion of AACs in each populations from global STRUCTURE analyses at K=14. Cells that are the maximum for each column and represent the highest proportion for each AAC are highlighted.

Table S9: Proportion of AACs in each populations from Africa STRUCTURE analyses at K=14. Cells that are the maximum for each column and represent the highest proportion for each AAC are highlighted.

Table S1. Populations used in the current study

Population Name	CEPH	n	%Missing 1327 markers	Continent	Region (UN Statistics Division geoscheme with exceptions)	Country	Latitude	Longitude	Language Family	Language Major Subgrouping	ISO 639-3 Language Code	Subsistence	Notes
Banuamir Beja		23	13	Africa	Saharan Africa	Sudan	21.0	36.0	Afroasiatic	Cushitic	BEJ	Herder	
Hadandawa Beja		19	14	Africa	Saharan Africa	Sudan	21.0	36.0	Afroasiatic	Cushitic	BEJ	Herder	
Mozabite	x	29	28	Africa	Saharan Africa	Algeria	32.0	3.0	Afroasiatic	Berber	MZB	Farmer	Cann et al. 2002, Rosenberg et al. 2005
Mandinka	x	22	28	Africa	Western Africa	Senegal	12.0	-12.0	Niger-Kordofanian	Mande	MNK	Farmer	Cann et al. 2002, Rosenberg et al. 2005
Dogon		9	43	Africa	Western Africa	Mali	14.0	-3.0	Niger-Kordofanian	Dogon	DDS	Farmer	
Dioula		5	19	Africa	Western Africa	Ivory Coast	9.5	-4.5	Niger-Kordofanian	Mande	DYU	Farmer	
Ashanti		15	14	Africa	Western Africa	Ghana	6.0	-1.0	Niger-Kordofanian	Kwa	AKA	Farmer	
Brong		26	13	Africa	Western Africa	Ghana	7.5	-2.0	Niger-Kordofanian	Kwa	ABR	Farmer	
Hausa		16	13	Africa	Western Africa	Nigeria	12.0	8.0	Afroasiatic	Chadic	HAU	Farmer	
Fulani		4	13	Africa	Western Africa	Nigeria	11.0	11.0	Niger-Kordofanian	Senegambian	FUV	Herder	
Yoruba		25	14	Africa	Western Africa	Nigeria	8.0	4.0	Niger-Kordofanian	Defoid	YOR	Farmer	new sample, not CEPH
Yoruba (CEPH)	x	22	28	Africa	Western Africa	Nigeria	8.0	5.0	Niger-Kordofanian	Defoid	YOR	Farmer	Cann et al. 2002, Rosenberg et al. 2005, The Y
Igala		17	13	Africa	Western Africa	Nigeria	7.0	7.0	Niger-Kordofanian	Defoid	IGL	Farmer	
Igbo		28	13	Africa	Western Africa	Nigeria	6.0	7.0	Niger-Kordofanian	Igboid	IBO	Farmer	
Gwari		22	15	Africa	Western Africa	Nigeria	10.0	7.0	Niger-Kordofanian	Nupoid	GBR	Farmer	
Bassange		20	13	Africa	Western Africa	Nigeria	9.0	5.5	Niger-Kordofanian	Nupoid	NUP	Farmer	
Koma		12	28	Africa	Western Africa	Nigeria	8.5	12.7	Niger-Kordofanian	Adamawa-Ubangi	KMY	Mixed Farmer	
Mada		28	19	Africa	Central Africa	Cameroon	10.8	14.1	Afroasiatic	Chadic	MXU	Farmer	
Ouldeme		26	17	Africa	Central Africa	Cameroon	11.0	14.3	Afroasiatic	Chadic	UDL	Farmer	
Giziga		24	17	Africa	Central Africa	Cameroon	10.3	14.3	Afroasiatic	Chadic	GIZ	Farmer	
Mandara		26	18	Africa	Central Africa	Cameroon	11.3	14.0	Afroasiatic	Chadic	MFI	Farmer	
Kotoko		17	16	Africa	Central Africa	Cameroon	11.8	14.8	Afroasiatic	Chadic	KOT	Farmer	
Zulgo		22	18	Africa	Central Africa	Cameroon	10.8	14.0	Afroasiatic	Chadic	GND	Farmer	
Podokwo		30	19	Africa	Central Africa	Cameroon	11.0	12.1	Afroasiatic	Chadic	PDI	Farmer	
Massa		15	15	Africa	Central Africa	Cameroon	10.3	15.3	Afroasiatic	Chadic	MCN	Mixed Farmer (with fishing)	
Hausa		27	19	Africa	Central Africa	Cameroon	10.5	14.5	Afroasiatic	Chadic	HAU	Farmer	
Baggara		23	18	Africa	Central Africa	Cameroon	12.5	14.5	Afroasiatic	Semitic	SHU	Herder	
Adamawa Fulani		41	18	Africa	Central Africa	Cameroon	9.0	13.5	Niger-Kordofanian	Senegambian	FUB	Herder	
Mbororo Fulani		13	18	Africa	Central Africa	Cameroon	11.8	14.8	Niger-Kordofanian	Senegambian	FUV	Herder	
Tupuri		22	18	Africa	Central Africa	Cameroon	10.3	14.8	Niger-Kordofanian	Adamawa-Ubangi	TUI	Farmer	
Baka		48	19	Africa	Central Africa	Cameroon	2.5	13.5	Niger-Kordofanian	Adamawa-Ubangi	BKC	Hunter-Gatherer	
Bakola		42	17	Africa	Central Africa	Cameroon	2.8	10.0	Niger-Kordofanian	Bantoid	GYI	Hunter-Gatherer	
Bedzan		17	17	Africa	Central Africa	Cameroon	5.5	11.6	Niger-Kordofanian	Bantoid	TIK	Hunter-Gatherer	
Mvae		24	18	Africa	Central Africa	Cameroon	3.0	12.0	Niger-Kordofanian	Bantoid	FAN	Farmer	
Ngumba		27	21	Africa	Central Africa	Cameroon	3.0	10.3	Niger-Kordofanian	Bantoid	NMG	Farmer	
Zime		30	19	Africa	Central Africa	Cameroon	9.0	14.5	Afroasiatic	Chadic	ZIM	Farmer	
Bamoun		31	19	Africa	Central Africa	Cameroon	5.5	10.8	Niger-Kordofanian	Bantoid	BAX	Farmer	
Banen		25	18	Africa	Central Africa	Cameroon	4.8	10.8	Niger-Kordofanian	Bantoid	BAZ	Farmer	
Bafia		30	19	Africa	Central Africa	Cameroon	4.8	11.0	Niger-Kordofanian	Bantoid	KSF	Farmer	
Lemande		26	19	Africa	Central Africa	Cameroon	4.5	11.0	Niger-Kordofanian	Bantoid	LEM	Farmer	
Batanga		20	18	Africa	Central Africa	Cameroon	3.0	10.0	Niger-Kordofanian	Bantoid	BNM	Farmer	
Iyassa		37	13	Africa	Central Africa	Cameroon	2.5	9.8	Niger-Kordofanian	Bantoid	YKO	Farmer (with fishing)	
Fang		19	14	Africa	Central Africa	Cameroon	2.5	13.0	Niger-Kordofanian	Bantoid	FAN	Farmer	
Mabea		13	13	Africa	Central Africa	Cameroon	2.9	10.3	Niger-Kordofanian	Bantoid	NMG	Farmer (with fishing)	
Yambassa		17	15	Africa	Central Africa	Cameroon	4.8	11.3	Niger-Kordofanian	Bantoid	YAS	Farmer	
South Tikar		21	13	Africa	Central Africa	Cameroon	5.5	11.5	Niger-Kordofanian	Bantoid	TIK	Farmer	
North Tikar		13	13	Africa	Central Africa	Cameroon	6.3	11.5	Niger-Kordofanian	Bantoid	TIK	Farmer	
Ntumu		11	12	Africa	Central Africa	Cameroon	2.3	10.5	Niger-Kordofanian	Bantoid	FAN	Farmer	
Bulu		22	13	Africa	Central Africa	Cameroon	3.0	11.0	Niger-Kordofanian	Bantoid	BUM	Farmer	
Eton		4	17	Africa	Central Africa	Cameroon	4.3	11.5	Niger-Kordofanian	Bantoid	ETO	Farmer	
Wimbum		15	21	Africa	Central Africa	Cameroon	6.5	10.8	Niger-Kordofanian	Bantoid	LMP	Farmer	
Batie		16	20	Africa	Central Africa	Cameroon	4.3	11.0	Niger-Kordofanian	Bantoid	BBJ	Farmer	
Ewondo		3	16	Africa	Central Africa	Cameroon	3.8	11.5	Niger-Kordofanian	Bantoid	EWO	Farmer	
Kanuri		31	18	Africa	Central Africa	Cameroon	11.3	14.3	Nilo-Saharan	Saharan	KNC	Farmer	
Biaka	x	23	27	Africa	Central Africa	C.A.R	4.0	17.0	Niger-Kordofanian	Adamawa-Ubangi	GDI	Hunter-Gatherer	Cann et al. 2002, Rosenberg et al. 2005
Mbum		13	20	Africa	Central Africa	C.A.R	5.5	13.5	Niger-Kordofanian	Adamawa-Ubangi	MDD	Farmer	
Yakoma		6	20	Africa	Central Africa	C.A.R	4.3	22.3	Niger-Kordofanian	Adamawa-Ubangi	YKY	Farmer	
Gbaya		15	20	Africa	Central Africa	C.A.R	5.0	15.0	Niger-Kordofanian	Adamawa-Ubangi	GYA	Farmer	
Laka		33	18	Africa	Central Africa	Chad	8.0	16.0	Nilo-Saharan	Central Sudanic	LAP	Farmer	
Ngambaye		30	19	Africa	Central Africa	Chad	9.0	16.0	Nilo-Saharan	Central Sudanic	SBA	Farmer	
Kaba		27	19	Africa	Central Africa	Chad	8.0	16.8	Nilo-Saharan	Central Sudanic	KSP	Farmer	
Bulala		15	18	Africa	Central Africa	Chad	13.0	18.0	Nilo-Saharan	Central Sudanic	MNE	Farmer (with fishing)	
Kanembou		5	18	Africa	Central Africa	Chad	14.0	15.0	Nilo-Saharan	Saharan	KBL	Mixed Farmer	

Table S1. Populations used in the current study

Population Name	CEPH	n	%Missing 1327 markers	Continent	Region (UN Statistics Division geoscheme with exceptions)	Country	Latitude	Longitude	Language Family	Language Major Subgrouping	ISO 639-3 Language Code	Subsistence	Notes
Sara		27	20	Africa	Central Africa	Chad	8.0	17.5	Nilo-Saharan	Central Sudanic		Farmer	combination of Sara Madjingaye, Sara Gor, Sa
Baluba		6	23	Africa	Central Africa	D.R.C.	-9.0	25.0	Niger-Kordofanian	Bantoid	LUA	Farmer	
Kongo		17	19	Africa	Central Africa	D.R.C.	-5.5	15.0	Niger-Kordofanian	Bantoid	KNG	Farmer	
Barega		4	18	Africa	Central Africa	D.R.C.	-3.5	28.0	Niger-Kordofanian	Bantoid	LEA	Farmer	
Mbuti	x	13	28	Africa	Central Africa	D.R.C.	1.0	29.0	Nilo-Saharan	Central Sudanic	LES	Hunter-Gatherer	Cann et al. 2002, Rosenberg et al. 2005
Tutsi/Hutu		8	18	Africa	Central Africa	Rwanda	-2.0	30.0	Niger-Kordofanian	Bantoid	KIN	Mixed Farmer	Mixed Sample
Dinka		17	14	Africa	Central Africa (deviation from UN cla	Sudan	8.0	30.0	Nilo-Saharan	Eastern Sudanic	DIP	Herder	
Nuer		18	46	Africa	Central Africa (deviation from UN cla	Sudan	8.5	31.0	Nilo-Saharan	Eastern Sudanic	NUS	Herder	
Shilluk		15	45	Africa	Central Africa (deviation from UN cla	Sudan	10.0	32.0	Nilo-Saharan	Eastern Sudanic	SHK	Herder	
Nyimang		12	42	Africa	Central Africa (deviation from UN cla	Sudan	12.3	29.5	Nilo-Saharan	Eastern Sudanic	NYI	Herder	
Hadza		63	18	Africa	Eastern Africa	Tanzania	-3.8	35.3	Khoesan	Hadza	HTS	Hunter-Gatherer	
Sandawe		51	18	Africa	Eastern Africa	Tanzania	-5.5	35.5	Khoesan	Sandawe	SAD	Hunter-Gatherer	
Iraqw		46	18	Africa	Eastern Africa	Tanzania	-4.0	35.5	Afroasiatic	Cushitic	IRK	Mixed Farmer	
Burunge		22	19	Africa	Eastern Africa	Tanzania	-5.3	36.0	Afroasiatic	Cushitic	BDS	Mixed Farmer	
Fiome		22	15	Africa	Eastern Africa	Tanzania	-4.3	35.8	Afroasiatic	Cushitic	GOW	Mixed Farmer	
Mbugu		22	18	Africa	Eastern Africa	Tanzania	-4.8	38.5	mixed A.A. and N.K.		MHD	Mixed Farmer	
Maasai		36	18	Africa	Eastern Africa	Tanzania	-4.0	37.0	Nilo-Saharan	Eastern Sudanic	MAS	Herder	
Akie		23	16	Africa	Eastern Africa	Tanzania	-5.0	37.5	Nilo-Saharan	Eastern Sudanic	OKI	Hunter-Gatherer	
Datog		54	15	Africa	Eastern Africa	Tanzania	-4.5	35.5	Nilo-Saharan	Eastern Sudanic	TCC	Herder	
Turu		32	18	Africa	Eastern Africa	Tanzania	-5.0	35.0	Niger-Kordofanian	Bantoid	RIM	Mixed Farmer	
Sukuma		10	13	Africa	Eastern Africa	Tanzania	-3.0	33.5	Niger-Kordofanian	Bantoid	SUK	Mixed Farmer	
Gogo		13	15	Africa	Eastern Africa	Tanzania	-6.0	36.0	Niger-Kordofanian	Bantoid	GOG	Mixed Farmer	
Mbugwe		21	18	Africa	Eastern Africa	Tanzania	-3.8	35.8	Niger-Kordofanian	Bantoid	MGZ	Mixed Farmer	
Rangi		36	19	Africa	Eastern Africa	Tanzania	-5.0	36.0	Niger-Kordofanian	Bantoid	LAG	Mixed Farmer	
Sambaa		18	14	Africa	Eastern Africa	Tanzania	-4.5	38.3	Niger-Kordofanian	Bantoid	KSB	Mixed Farmer	
Pare		23	18	Africa	Eastern Africa	Tanzania	-4.5	38.0	Niger-Kordofanian	Bantoid	ASA	Mixed Farmer	
Gabra		17	19	Africa	Eastern Africa	Kenya	3.0	37.5	Afroasiatic	Cushitic	GAX	Herder	
Borana		32	21	Africa	Eastern Africa	Kenya	3.0	38.0	Afroasiatic	Cushitic	GAX	Herder	
Wata		6	24	Africa	Eastern Africa	Kenya	3.5	37.0	Afroasiatic	Cushitic	BOB	Hunter-Gatherer	
Rendille		28	18	Africa	Eastern Africa	Kenya	2.3	37.5	Afroasiatic	Cushitic	REL	Herder	
El Molo		16	18	Africa	Eastern Africa	Kenya	2.8	36.8	Afroasiatic	Cushitic	ELO	Hunter-Gatherer	
Yaaku		19	13	Africa	Eastern Africa	Kenya	0.5	37.0	Afroasiatic	Cushitic	MUU	Hunter-Gatherer	
Mumonyot Maasai		12	13	Africa	Eastern Africa	Kenya	0.6	37.0	Nilo-Saharan	Eastern Sudanic	MAS	Herder	
Il'gwesi Maasai		21	14	Africa	Eastern Africa	Kenya	0.3	36.8	Nilo-Saharan	Eastern Sudanic	MAS	Herder	
Dorobo		10	16	Africa	Eastern Africa	Tanzania	-5.0	37.0	Nilo-Saharan	Eastern Sudanic	MWX	Hunter-Gatherer	
Tugen		22	19	Africa	Eastern Africa	Kenya	0.8	35.8	Nilo-Saharan	Eastern Sudanic	TUY	Herder	
Samburu		18	13	Africa	Eastern Africa	Kenya	1.5	37.0	Nilo-Saharan	Eastern Sudanic	SAQ	Herder	
Marakwet		14	19	Africa	Eastern Africa	Kenya	1.3	35.5	Nilo-Saharan	Eastern Sudanic	ENB	Herder	
Sengwer		21	18	Africa	Eastern Africa	Kenya	1.0	35.0	Nilo-Saharan	Eastern Sudanic	KLN	Hunter-Gatherer	
Okiek		22	18	Africa	Eastern Africa	Kenya	0.3	36.0	Nilo-Saharan	Eastern Sudanic	OKI	Hunter-Gatherer	
Nandi		11	21	Africa	Eastern Africa	Kenya	0.0	35.5	Nilo-Saharan	Eastern Sudanic	KLN	Herder	
Sabaot		20	19	Africa	Eastern Africa	Kenya	1.0	34.8	Nilo-Saharan	Eastern Sudanic	SPY	Herder	
Turkana		26	19	Africa	Eastern Africa	Kenya	3.0	36.0	Nilo-Saharan	Eastern Sudanic	TUV	Herder	
Pokot		23	19	Africa	Eastern Africa	Kenya	1.5	35.5	Nilo-Saharan	Eastern Sudanic	PKO	Herder	
Maasai Ilchamus		27	19	Africa	Eastern Africa	Kenya	1.5	37.1	Nilo-Saharan	Eastern Sudanic	MAS	Herder	
Luo		28	18	Africa	Eastern Africa	Kenya	-0.5	34.5	Nilo-Saharan	Eastern Sudanic	LUO	Herder	
Kenya Bantu	x	11	28	Africa	Eastern Africa	Kenya	-3.0	37.0	Niger-Kordofanian	Bantoid		Farmer	Cann et al. 2002, Rosenberg et al. 2005
Kikuyu		22	21	Africa	Eastern Africa	Kenya	-1.0	37.0	Niger-Kordofanian	Bantoid	KIK	Farmer	
Luhya		17	19	Africa	Eastern Africa	Kenya	0.5	34.5	Niger-Kordofanian	Bantoid	LUY	Farmer	
Burji		24	17	Africa	Eastern Africa	Ethiopia	5.5	37.8	Afroasiatic	Cushitic	BJI	Mixed Farmer	
Konso		14	17	Africa	Eastern Africa	Ethiopia	5.5	37.5	Afroasiatic	Cushitic	KXC	Mixed Farmer	
Beta Israel		17	19	Africa	Eastern Africa	Ethiopia	12.0	38.0	Afroasiatic	Semitic	AMH	Mixed Farmer	
Cape Mixed Ancestry		39	27	Africa	Southern Africa	South Africa	-34.0	18.5			ENG		
San	x	6	27	Africa	Southern Africa	Namibia	-21.0	20.0	Khoesan	Southern	KTZ	Hunter-Gatherer	Cann et al. 2002, Rosenberg et al. 2005
!Xun/Khoe		8	28	Africa	Southern Africa	South Africa	-30.0	18.0	Khoesan	Southern	XUU	Hunter-Gatherer	Mixed Sample, Tishkoff et al. 2007
Xhosa		28	22	Africa	Southern Africa	South Africa	-32.0	28.0	Niger-Kordofanian	Bantoid	XHO	Mixed Farmer	
Venda		13	27	Africa	Southern Africa	South Africa	-22.5	30.0	Niger-Kordofanian	Bantoid	VEN	Mixed Farmer	
South Bantu	x	8	27	Africa	Southern Africa	South Africa a	-25.6	24.3	Niger-Kordofanian	Bantoid		Mixed Farmer	Cann et al. 2002, Rosenberg et al. 2005
Chicago		15	14			U.S.A.					ENG		
Pittsburgh		21	13			U.S.A.					ENG		
Baltimore		44	13			U.S.A.					ENG		
North Carolina		18	14			U.S.A.					ENG		

Table S1. Populations used in the current study

Population Name	CEPH	n	%Missing 1327 markers	Continent	Region (UN Statistics Division geoscheme with exceptions)	Country	Latitude	Longitude	Language Family	Language Major Subgrouping	ISO 639-3 Language Code	Subsistence	Notes
Orcadian	x	15	27	Eurasia	Northern Europe	U.K.	59.0	-3.0					Cann et al. 2002, Rosenberg et al. 2005
Adygei	x	17	27	Eurasia	Eastern Europe	Russia	44.0	39.0					Cann et al. 2002, Rosenberg et al. 2005
Russian	x	25	31	Eurasia	Eastern Europe	Russia	61.0	40.0					Cann et al. 2002, Rosenberg et al. 2005
Basque	x	24	28	Eurasia	Western Europe	France	43.0	0.0					Cann et al. 2002, Rosenberg et al. 2005
French	x	28	28	Eurasia	Western Europe	France	46.0	2.0					Cann et al. 2002, Rosenberg et al. 2005
Italian	x	21	28	Eurasia	Southern Europe	Italy	44.5	10.5					Cann et al. 2002, Rosenberg et al. 2005
Sardinian	x	28	28	Eurasia	Southern Europe	Italy	40.0	9.0					Cann et al. 2002, Rosenberg et al. 2005
Bedouin	x	46	27	Eurasia	Western Asia	Israel	31.0	35.0					Cann et al. 2002, Rosenberg et al. 2005
Druze	x	42	28	Eurasia	Western Asia	Israel	32.0	35.0					Cann et al. 2002, Rosenberg et al. 2005
Palestinian	x	46	28	Eurasia	Western Asia	Israel	32.0	35.0					Cann et al. 2002, Rosenberg et al. 2005
Temani		21	24	Eurasia	Western Asia	Yemen	15.0	45.0	Afroasiatic	Semitic	JYE		
Balochi	x	24	28	Eurasia	Southern Asia	Pakistan	30.5	66.5					Cann et al. 2002, Rosenberg et al. 2005
Brahui	x	25	28	Eurasia	Southern Asia	Pakistan	30.5	66.5	Dravidian		BRH		Cann et al. 2002, Rosenberg et al. 2005
Makrani	x	25	28	Eurasia	Southern Asia	Pakistan	25.0	20.0					Cann et al. 2002, Rosenberg et al. 2005
Sindhi	x	24	27	Eurasia	Southern Asia	Pakistan	26.0	69.0					Cann et al. 2002, Rosenberg et al. 2005
Pathan	x	24	27	Eurasia	Southern Asia	Pakistan	25.0	20.0					Cann et al. 2002, Rosenberg et al. 2005
Burusho	x	25	27	Eurasia	Southern Asia	Pakistan	36.5	74.0					Cann et al. 2002, Rosenberg et al. 2005
Hazara	x	22	29	Eurasia	Southern Asia	Pakistan	33.5	70.0					Cann et al. 2002, Rosenberg et al. 2005
Uygur	x	10	28	Eurasia	Eastern Asia	China	44.0	81.0					Cann et al. 2002, Rosenberg et al. 2005
Kalash	x	23	28	Eurasia	Southern Asia	Pakistan	36.0	71.5					Cann et al. 2002, Rosenberg et al. 2005
Parsi		25	11	Eurasia	Southern Asia	India	19.0	72.8					Rosenberg et al. 2006
Kashmiri		25	11	Eurasia	Southern Asia	India	32.4	74.5					Rosenberg et al. 2006
Punjabi		28	12	Eurasia	Southern Asia	India	30.5	75.3					Rosenberg et al. 2006
Hindi		28	11	Eurasia	Southern Asia	India	27.0	78.4					Rosenberg et al. 2006
Marwari		25	11	Eurasia	Southern Asia	India	26.7	74.3					Rosenberg et al. 2006
Gujarat		50	11	Eurasia	Southern Asia	India	23.0	72.0					Rosenberg et al. 2006
Marathi		26	11	Eurasia	Southern Asia	India	19.8	75.9					Rosenberg et al. 2006
Kannada		24	11	Eurasia	Southern Asia	India	15.0	75.0					Rosenberg et al. 2006
Konkani		42	11	Eurasia	Southern Asia	India	14.6	75.1					Rosenberg et al. 2006
Malayalam		25	11	Eurasia	Southern Asia	India	10.0	76.3					Rosenberg et al. 2006
Tamil		29	11	Eurasia	Southern Asia	India	11.1	78.0					Rosenberg et al. 2006
Telugu		27	11	Eurasia	Southern Asia	India	15.9	79.6					Rosenberg et al. 2006
Oriya		26	11	Eurasia	Southern Asia	India	20.0	85.0					Rosenberg et al. 2006
Bengali		27	11	Eurasia	Southern Asia	India	23.3	87.7					Rosenberg et al. 2006
Assamese		25	11	Eurasia	Southern Asia	India	26.0	93.0					Rosenberg et al. 2006
Han	x	44	28	Eurasia	Eastern Asia	China	32.5	114.0					Cann et al. 2002, Rosenberg et al. 2005
Dai	x	10	28	Eurasia	Eastern Asia	China	21.0	100.0					Cann et al. 2002, Rosenberg et al. 2005
Daur	x	10	28	Eurasia	Eastern Asia	China	48.5	124.0					Cann et al. 2002, Rosenberg et al. 2005
Hezhen	x	9	28	Eurasia	Eastern Asia	China	47.5	133.5					Cann et al. 2002, Rosenberg et al. 2005
Lahu	x	8	28	Eurasia	Eastern Asia	China	22.0	100.0					Cann et al. 2002, Rosenberg et al. 2005
Miao	x	10	27	Eurasia	Eastern Asia	China	28.0	109.0					Cann et al. 2002, Rosenberg et al. 2005
Oroqen	x	9	27	Eurasia	Eastern Asia	China	50.5	126.5					Cann et al. 2002, Rosenberg et al. 2005
She	x	10	28	Eurasia	Eastern Asia	China	27.0	119.0					Cann et al. 2002, Rosenberg et al. 2005
Tujia	x	10	29	Eurasia	Eastern Asia	China	29.0	109.0					Cann et al. 2002, Rosenberg et al. 2005
Tu	x	10	27	Eurasia	Eastern Asia	China	36.0	101.0					Cann et al. 2002, Rosenberg et al. 2005
Xibo	x	9	28	Eurasia	Eastern Asia	China	43.5	81.5					Cann et al. 2002, Rosenberg et al. 2005
Yi	x	10	27	Eurasia	Eastern Asia	China	28.0	103.0					Cann et al. 2002, Rosenberg et al. 2005
Mongola	x	10	28	Eurasia	Eastern Asia	China	48.5	119.0					Cann et al. 2002, Rosenberg et al. 2005
Naxi	x	9	28	Eurasia	Eastern Asia	China	26.0	100.0					Cann et al. 2002, Rosenberg et al. 2005
Cambodian	x	10	29	Eurasia	Southeastern Asia	Cambodia	12.0	105.0					Cann et al. 2002, Rosenberg et al. 2005
Japanese	x	29	28	Eurasia	Eastern Asia	Japan	38.0	138.0					Cann et al. 2002, Rosenberg et al. 2005
Yakut	x	25	28	Eurasia	Eastern Asia (deviation from UN cla	Russia	63.0	129.5					Cann et al. 2002, Rosenberg et al. 2005
Melanesian	x	11	28	Oceania	Melanesia	Papua New Guinea							Cann et al. 2002, Rosenberg et al. 2005
Papuan	x	17	28	Oceania	Melanesia	Papua New Guinea							Cann et al. 2002, Rosenberg et al. 2005
Australian		10	13	Oceania	Australia	Australia						Hunter-Gatherer	
Karitiana	x	14	29	America	South America	Brazil							Cann et al. 2002, Rosenberg et al. 2005
Surui	x	8	30	America	South America	Brazil							Cann et al. 2002, Rosenberg et al. 2005
Columbian	x	7	28	America	South America	Columbia							Cann et al. 2002, Rosenberg et al. 2005
Maya	x	21	30	America	Central America	Mexico							Cann et al. 2002, Rosenberg et al. 2005
Pima	x	14	29	America	Central America	Mexico							Cann et al. 2002, Rosenberg et al. 2005

Table S2a. Allele length adjustments between the CEPH-HGDP data and the Africa1 dataset

ID, previous (CEPH-HGDP)	ID, new (Africa1)	adjustment (bp, added to previous length)	Other ID
ATA31G11	ATA31G11P	7	D10S1412
UT2095	UT2095M	1	D11S1304
GATA6B09	GATA6B09P	7	D11S1392
GGAA17G05	GGAA17G05P	7	D11S1984
GATA23E06	GATA23E06L	3	D11S1998
GATA23F06	GATA23F06L	3	D11S1999
GATA28D01	GATA28D01M	1	D11S2000
ATA27A06	ATA27A06P	7	D12S1042
ATA29A06	ATA29A06P	7	D12S1045
GATA26D02	GATA26D02M	1	D12S1052
GATA63D12	GATA63D12P	7	D12S1064
GATA85A04	GATA85A04M	1	D12S1300
GATA91H06	GATA91H06M	1	D12S1301
ATA25F09	ATA25F09M	1	D12S2070
GATA11C08	GATA11C08P	7	D13S1807
GATA23C03	GATA23C03P	7	D13S787
GATA51B02	GATA51B02M	1	D13S796
GATA136B01	GATA136B01M	1	D14S1426
GGAA10C09	GGAA10C09P	7	D14S587
GATA43H01	GATA43H01M	1	D14S608
GGAA21G11	GGAA21G11L	2	D14S617
GATA197B10	GATA197B10P	7	D15S1515
ATA28G05	ATA28G05P	7	D15S655
GATA73F01	GATA73F01M	1	D15S816
GATA81D12	GATA81D12M	1	D16S2624
GATA22F09	GATA22F09P	7	D16S3253
GAAT2C03	GAAT2C03P	7	D17S1298
ATC6A06	ATC6A06M	1	D17S2180
ATA43A10Z	ATA43A10M	1	D17S2193
ATA58A02	ATA58A02P	7	D17S2195
GATA29B01	GATA29B01L	-1	D19S589
GATA44F10	GATA44F10P	7	D19S591
GATA26G09	GATA26G09P	7	D1S1596
GGAA3A07	GGAA3A07M	1	D1S1612
ATA25E07	ATA25E07M	1	D1S1627
ATA29C07	ATA29C07L	-2	D1S3462
GATA29A05	GATA29A05P	7	D1S3669
GATA29F06	GATA29F06z	1	D20S477
UT1355	UT1355z	-1	D21S1411
GGAA10F06	GGAA10F06M	1	D22S686
ATA27D04	ATA27D04P	7	D2S1352
GATA52A04	GATA52A04M	1	D2S1384
GATA65C03	GATA65C03M	1	D2S1391
GATA69E12	GATA69E12M	1	D2S1394
GGAA20G10	GGAA20G10M	1	D2S1400
GATA86E02	GATA86E02P	7	D2S1788
GATA30E06	GATA30E06P	7	D2S2944
GATA178G09	GATA178G09M	1	D2S2968
GATA3C02	GATA3C02M	1	D3S1744
GATA8B05	GATA8B05M	1	D3S1768
GGAA4B09	GGAA4B09N	1	D3S2403
GATA22F11	GATA22F11NZ	-58	D3S2427
GATA27C08	GATA27C08P	7	D3S2432
GATA128C02	GATA128C02M	1	D3S4529
GATA164B08	GATA164B08P	7	D3S4545
GATA7D01	GATA7D01ZP	-37	D4S1627
GATA5B02	GATA5B02M	1	D4S1652

Table S2a. Allele length adjustments between the CEPH-HGDP data and the Africa1 dataset

ID, previous (CEPH-HGDP)	ID, new (Africa1)	adjustment (bp, added to previous length)	Other ID
GATA22G05	GATA22G05M	1	D4S2366
ATA27C07	ATA27C07P	7	D4S2397
GATA42H02	GATA42H02P	8	D4S2417
GATA11A11	GATA11A11P	7	D5S1456
GATA3H06	GATA3H06M	1	D5S1462
GATA7C06	GATA7C06M	1	D5S1470
ATA23A10	ATA23A10M	1	D5S1480
GATA89G08	GATA89G08z	-27	D5S1725
ATA20G07	ATA20G07M	1	D5S2488
GGAT3H10	GGAT3H10M	1	D6S1017
ATA11D10	ATA11D10M	1	D6S1021
ATA22G07	ATA22G07P	7	D6S1027
GATA165G02	GATA165G02M	1	D6S2436
GATA41G07	GATA41G07M	1	D7S1802
GATA24D12	GATA24D12P	7	D7S1818
GATA73D10	GATA73D10L	-4	D7S2204
GATA189C06	GATA189C06M	1	D7S3070
UT7129	UT7129L	-1	D8S1048
GATA8G10	GATA8G10M	1	D8S1110
GATA26E03	GATA26E03M	1	D8S1132
GGAA20C10	GGAA20C10M	1	D8S1477
UT721	UT721M	1	D8S373
GATA6B02	GATA6B02P	7	D8S592
GATA81C04	GATA81C04M	1	D9S1120
GATA62F03	GATA62F03M	1	D9S2169
ATA18A07	ATA18A07M	1	D9S910
GATA178F11	GATA178F11z	-67	NA-D18S-1
GATA133A08	GATA133A08Q	4	NA-D1S-3
ATA42G12	ATA42G12M	1	NA-D1S-4
SRA	SraP	7	TPO-D2S

Table S2b. Allele length adjustments between the Africa1 dataset and the Africa 2-4 datasets

ID, previous	ID, new	adjustment
AATA019	AATA019ZP	-131
ATA70B03P	ATA70B03ZP	-124
AFM248VC5	248VC5ZP	-135
AFM273YF9	273yf9P	-5
GATA109	GATA109Z	-130
GATA66D01	GATA66D01ZP	-110
GATA23D03N	GATA23D03ZP	-107
GATA3C02M	GATA3C02ZP	-25
079yg5P	079YG5ZP	124
TTTAT002Z	tttat002	124
GATA72G09N	GATA72G09Z	25
GATA62A12	GATA62A12Z	-35
165zf8P	165zf8ZP	90
GATA138B05	GATA138B05ZP	49
ATA50C05	ATA50C05ZP	116
035xb9	035xb9ZP	-70
GATA24F03	GATA24F03ZP	7
GATA13G11	GATA13G11ZP	45
AATA019	AATA019ZP	-131
ATT070	ATT070Z	93
GGAA20C10M	GGAA20C10Z	-103
GATA84C01	GATA84C01ZP	-39
ATA21A03	ATA21A03Z	73
ATA27C11	ATA27C11ZP	78
ATA27A06P	ATA27A06N	7
GATA51B02M	GATA51B02ZP	-45
GGAA22G01	GGAA22G01ZP	17
ATA70B03P	ATA70B03ZP	-124
ATA77F05	ATA77F05Z	74
GGAA30H04	GGAA30H04ZP	-103
GATA169E06	GATA169E06ZP	38
ATAC026P	ATAC026	7
AFM273YF9	273yf9P	-5
AFM248VC5	248VC5ZP	-135
204zg5P	204ZG5ZP	115
AAT107	AAT107Z	-53
095tc5P	095TC5ZP	31
AAT200	AAT200ZP	-137
AAT263P	AAT263ZP	-127
GATA138B05	GATA138B05P	-7
ATA11D10M	ATA11D10Z	26

Table S2c. Allele specific adjustments (only the listed alleles within the locus were changed)

marker	from bp	to bp
GATA88H02N	275	274
GATA88H02N	271	270
GATA88H02N	267	266
GATA88H02N	263	262
GATA88H02N	259	258
GATA88H02N	255	254
GATA88H02N	251	250
GATA88H02N	247	246
GATAH05	219	220
GATAH05	223	224
GATAH05	227	228
GATAH05	231	232
GATAH05	235	236
GATAH05	239	240
GATAH05	243	244
GATAH05	247	248
GATAH05	251	252
GATA86E02P	181	182
GATA86E02P	177	178
GATA86E02P	173	174
GATA86E02P	169	170
GATA86E02P	165	166
GATA86E02P	161	162
ATA28B11	240	239
TAGA031z	363	362
TAGA031z	359	358
TAGA031z	357	356
TAGA031z	355	354
GGAA10C09P	288	289
GGAA10C09P	284	285
GGAA10C09P	280	281
GGAA10C09P	276	277
GGAA10C09P	272	273
GGAA10C09P	268	269
GGAA10C09P	264	265
GGAA10C09P	260	261
GGAA10C09P	256	257
GGAA10C09P	252	253
GATA8B01	255	254
GATA8B01	251	250
GATA8B01	247	246
GATA8B01	243	242
GATA8B01	239	238
GATA8B01	235	234
GATA8B01	225	226
UT5029	276	277
UT5029	272	273
UT5029	268	269
UT5029	264	265
UT5029	260	261
ATA33B03Z	101	100
ATA33B03Z	104	103
AGAT114	326	327
AGAT114	322	323

Table S2c. Allele specific adjustments (only the listed alleles within the locus were changed)

marker	from bp	to bp
AGAT114	318	319
AGAT114	314	315
AGAT114	310	311
AGAT114	306	307
AGAT114	302	303
GATA164B08P	223	224
GATA164B08P	219	220
GATA164B08P	215	216
GATA164B08P	211	212
GATA164B08P	207	208
GATA164B08P	203	204
GATA164B08P	199	200
GATA164B08P	195	196
GATA164B08P	187	188
UT1355z	283	284
UT1355z	279	280
UT1355z	275	276
UT1355z	271	272
UT1355z	267	268
GATA91H06M	98	97

Table S3. Analysis of Molecular Variance (AMOVA) by regions for 848 STRs using GDA

						Variance Components and 95% confidence intervals (%)			
Continent			***n***	**Populations**	**Group**		**Among Populations within Regions**	**Within Populations**	**# STRs**
Africa		**CEPH only**	**105**	**7**	**1**	-	**2.59** (2.45,2.74)	**97.41** (97.26,97.55)	783
Africa			**2527**	**120**	**1**	-	**1.71** (1.63,1.81)	**98.29** (98.20,98.37)	
	Region								
	Africa - Eastern		**990**	**42**	**1**	-	**1.17** (1.11,1.23)	**98.83** (98.77,98.89)	
	Africa - Western		**243**	**14**	**1**	-	**1.18** (1.01,1.38)	**98.82** (98.62,99.00)	
	Africa - Central		**1160**	**56**	**1**	-	**1.31** (1.20,1.43)	**98.69** (98.57,98.80)	
	Africa - Saharan		**71**	**3**	**1**	-	**1.29** (1.08,1.52)	**98.71** (98.48,98.92)	847
	Africa - Southern		**63**	**5**	**1**	-	**2.13** (1.86,2.38)	**97.87** (97.62,98.14)	845
	Language Family								
	Afroasiatic		**618**	**27**	**1**	-	**1.29** (1.21,1.38)	**98.71** (98.63,98.79)	
	Khoesan		**128**	**4**	**1**	-	**3.39** (3.20,3.58)	**96.61** (96.42,96.81)	847
	Niger-Kordofanian		**1128**	**59**	**1**	-	**1.17** (1.11,1.26)	**98.83** (98.75,98.89)	
	Nilo-Saharan		**631**	**29**	**1**	-	**1.13** (0.98,1.31)	**98.87** (98.69,99.02)	
	Subsistence								
	Farmers		**1156**	**57**	**1**	-	**0.97** (0.84,1.10)	**99.03** (98.90,99.16)	
	Mixed Farmers		**409**	**21**	**1**	-	**1.07** (1.00,1.15)	**98.93** (98.85,99.00)	
	Hunter-Gatherers		**388**	**16**	**1**	-	**3.18** (3.06,3.32)	**96.82** (96.68,96.94)	
	Herders		**574**	**26**	**1**	-	**0.94** (0.84,1.03)	**99.07** (98.97,99.16)	798
America			**64**	**5**	**1**	-	**8.36** (7.97,8.75)	**91.64** (91.25,92.03)	783
Oceania			**38**	**3**	**1**	-	**4.59** (4.17,5.03)	**95.41** (94.97,95.84)	786
Eurasia			**1179**	**52**	**1**	-	**1.97** (1.86,2.09)	**98.03** (97.91,98.14)	847
	Region								
	Asia		**1021**	**45**	**1**	-	**1.94** (1.83,2.05)	**98.06** (97.95,98.18)	
	Europe		**128**	**7**	**1**	-	**0.74** (0.67,0.82)	**99.26** (99.18,99.33)	783

					Variance Components and 95% confidence intervals (%)		
Continent	**Grouping**	***n***	**Populations**	**Groups**	**Among Groups**	**Among Populations within Groups**	**Within Populations**
Africa	**Geography (5 regions)**	**2527**	**120**	**5**	**0.65** (0.60,0.71)	**1.28** (1.24,1.32)	**98.07** (97.97,98.16)
Africa	**Language (4 families)**	**2505**	**119**	**4**	**0.62** (0.54,0.69)	**1.30** (1.27,1.33)	**98.08** (97.98,98.19)
Africa	**Subsistence (4)**	**2527**	**120**	**4**	**0.44** (0.41,0.48)	**1.41** (1.35,1.47)	**98.15** (98.05,98.24)

Table S4. Genetic by geographic distances analysis

a)

Model	**N**	**Spearman's Rho**	**p-value**
Global Region			
Africa	14042	0.40153	**<0.0001**
Central Asia	56	-0.16600	0.2214
Europe	56	0.82813	**<0.0001**
Middle East	12	0.88273	**0.0001**
East Asia	272	0.33515	**<0.0001**
India	210	0.05653	0.4151
African Region			
Eastern Africa	1722	0.19239	**<0.0001**
Central Africa	2550	0.55188	**<0.0001**
Saharan Africa	42	0.75968	**<0.0001**
Southern Africa	20	-0.11515	0.6288
Western Africa	182	0.42623	**<0.0001**

b)

Correlation1	**Correlation2**	**p-value**
Eastern Africa	Central Africa	**4.62E-42**
Eastern Africa	Saharan Africa	**7.77E-07**
Eastern Africa	Southern Africa	0.2027
Eastern Africa	Western Africa	**0.0009**
Central Africa	Saharan Africa	**0.0203**
Central Africa	Southern Africa	**0.0025**
Central Africa	Western Africa	**0.0321**
Saharan Africa	Southern Africa	**0.0001**
Saharan Africa	Western Africa	**0.0022**
Southern Africa	Western Africa	**0.0245**
Africa	Central Asia	**1.65E-05**
Africa	Europe	**7.78E-13**
Africa	Middle East	**0.0003**
Africa	East Asia	**0.0125**
Africa	India	**8.45E-16**
Central Asia	Europe	**3.93E-12**
Central Asia	Middle East	**1.61E-05**
Central Asia	East Asia	**0.0006**
Central Asia	India	0.1454
Europe	Middle East	0.5680
Europe	East Asia	**2.99E-08**
Europe	India	**2.84E-13**
Middle East	East Asia	**0.0022**
Middle East	India	**9.29E-05**
East Asia	India	**0.0016**

Table S5. Regression of genetic distance on language and geography

	Linguistic Distance[a] R^2	p-value	Geographic Distance[b] R^2	p-value	Full model[c] R^2	Language distance after adjustment p-value[d]	Geographic Distance after adjustment p-value[e]	Full model with Interaction[f] R^2	Language Distance p-value after interaction adjustment	Geographic Distance p-value after interaction adjustment [h]	Interaction p-value
NK without Pygmies	0.16	<0.0001	0.02	<0.0001	0.17	<0.0001	0.0003	0.17	<0.0001	0.0097	0.4591
NK with Pygmies	0.11	<0.0001	0.02	<0.0001	0.12	<0.0001	0.0002	0.12	<0.0001	0.0424	0.99976
NS	0.06	<0.0001	0.21	<0.0001	0.22	0.0051	<0.0001	0.26	0.0188	<0.0001	0.0002
AA all	0.20	<0.0001	0.34	<0.0001	0.34	0.0894	<0.0001	0.37	0.1052	<0.0001	0.0006
Chadic only	0.0012	0.7978	0.39	<0.0001	0.4	0.6028	<0.0001	0.4	0.8315	0.0012	0.9237
Cushitic only	0.29	<0.0001	0.27	<0.0001	0.29	0.0863	0.6723	0.3	0.7587	0.7874	0.1075

a - Proportion of variance in genetic distance explained by linguistic distance alone (unadjusted for anything else)

b - Proportion of variance in genetic distance explained by geographic distance alone (unadjusted for anything else)

c - Proportion of variance in genetic distance explained by linguistic distance and geographic distance together in regression model

d- Significance of linguistic distance effect on genetic distance after adjusting for geographic distance

e- Significance of geographic distance on genetic distance effect after adjusting for linguistic distance

f - Proportion of variance in genetic distance explained by linguistic distance, geographic distance, and interaction between geographic and linguistic distance

g – p-value for effect of linguistic distance on genetic distance after adjusting for geography and interaction

h - p-value for effect of geographic distance on genetic distance after adjusting for language and interaction

Table S6. Proportion of AACs from unsupervised STRUCTURE runs in Cape Mixed Ancestry and African American Populations

Populations		Fulani	Nilo-Saharan	Oceania	Chadic-Saharan	Southern African Khoesan	Niger-Kordofanian	East Asia	Hadza	Sandawe	Indian	Cushitic	Pygmy	European	Native American	n
Cape Mixed Ancestry	**ave**	**0.01**	**0.00**	**0.01**	**0.01**	**0.25**	**0.19**	**0.08**	**0.00**	**0.01**	**0.20**	**0.03**	**0.01**	**0.19**	**0.01**	39
	min	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	max	0.14	0.05	0.06	0.06	0.48	0.71	0.21	0.02	0.11	0.69	0.40	0.07	0.86	0.15	
Chicago	**ave**	**0.03**	**0.00**	**0.00**	**0.01**	**0.01**	**0.70**	**0.01**	**0.00**	**0.01**	**0.04**	**0.02**	**0.01**	**0.15**	**0.00**	15
	min	0.00	0.00	0.00	0.00	0.00	0.52	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.00	
	max	0.14	0.01	0.01	0.05	0.04	0.83	0.08	0.01	0.05	0.11	0.06	0.02	0.29	0.01	
Pittsburgh	**ave**	**0.01**	**0.01**	**0.00**	**0.02**	**0.01**	**0.70**	**0.01**	**0.00**	**0.02**	**0.04**	**0.02**	**0.01**	**0.15**	**0.01**	21
	min	0.00	0.00	0.00	0.00	0.00	0.45	0.00	0.00	0.00	0.01	0.00	0.00	0.03	0.00	
	max	0.03	0.03	0.02	0.05	0.04	0.86	0.03	0.01	0.10	0.15	0.05	0.05	0.36	0.03	
Baltimore	**ave**	**0.01**	**0.01**	**0.00**	**0.01**	**0.00**	**0.74**	**0.01**	**0.00**	**0.02**	**0.05**	**0.02**	**0.01**	**0.11**	**0.00**	44
	min	0.00	0.00	0.00	0.00	0.00	0.54	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	
	max	0.08	0.06	0.04	0.04	0.03	0.88	0.08	0.05	0.10	0.17	0.10	0.07	0.29	0.03	
North Carolina	**ave**	**0.00**	**0.00**	**0.00**	**0.01**	**0.00**	**0.69**	**0.02**	**0.00**	**0.03**	**0.06**	**0.02**	**0.01**	**0.15**	**0.00**	18
	min	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	
	max	0.03	0.01	0.01	0.06	0.01	0.86	0.05	0.04	0.12	0.15	0.07	0.05	0.41	0.02	
All African American	**ave**	**0.01**	**0.01**	**0.00**	**0.01**	**0.01**	**0.71**	**0.01**	**0.00**	**0.02**	**0.05**	**0.02**	**0.01**	**0.13**	**0.00**	98
	min	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	
	max	0.14	0.06	0.04	0.06	0.04	0.88	0.08	0.05	0.12	0.17	0.10	0.07	0.41	0.03	

Table S7. Proportion of AACs from supervised STRUCTURE runs in Cape Mixed Ancestry and African American Populations

Populations		Fulani Mbororo	Nilo-Saharan Dinka	Oceanic PNG	Chadic Mada	SAK San	NK non-Bantu Mandenka	NK Bantu Lemande	East Asian Han	Hadza	Indian	Cushitic Iraqw	Pygmy Baka	European French	Native American Pima	n
Cape Mixed Ancestry	**ave**	**0.02**	**0.02**	**0.02**	**0.03**	**0.19**	**0.04**	**0.10**	**0.09**	**0.01**	**0.15**	**0.04**	**0.02**	**0.24**	**0.01**	39
	min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	max	0.16	0.19	0.11	0.16	0.42	0.39	0.50	0.25	0.07	0.63	0.32	0.12	0.73	0.12	
Chicago	**ave**	**0.06**	**0.03**	**0.01**	**0.08**	**0.01**	**0.19**	**0.44**	**0.01**	**0.01**	**0.02**	**0.02**	**0.02**	**0.10**	**0.01**	15
	min	0.01	0.01	0.00	0.01	0.00	0.04	0.29	0.00	0.00	0.00	0.01	0.01	0.01	0.00	
	max	0.21	0.08	0.02	0.25	0.05	0.35	0.62	0.03	0.02	0.07	0.08	0.06	0.30	0.01	
Pittsburgh	**ave**	**0.04**	**0.03**	**0.00**	**0.06**	**0.01**	**0.21**	**0.45**	**0.01**	**0.01**	**0.03**	**0.03**	**0.02**	**0.10**	**0.01**	21
	min	0.01	0.00	0.00	0.01	0.00	0.05	0.26	0.00	0.00	0.00	0.01	0.00	0.00	0.00	
	max	0.24	0.16	0.01	0.16	0.01	0.37	0.65	0.03	0.04	0.17	0.07	0.05	0.36	0.02	
Baltimore	**ave**	**0.03**	**0.03**	**0.01**	**0.07**	**0.01**	**0.20**	**0.50**	**0.01**	**0.01**	**0.02**	**0.02**	**0.03**	**0.06**	**0.01**	44
	min	0.01	0.01	0.00	0.01	0.00	0.03	0.28	0.00	0.00	0.00	0.01	0.00	0.00	0.00	
	max	0.24	0.11	0.03	0.29	0.03	0.39	0.75	0.05	0.05	0.09	0.11	0.11	0.21	0.03	
North Carolina	**ave**	**0.04**	**0.03**	**0.00**	**0.06**	**0.01**	**0.24**	**0.41**	**0.01**	**0.01**	**0.03**	**0.02**	**0.02**	**0.11**	**0.01**	18
	min	0.01	0.00	0.00	0.00	0.00	0.10	0.08	0.00	0.00	0.00	0.01	0.00	0.00	0.00	
	max	0.20	0.21	0.01	0.15	0.03	0.41	0.63	0.04	0.06	0.09	0.08	0.06	0.39	0.04	

Table S8. Proportion of AACs from the Global Unsupervised Structure Run at K=14

Populations	Continent	Subregion	Fulani	Nilo Saharan	Oceania	Chadic	S.African Khoesan	Niger Kordofanian	East Asian	Hadza	Sandawe	Indian	Cushitic	W. Pygmy	European	Native American	n
Banuamir Beja	Africa	Saharan Africa	0.004	0.017	0.001	0.043	0.002	0.045	0.010	0.003	0.009	0.030	0.514	0.003	0.317	0.002	23
Hadandawa Beja	Africa	Saharan Africa	0.006	0.022	0.001	0.058	0.001	0.040	0.008	0.003	0.006	0.021	0.492	0.004	0.335	0.002	19
Mozabite	Africa	Saharan Africa	0.142	0.006	0.001	0.017	0.002	0.105	0.007	0.004	0.008	0.026	0.075	0.003	0.602	0.002	29
Mandinka	Africa	Western Africa	0.081	0.006	0.006	0.044	0.005	0.792	0.006	0.003	0.012	0.013	0.016	0.007	0.007	0.004	22
Dogon	Africa	Western Africa	0.004	0.005	0.006	0.010	0.164	0.174	0.009	0.002	0.030	0.026	0.117	0.004	0.445	0.004	9
Dioula	Africa	Western Africa	0.006	0.003	0.006	0.035	0.001	0.904	0.004	0.003	0.008	0.014	0.007	0.003	0.005	0.002	5
Ashanti	Africa	Western Africa	0.005	0.009	0.001	0.025	0.003	0.901	0.004	0.003	0.012	0.009	0.011	0.009	0.008	0.002	15
Brong	Africa	Western Africa	0.010	0.006	0.002	0.034	0.002	0.890	0.006	0.003	0.008	0.012	0.016	0.004	0.008	0.002	26
Hausa	Africa	Western Africa	0.019	0.007	0.003	0.048	0.002	0.864	0.006	0.002	0.008	0.011	0.016	0.005	0.007	0.002	16
Fulani	Africa	Western Africa	0.360	0.012	0.001	0.053	0.001	0.474	0.004	0.001	0.008	0.012	0.016	0.003	0.054	0.002	4
Yoruba	Africa	Western Africa	0.003	0.006	0.001	0.015	0.002	0.932	0.004	0.003	0.008	0.008	0.007	0.003	0.006	0.003	25
Yoruba (CEPH)	Africa	Western Africa	0.004	0.005	0.002	0.016	0.004	0.915	0.004	0.003	0.011	0.009	0.010	0.005	0.009	0.003	22
Igala	Africa	Western Africa	0.001	0.005	0.001	0.013	0.002	0.931	0.003	0.003	0.009	0.010	0.010	0.005	0.004	0.003	17
Igbo	Africa	Western Africa	0.004	0.007	0.002	0.009	0.003	0.920	0.003	0.003	0.013	0.009	0.014	0.008	0.005	0.001	28
Gwari	Africa	Western Africa	0.006	0.008	0.002	0.032	0.002	0.895	0.003	0.003	0.010	0.008	0.019	0.004	0.008	0.002	22
Bassange	Africa	Western Africa	0.002	0.004	0.001	0.030	0.003	0.913	0.005	0.004	0.006	0.009	0.009	0.003	0.006	0.004	20
Koma	Africa	Western Africa	0.027	0.068	0.003	0.049	0.071	0.717	0.006	0.004	0.010	0.011	0.008	0.003	0.020	0.004	12
Mada	Africa	Central Africa	0.002	0.009	0.001	0.477	0.002	0.471	0.004	0.002	0.008	0.005	0.008	0.005	0.004	0.001	28
Ouldeme	Africa	Central Africa	0.004	0.004	0.001	0.471	0.002	0.477	0.003	0.003	0.006	0.006	0.009	0.006	0.005	0.002	26
Giziga	Africa	Central Africa	0.005	0.011	0.002	0.371	0.004	0.557	0.003	0.002	0.010	0.009	0.014	0.004	0.007	0.001	24
Mandara	Africa	Central Africa	0.016	0.023	0.001	0.379	0.003	0.512	0.004	0.003	0.010	0.011	0.023	0.005	0.008	0.002	26
Kotoko	Africa	Central Africa	0.008	0.023	0.001	0.401	0.002	0.473	0.002	0.005	0.006	0.011	0.035	0.009	0.019	0.003	17
Zulgo	Africa	Central Africa	0.002	0.007	0.001	0.452	0.001	0.495	0.002	0.002	0.006	0.006	0.010	0.011	0.003	0.001	22
Podokwo	Africa	Central Africa	0.001	0.009	0.001	0.467	0.002	0.479	0.003	0.004	0.005	0.006	0.010	0.007	0.004	0.002	30
Massa	Africa	Central Africa	0.002	0.063	0.001	0.352	0.002	0.528	0.003	0.004	0.009	0.006	0.020	0.004	0.004	0.002	15
Hausa	Africa	Central Africa	0.019	0.007	0.002	0.150	0.003	0.748	0.004	0.003	0.010	0.012	0.021	0.007	0.012	0.002	27
Baggara	Africa	Central Africa	0.085	0.075	0.002	0.265	0.002	0.360	0.007	0.002	0.010	0.019	0.113	0.006	0.051	0.002	23
Adamawa Fulani	Africa	Central Africa	0.289	0.008	0.001	0.176	0.002	0.449	0.004	0.002	0.008	0.013	0.016	0.004	0.025	0.004	41
Mbororo Fulani	Africa	Central Africa	0.627	0.002	0.001	0.016	0.001	0.283	0.005	0.001	0.006	0.013	0.016	0.002	0.026	0.002	13
Tupuri	Africa	Central Africa	0.003	0.027	0.001	0.325	0.002	0.590	0.003	0.002	0.008	0.008	0.019	0.005	0.007	0.002	22
Baka	Africa	Central Africa	0.001	0.003	0.001	0.005	0.002	0.326	0.002	0.002	0.004	0.004	0.005	0.642	0.002	0.001	48
Bakola	Africa	Central Africa	0.001	0.002	0.001	0.005	0.003	0.366	0.002	0.003	0.008	0.005	0.004	0.596	0.002	0.001	42
Bedzan	Africa	Central Africa	0.002	0.002	0.000	0.004	0.002	0.536	0.002	0.004	0.004	0.006	0.006	0.427	0.003	0.002	17
Mvae	Africa	Central Africa	0.001	0.009	0.002	0.006	0.004	0.887	0.003	0.002	0.015	0.007	0.016	0.040	0.007	0.001	24
Ngumba	Africa	Central Africa	0.003	0.007	0.001	0.004	0.002	0.862	0.003	0.005	0.016	0.009	0.013	0.068	0.005	0.002	27
Zime	Africa	Central Africa	0.001	0.008	0.001	0.018	0.003	0.852	0.003	0.005	0.022	0.008	0.013	0.059	0.005	0.002	30
Bamoun	Africa	Central Africa	0.003	0.011	0.001	0.010	0.003	0.919	0.003	0.002	0.011	0.007	0.015	0.009	0.005	0.001	31
Banen	Africa	Central Africa	0.002	0.009	0.002	0.006	0.006	0.914	0.003	0.003	0.013	0.006	0.014	0.015	0.005	0.002	25
Bafia	Africa	Central Africa	0.002	0.005	0.001	0.010	0.004	0.914	0.004	0.004	0.013	0.007	0.011	0.017	0.005	0.002	30
Lemande	Africa	Central Africa	0.001	0.005	0.001	0.008	0.002	0.935	0.005	0.004	0.009	0.005	0.008	0.010	0.004	0.002	26
Batanga	Africa	Central Africa	0.003	0.006	0.001	0.012	0.004	0.896	0.004	0.003	0.016	0.007	0.013	0.030	0.004	0.001	20
Iyassa	Africa	Central Africa	0.001	0.004	0.001	0.006	0.004	0.891	0.003	0.004	0.018	0.011	0.013	0.026	0.016	0.002	37
Fang	Africa	Central Africa	0.002	0.009	0.001	0.006	0.003	0.887	0.003	0.005	0.012	0.006	0.016	0.045	0.004	0.001	19
Mabea	Africa	Central Africa	0.003	0.006	0.001	0.005	0.003	0.883	0.003	0.003	0.010	0.006	0.008	0.064	0.005	0.002	13
Yambassa	Africa	Central Africa	0.002	0.006	0.001	0.010	0.002	0.917	0.004	0.002	0.019	0.009	0.009	0.012	0.004	0.002	17
South Tikar	Africa	Central Africa	0.002	0.006	0.002	0.017	0.009	0.897	0.003	0.004	0.013	0.007	0.010	0.023	0.005	0.002	21
North Tikar	Africa	Central Africa	0.003	0.009	0.001	0.018	0.006	0.891	0.003	0.002	0.023	0.009	0.014	0.015	0.005	0.001	13
Ntumu	Africa	Central Africa	0.001	0.004	0.001	0.008	0.002	0.922	0.003	0.007	0.011	0.006	0.007	0.018	0.009	0.001	11
Bulu	Africa	Central Africa	0.001	0.005	0.001	0.008	0.005	0.901	0.002	0.004	0.013	0.010	0.011	0.030	0.006	0.001	22
Eton	Africa	Central Africa	0.003	0.013	0.002	0.010	0.001	0.895	0.009	0.004	0.008	0.012	0.009	0.029	0.004	0.001	4
Wimbum	Africa	Central Africa	0.002	0.009	0.004	0.007	0.004	0.908	0.004	0.003	0.019	0.011	0.015	0.007	0.005	0.002	15

Table S8. Proportion of AACs from the Global Unsupervised Structure Run at K=14

Populations	Continent	Subregion	Fulani	Nilo Saharan	Oceania	Chadic	S.African Khoesan	Niger Kordofanian	East Asian	Hadza	Sandawe	Indian	Cushitic	W. Pygmy	European	Native American	n
Batie	Africa	Central Africa	0.001	0.009	0.002	0.008	0.007	0.921	0.003	0.002	0.014	0.008	0.008	0.009	0.003	0.003	16
Ewondo	Africa	Central Africa	0.002	0.004	0.001	0.021	0.002	0.912	0.002	0.001	0.009	0.005	0.007	0.032	0.003	0.001	3
Kanuri	Africa	Central Africa	0.023	0.037	0.001	0.296	0.004	0.556	0.005	0.004	0.016	0.008	0.034	0.007	0.008	0.002	31
Biaka	Africa	Central Africa	0.001	0.003	0.001	0.011	0.006	0.310	0.002	0.001	0.006	0.005	0.005	0.646	0.003	0.001	23
Mbum	Africa	Central Africa	0.003	0.014	0.002	0.132	0.004	0.809	0.002	0.003	0.007	0.006	0.010	0.003	0.004	0.002	13
Yakoma	Africa	Central Africa	0.001	0.020	0.002	0.123	0.008	0.749	0.008	0.004	0.014	0.013	0.012	0.043	0.004	0.001	6
Gbaya	Africa	Central Africa	0.002	0.022	0.001	0.086	0.003	0.822	0.004	0.003	0.006	0.011	0.020	0.012	0.006	0.002	15
Laka	Africa	Central Africa	0.004	0.010	0.001	0.153	0.005	0.767	0.005	0.004	0.009	0.009	0.019	0.008	0.004	0.002	33
Ngambaye	Africa	Central Africa	0.003	0.023	0.001	0.192	0.002	0.722	0.005	0.003	0.009	0.011	0.016	0.005	0.005	0.003	30
Kaba	Africa	Central Africa	0.002	0.016	0.002	0.147	0.004	0.768	0.005	0.003	0.010	0.011	0.019	0.006	0.005	0.003	27
Bulala	Africa	Central Africa	0.009	0.127	0.001	0.404	0.003	0.342	0.004	0.003	0.009	0.012	0.061	0.004	0.019	0.002	15
Kanembou	Africa	Central Africa	0.006	0.060	0.004	0.311	0.005	0.408	0.002	0.002	0.011	0.023	0.112	0.003	0.051	0.003	5
Sara	Africa	Central Africa	0.005	0.026	0.001	0.163	0.002	0.736	0.004	0.004	0.010	0.008	0.027	0.007	0.005	0.002	27
Baluba	Africa	Central Africa	0.002	0.014	0.001	0.006	0.003	0.909	0.006	0.003	0.012	0.007	0.007	0.025	0.003	0.002	6
Kongo	Africa	Central Africa	0.001	0.006	0.001	0.007	0.003	0.907	0.003	0.003	0.023	0.010	0.012	0.015	0.006	0.003	17
Barega	Africa	Central Africa	0.002	0.013	0.001	0.030	0.041	0.826	0.002	0.009	0.042	0.010	0.009	0.011	0.002	0.002	4
Mbuti	Africa	Central Africa	0.000	0.031	0.001	0.006	0.568	0.134	0.001	0.005	0.011	0.003	0.007	0.230	0.002	0.001	13
Tutsi/Hutu	Africa	Central Africa	0.001	0.038	0.001	0.004	0.007	0.655	0.005	0.002	0.064	0.016	0.177	0.018	0.011	0.002	8
Dinka	Africa	Central Africa	0.001	0.522	0.001	0.266	0.003	0.149	0.003	0.004	0.006	0.003	0.033	0.006	0.002	0.001	17
Nuer	Africa	Central Africa	0.002	0.531	0.001	0.239	0.004	0.142	0.003	0.003	0.010	0.009	0.042	0.008	0.004	0.002	18
Shilluk	Africa	Central Africa	0.002	0.499	0.001	0.233	0.002	0.198	0.003	0.008	0.006	0.005	0.034	0.004	0.003	0.001	15
Nyimang	Africa	Central Africa	0.010	0.375	0.001	0.267	0.022	0.154	0.003	0.002	0.011	0.009	0.121	0.005	0.018	0.002	12
Hadza	Africa	Eastern Africa	0.002	0.006	0.001	0.005	0.003	0.206	0.003	0.671	0.025	0.008	0.057	0.005	0.006	0.003	63
Sandawe	Africa	Eastern Africa	0.002	0.008	0.001	0.007	0.004	0.171	0.003	0.006	0.688	0.011	0.083	0.004	0.010	0.002	51
Iraqw	Africa	Eastern Africa	0.002	0.003	0.002	0.004	0.002	0.054	0.009	0.008	0.085	0.013	0.788	0.007	0.018	0.003	46
Burunge	Africa	Eastern Africa	0.003	0.013	0.001	0.005	0.007	0.248	0.007	0.014	0.285	0.013	0.375	0.016	0.010	0.002	22
Fiome	Africa	Eastern Africa	0.005	0.008	0.001	0.009	0.004	0.109	0.009	0.012	0.101	0.019	0.691	0.005	0.023	0.003	22
Mbugu	Africa	Eastern Africa	0.002	0.007	0.002	0.011	0.003	0.293	0.005	0.003	0.034	0.018	0.605	0.004	0.012	0.003	22
Maasai	Africa	Eastern Africa	0.002	0.136	0.001	0.023	0.006	0.230	0.006	0.008	0.050	0.018	0.486	0.008	0.021	0.004	36
Akie	Africa	Eastern Africa	0.003	0.032	0.001	0.007	0.004	0.143	0.008	0.006	0.151	0.016	0.600	0.006	0.019	0.004	23
Datog	Africa	Eastern Africa	0.005	0.050	0.002	0.026	0.003	0.128	0.009	0.007	0.044	0.016	0.683	0.007	0.015	0.005	54
Turu	Africa	Eastern Africa	0.003	0.010	0.002	0.007	0.006	0.374	0.009	0.009	0.232	0.013	0.312	0.010	0.009	0.004	32
Sukuma	Africa	Eastern Africa	0.002	0.023	0.004	0.003	0.011	0.677	0.006	0.016	0.106	0.010	0.125	0.006	0.008	0.002	10
Gogo	Africa	Eastern Africa	0.002	0.014	0.001	0.007	0.017	0.515	0.005	0.009	0.196	0.009	0.205	0.007	0.010	0.004	13
Mbugwe	Africa	Eastern Africa	0.002	0.013	0.002	0.006	0.006	0.513	0.007	0.008	0.141	0.011	0.265	0.011	0.012	0.004	21
Rangi	Africa	Eastern Africa	0.001	0.008	0.001	0.007	0.007	0.424	0.006	0.008	0.195	0.016	0.306	0.008	0.010	0.002	36
Sambaa	Africa	Eastern Africa	0.002	0.007	0.001	0.006	0.007	0.609	0.007	0.002	0.083	0.011	0.242	0.007	0.013	0.004	18
Pare	Africa	Eastern Africa	0.003	0.006	0.001	0.004	0.011	0.657	0.004	0.005	0.077	0.009	0.206	0.006	0.009	0.002	23
Gabra	Africa	Eastern Africa	0.005	0.028	0.003	0.037	0.006	0.026	0.016	0.002	0.010	0.055	0.754	0.004	0.052	0.002	17
Borana	Africa	Eastern Africa	0.005	0.022	0.003	0.027	0.004	0.055	0.011	0.004	0.019	0.037	0.749	0.006	0.055	0.003	32
Wata	Africa	Eastern Africa	0.005	0.032	0.001	0.040	0.003	0.053	0.011	0.004	0.010	0.032	0.782	0.005	0.021	0.002	6
Rendille	Africa	Eastern Africa	0.002	0.033	0.002	0.028	0.002	0.042	0.005	0.002	0.006	0.036	0.743	0.003	0.089	0.006	28
El Molo	Africa	Eastern Africa	0.002	0.217	0.001	0.022	0.003	0.109	0.004	0.002	0.020	0.015	0.576	0.008	0.020	0.002	16
Yaaku	Africa	Eastern Africa	0.002	0.113	0.001	0.011	0.003	0.098	0.005	0.002	0.026	0.014	0.694	0.006	0.018	0.006	19
Mumonyot Maasai	Africa	Eastern Africa	0.009	0.182	0.001	0.017	0.005	0.156	0.005	0.011	0.030	0.016	0.535	0.007	0.023	0.002	12
Il'gwesi Maasai	Africa	Eastern Africa	0.005	0.166	0.005	0.020	0.002	0.165	0.007	0.007	0.028	0.017	0.558	0.007	0.010	0.002	21
Dorobo	Africa	Eastern Africa	0.003	0.154	0.004	0.017	0.004	0.160	0.005	0.008	0.078	0.012	0.531	0.009	0.012	0.003	10
Tugen	Africa	Eastern Africa	0.007	0.506	0.002	0.025	0.002	0.087	0.005	0.006	0.012	0.015	0.312	0.005	0.012	0.003	22
Samburu	Africa	Eastern Africa	0.003	0.160	0.002	0.031	0.005	0.090	0.005	0.004	0.017	0.023	0.625	0.006	0.027	0.002	18
Marakwet	Africa	Eastern Africa	0.003	0.567	0.001	0.008	0.003	0.053	0.003	0.008	0.018	0.009	0.312	0.007	0.006	0.002	14
Sengwer	Africa	Eastern Africa	0.002	0.531	0.002	0.019	0.003	0.069	0.008	0.004	0.014	0.008	0.323	0.007	0.009	0.002	21

Table S8. Proportion of AACs from the Global Unsupervised Structure Run at K=14

Populations	Continent	Subregion	Fulani	Nilo Saharan	Oceania	Chadic	S.African Khoesan	Niger Kordofanian	East Asian	Hadza	Sandawe	Indian	Cushitic	W. Pygmy	European	Native American	n
Okiek	Africa	Eastern Africa	0.005	0.280	0.001	0.007	0.003	0.105	0.006	0.006	0.034	0.011	0.522	0.005	0.012	0.002	22
Nandi	Africa	Eastern Africa	0.002	0.345	0.001	0.015	0.003	0.247	0.005	0.003	0.014	0.020	0.322	0.008	0.012	0.003	11
Sabaot	Africa	Eastern Africa	0.002	0.400	0.001	0.019	0.004	0.310	0.003	0.010	0.031	0.009	0.192	0.008	0.009	0.003	20
Turkana	Africa	Eastern Africa	0.002	0.361	0.001	0.090	0.007	0.220	0.004	0.006	0.029	0.012	0.245	0.010	0.008	0.002	26
Pokot	Africa	Eastern Africa	0.002	0.544	0.001	0.029	0.006	0.069	0.003	0.004	0.014	0.007	0.306	0.007	0.008	0.001	23
Maasai Ilchamus	Africa	Eastern Africa	0.003	0.296	0.002	0.013	0.004	0.119	0.009	0.005	0.020	0.022	0.477	0.006	0.019	0.005	27
Luo	Africa	Eastern Africa	0.002	0.129	0.003	0.022	0.013	0.680	0.004	0.005	0.066	0.011	0.047	0.012	0.004	0.002	28
Kenya Bantu	Africa	Eastern Africa	0.001	0.118	0.001	0.008	0.016	0.681	0.006	0.003	0.054	0.009	0.064	0.018	0.017	0.003	11
Kikuyu	Africa	Eastern Africa	0.003	0.078	0.001	0.010	0.011	0.433	0.006	0.003	0.062	0.011	0.356	0.010	0.010	0.005	22
Luhya	Africa	Eastern Africa	0.002	0.135	0.002	0.008	0.012	0.664	0.003	0.004	0.080	0.007	0.065	0.010	0.005	0.002	17
Burji	Africa	Eastern Africa	0.006	0.015	0.004	0.010	0.007	0.073	0.007	0.006	0.038	0.032	0.736	0.017	0.045	0.004	24
Konso	Africa	Eastern Africa	0.003	0.008	0.006	0.018	0.011	0.032	0.010	0.008	0.045	0.032	0.764	0.031	0.029	0.002	14
Beta Israel	Africa	Eastern Africa	0.003	0.023	0.001	0.028	0.013	0.030	0.007	0.003	0.012	0.030	0.565	0.007	0.274	0.003	17
Cape Mixed Ancestry	Africa	Southern Africa	0.006	0.005	0.006	0.005	0.247	0.190	0.081	0.003	0.014	0.197	0.031	0.012	0.192	0.010	39
San	Africa	Southern Africa	0.001	0.002	0.001	0.003	0.806	0.028	0.001	0.008	0.011	0.003	0.019	0.114	0.002	0.001	6
!Xun/Khoe	Africa	Southern Africa	0.001	0.007	0.001	0.003	0.689	0.181	0.001	0.003	0.007	0.003	0.007	0.091	0.003	0.001	8
Xhosa	Africa	Southern Africa	0.002	0.004	0.001	0.004	0.367	0.566	0.003	0.002	0.011	0.009	0.008	0.006	0.013	0.003	28
Venda	Africa	Southern Africa	0.002	0.002	0.002	0.006	0.173	0.771	0.003	0.001	0.006	0.012	0.005	0.010	0.005	0.002	13
South Bantu	Africa	Southern Africa	0.002	0.004	0.001	0.004	0.180	0.762	0.002	0.002	0.013	0.005	0.011	0.010	0.005	0.001	8
Chicago			0.028	0.004	0.002	0.012	0.007	0.702	0.012	0.002	0.012	0.036	0.022	0.006	0.149	0.005	15
Pittsburgh			0.006	0.006	0.002	0.015	0.008	0.700	0.009	0.003	0.018	0.045	0.019	0.008	0.155	0.006	21
Baltimore			0.007	0.009	0.003	0.010	0.004	0.735	0.014	0.004	0.018	0.049	0.019	0.012	0.112	0.005	44
North Carolina			0.003	0.004	0.002	0.011	0.004	0.685	0.017	0.004	0.027	0.056	0.016	0.011	0.154	0.005	18
Orcadian	Eurasia	Northern Europe	0.001	0.001	0.002	0.002	0.004	0.007	0.007	0.001	0.003	0.127	0.003	0.002	0.836	0.005	15
Adygei	Eurasia	Eastern Europe	0.001	0.001	0.002	0.002	0.001	0.007	0.034	0.001	0.003	0.244	0.005	0.001	0.692	0.004	17
Russian	Eurasia	Eastern Europe	0.001	0.002	0.002	0.002	0.002	0.007	0.069	0.002	0.003	0.174	0.004	0.002	0.717	0.013	25
Basque	Eurasia	Western Europe	0.001	0.001	0.002	0.002	0.001	0.007	0.011	0.003	0.004	0.048	0.004	0.002	0.907	0.005	24
French	Eurasia	Western Europe	0.001	0.002	0.003	0.002	0.001	0.007	0.016	0.002	0.005	0.101	0.004	0.002	0.848	0.007	28
Italian	Eurasia	Southern Europe	0.002	0.003	0.003	0.002	0.002	0.008	0.014	0.002	0.005	0.086	0.007	0.002	0.862	0.003	21
Sardinian	Eurasia	Southern Europe	0.003	0.004	0.002	0.002	0.002	0.008	0.008	0.002	0.004	0.035	0.006	0.002	0.919	0.003	28
Bedouin	Eurasia	Western Asia	0.003	0.007	0.003	0.014	0.003	0.032	0.008	0.003	0.015	0.047	0.106	0.002	0.754	0.003	46
Druze	Eurasia	Western Asia	0.001	0.004	0.002	0.004	0.001	0.015	0.007	0.002	0.006	0.103	0.019	0.002	0.831	0.002	42
Palestinian	Eurasia	Western Asia	0.006	0.004	0.003	0.013	0.003	0.027	0.012	0.002	0.012	0.094	0.061	0.004	0.757	0.003	46
Temani	Eurasia	Western Asia	0.003	0.002	0.001	0.005	0.003	0.014	0.008	0.003	0.007	0.056	0.120	0.002	0.774	0.002	21
Balochi	Eurasia	Southern Asia	0.002	0.003	0.001	0.003	0.001	0.021	0.008	0.002	0.006	0.503	0.008	0.003	0.435	0.004	24
Brahui	Eurasia	Southern Asia	0.001	0.002	0.001	0.003	0.003	0.024	0.008	0.001	0.004	0.476	0.007	0.002	0.462	0.005	25
Makrani	Eurasia	Southern Asia	0.001	0.003	0.003	0.004	0.006	0.038	0.013	0.003	0.013	0.465	0.014	0.006	0.430	0.003	25
Sindhi	Eurasia	Southern Asia	0.001	0.002	0.005	0.002	0.001	0.032	0.011	0.001	0.005	0.606	0.007	0.003	0.316	0.007	24
Pathan	Eurasia	Southern Asia	0.001	0.002	0.002	0.002	0.001	0.007	0.030	0.001	0.005	0.582	0.006	0.003	0.351	0.007	24
Burusho	Eurasia	Southern Asia	0.001	0.002	0.009	0.002	0.002	0.006	0.098	0.002	0.004	0.555	0.004	0.002	0.300	0.013	25
Hazara	Eurasia	Southern Asia	0.001	0.002	0.003	0.001	0.001	0.006	0.427	0.001	0.004	0.227	0.005	0.002	0.307	0.013	22
Uygur	Eurasia	Eastern Asia	0.001	0.001	0.002	0.001	0.001	0.007	0.454	0.002	0.003	0.235	0.003	0.001	0.280	0.010	10
Kalash	Eurasia	Southern Asia	0.001	0.001	0.001	0.002	0.001	0.004	0.006	0.001	0.002	0.567	0.002	0.001	0.398	0.012	23
Parsi	Eurasia	Southern Asia	0.001	0.002	0.001	0.003	0.001	0.009	0.003	0.001	0.004	0.490	0.006	0.002	0.476	0.002	25
Kashmiri	Eurasia	Southern Asia	0.001	0.002	0.003	0.002	0.002	0.008	0.015	0.002	0.005	0.724	0.006	0.002	0.226	0.003	25
Punjabi	Eurasia	Southern Asia	0.001	0.001	0.002	0.002	0.001	0.007	0.009	0.001	0.004	0.753	0.004	0.002	0.209	0.004	28
Hindi	Eurasia	Southern Asia	0.001	0.002	0.004	0.003	0.002	0.009	0.011	0.002	0.005	0.853	0.006	0.002	0.095	0.006	28
Marwari	Eurasia	Southern Asia	0.001	0.001	0.003	0.003	0.001	0.007	0.009	0.001	0.004	0.929	0.006	0.002	0.030	0.003	25
Gujarat	Eurasia	Southern Asia	0.001	0.002	0.002	0.002	0.001	0.007	0.011	0.002	0.003	0.929	0.005	0.002	0.032	0.002	50
Marathi	Eurasia	Southern Asia	0.001	0.002	0.003	0.002	0.001	0.007	0.008	0.001	0.004	0.867	0.006	0.003	0.092	0.004	26
Kannada	Eurasia	Southern Asia	0.001	0.002	0.003	0.003	0.001	0.008	0.026	0.003	0.005	0.874	0.006	0.002	0.061	0.005	24

Table S8. Proportion of AACs from the Global Unsupervised Structure Run at K=14

Populations	Continent	Subregion	Fulani	Nilo Saharan	Oceania	Chadic	S.African Khoesan	Niger Kordofanian	East Asian	Hadza	Sandawe	Indian	Cushitic	W. Pygmy	European	Native American	n
Konkani	Eurasia	Southern Asia	0.001	0.002	0.001	0.002	0.001	0.006	0.011	0.001	0.004	0.866	0.005	0.002	0.096	0.002	42
Malayalam	Eurasia	Southern Asia	0.001	0.002	0.002	0.002	0.002	0.010	0.020	0.002	0.005	0.901	0.006	0.002	0.041	0.004	25
Tamil	Eurasia	Southern Asia	0.001	0.002	0.004	0.002	0.002	0.008	0.015	0.001	0.005	0.855	0.006	0.002	0.095	0.003	29
Telugu	Eurasia	Southern Asia	0.001	0.002	0.002	0.002	0.002	0.008	0.019	0.001	0.004	0.903	0.005	0.002	0.044	0.004	27
Oriya	Eurasia	Southern Asia	0.001	0.002	0.004	0.002	0.003	0.008	0.049	0.002	0.005	0.861	0.005	0.003	0.050	0.006	26
Bengali	Eurasia	Southern Asia	0.001	0.002	0.005	0.002	0.002	0.009	0.063	0.001	0.006	0.834	0.007	0.001	0.062	0.005	27
Assamese	Eurasia	Southern Asia	0.001	0.003	0.002	0.003	0.002	0.010	0.139	0.001	0.005	0.772	0.006	0.002	0.049	0.004	25
Han	Eurasia	Eastern Asia	0.001	0.002	0.003	0.002	0.001	0.007	0.934	0.001	0.004	0.022	0.005	0.002	0.009	0.007	44
Dai	Eurasia	Eastern Asia	0.001	0.001	0.002	0.002	0.001	0.009	0.923	0.003	0.003	0.034	0.004	0.002	0.009	0.004	10
Daur	Eurasia	Eastern Asia	0.001	0.001	0.003	0.001	0.001	0.005	0.896	0.001	0.003	0.045	0.003	0.002	0.022	0.016	10
Hezhen	Eurasia	Eastern Asia	0.000	0.001	0.002	0.001	0.001	0.005	0.901	0.001	0.002	0.049	0.002	0.001	0.013	0.022	9
Lahu	Eurasia	Eastern Asia	0.001	0.002	0.001	0.002	0.001	0.007	0.920	0.001	0.006	0.032	0.006	0.002	0.017	0.003	8
Miao	Eurasia	Eastern Asia	0.001	0.001	0.007	0.001	0.001	0.004	0.947	0.002	0.002	0.021	0.003	0.002	0.006	0.003	10
Oroqen	Eurasia	Eastern Asia	0.001	0.002	0.003	0.002	0.001	0.005	0.893	0.002	0.007	0.030	0.005	0.001	0.022	0.027	9
She	Eurasia	Eastern Asia	0.001	0.001	0.001	0.001	0.001	0.005	0.963	0.001	0.002	0.011	0.004	0.001	0.005	0.003	10
Tujia	Eurasia	Eastern Asia	0.001	0.001	0.001	0.002	0.002	0.006	0.942	0.001	0.003	0.017	0.004	0.002	0.009	0.008	10
Tu	Eurasia	Eastern Asia	0.002	0.002	0.001	0.003	0.002	0.007	0.866	0.002	0.004	0.068	0.004	0.002	0.026	0.010	10
Xibo	Eurasia	Eastern Asia	0.001	0.001	0.003	0.002	0.002	0.005	0.853	0.001	0.004	0.057	0.004	0.001	0.048	0.019	9
Yi	Eurasia	Eastern Asia	0.001	0.002	0.005	0.001	0.002	0.007	0.910	0.002	0.005	0.044	0.004	0.002	0.009	0.005	10
Mongola	Eurasia	Eastern Asia	0.001	0.002	0.002	0.002	0.001	0.008	0.849	0.001	0.008	0.062	0.004	0.003	0.040	0.019	10
Naxi	Eurasia	Eastern Asia	0.001	0.002	0.003	0.002	0.002	0.008	0.907	0.002	0.004	0.040	0.004	0.002	0.019	0.005	9
Cambodian	Eurasia	Southeastern Asia	0.002	0.002	0.017	0.002	0.004	0.011	0.820	0.001	0.004	0.089	0.014	0.002	0.025	0.007	10
Japanese	Eurasia	Eastern Asia	0.001	0.002	0.002	0.003	0.002	0.007	0.933	0.002	0.004	0.020	0.004	0.002	0.007	0.010	29
Yakut	Eurasia	Eastern Asia	0.001	0.001	0.001	0.002	0.001	0.006	0.798	0.001	0.003	0.036	0.003	0.002	0.123	0.023	25
Melanesian	Oceania	Melanesia	0.001	0.001	0.621	0.001	0.001	0.005	0.339	0.001	0.002	0.017	0.002	0.001	0.005	0.003	11
Papuan	Oceania	Melanesia	0.001	0.001	0.854	0.001	0.001	0.003	0.099	0.001	0.002	0.028	0.002	0.001	0.003	0.003	17
Australian	Oceania	Australia	0.001	0.002	0.440	0.002	0.001	0.005	0.142	0.001	0.005	0.234	0.004	0.001	0.161	0.001	10
Karitiana	America	South America	0.000	0.000	0.000	0.000	0.000	0.002	0.010	0.000	0.001	0.006	0.001	0.000	0.001	0.978	14
Surui	America	South America	0.000	0.000	0.000	0.000	0.000	0.001	0.008	0.000	0.001	0.005	0.001	0.000	0.001	0.981	8
Colombian	America	South America	0.000	0.001	0.001	0.001	0.001	0.004	0.143	0.001	0.001	0.036	0.002	0.002	0.013	0.794	7
Maya	America	Central America	0.001	0.001	0.005	0.002	0.001	0.011	0.177	0.001	0.003	0.025	0.006	0.002	0.100	0.665	21
Pima	America	Central America	0.000	0.001	0.001	0.001	0.000	0.003	0.170	0.001	0.001	0.012	0.002	0.001	0.010	0.799	14
Maximum			0.627	0.567	0.854	0.477	0.806	0.935	0.963	0.671	0.688	0.929	0.788	0.646	0.919	0.981	

Table S9. Proportion of AACs from the Africa Structure Run at K=14

Population Name	Mbugu	Chadic	Cushitic	Eastern Bantu	Nilo Saharan	Saharan /Dogon	Fulani	Non-Bantu NK	S.African Khoesan /Mbuti	Niger Kordofanian	Sandawe	Central Sudanic	Hadza	W.Pygmy	n
Beja_Banuamir	0.001	0.035	0.678	0.005	0.020	0.185	0.007	0.009	0.003	0.039	0.005	0.009	0.002	0.003	23
Beja_Hadandawa	0.001	0.035	0.651	0.004	0.023	0.213	0.010	0.007	0.002	0.035	0.004	0.007	0.002	0.004	19
Mozabite	0.001	0.013	0.352	0.005	0.006	0.510	0.019	0.009	0.003	0.061	0.003	0.011	0.003	0.003	29
Mandinka	0.003	0.045	0.022	0.006	0.008	0.005	0.025	0.373	0.006	0.481	0.007	0.013	0.002	0.007	22
Dogon	0.218	0.009	0.007	0.004	0.006	0.551	0.001	0.007	0.036	0.115	0.038	0.005	0.001	0.002	9
Dioula	0.001	0.050	0.018	0.006	0.006	0.003	0.005	0.138	0.002	0.728	0.006	0.029	0.002	0.004	5
Ashanti	0.002	0.035	0.015	0.009	0.015	0.003	0.004	0.191	0.006	0.671	0.006	0.033	0.003	0.007	15
Brong	0.001	0.043	0.021	0.006	0.010	0.002	0.009	0.222	0.003	0.644	0.004	0.028	0.002	0.004	26
Hausa_Nigeria	0.004	0.086	0.029	0.010	0.015	0.005	0.019	0.044	0.004	0.711	0.008	0.055	0.002	0.007	16
Fulani_Nigeria	0.001	0.070	0.089	0.009	0.021	0.029	0.376	0.020	0.003	0.308	0.005	0.064	0.001	0.003	4
Yoruba	0.001	0.030	0.015	0.006	0.010	0.005	0.003	0.141	0.003	0.748	0.007	0.026	0.002	0.004	25
Yoruba	0.001	0.027	0.021	0.008	0.009	0.003	0.005	0.142	0.006	0.726	0.007	0.036	0.002	0.005	22
Igala	0.001	0.037	0.021	0.007	0.015	0.002	0.002	0.072	0.005	0.788	0.011	0.026	0.003	0.008	17
Igbo	0.002	0.023	0.019	0.007	0.013	0.003	0.004	0.100	0.006	0.760	0.011	0.039	0.002	0.010	28
Gwari	0.002	0.056	0.028	0.007	0.017	0.004	0.011	0.069	0.004	0.725	0.007	0.061	0.002	0.006	22
Bassange	0.001	0.064	0.015	0.006	0.011	0.006	0.005	0.107	0.004	0.742	0.006	0.025	0.003	0.004	20
Koma	0.002	0.037	0.018	0.005	0.082	0.025	0.029	0.013	0.051	0.511	0.013	0.207	0.003	0.003	12
Mada	0.001	0.591	0.023	0.007	0.018	0.002	0.004	0.012	0.004	0.309	0.006	0.014	0.002	0.007	28
Ouldeme	0.001	0.580	0.023	0.007	0.013	0.004	0.005	0.021	0.005	0.311	0.006	0.012	0.002	0.009	26
Giziga	0.003	0.442	0.033	0.008	0.025	0.003	0.007	0.019	0.007	0.390	0.007	0.049	0.002	0.004	24
Mandara	0.001	0.446	0.045	0.009	0.047	0.005	0.023	0.026	0.006	0.347	0.008	0.029	0.002	0.006	26
Kotoko	0.001	0.464	0.066	0.007	0.043	0.016	0.012	0.032	0.005	0.291	0.005	0.042	0.003	0.012	17
Zulgo	0.002	0.559	0.024	0.006	0.019	0.002	0.004	0.008	0.003	0.344	0.004	0.012	0.002	0.011	22
Podokwo	0.002	0.576	0.023	0.008	0.017	0.002	0.004	0.014	0.004	0.312	0.005	0.023	0.003	0.007	30
Massa	0.001	0.405	0.042	0.008	0.090	0.001	0.006	0.014	0.004	0.368	0.008	0.044	0.004	0.004	15
Hausa_Cameroon	0.001	0.182	0.043	0.012	0.016	0.006	0.027	0.030	0.007	0.593	0.006	0.065	0.003	0.011	27
Baggara	0.002	0.302	0.187	0.011	0.091	0.029	0.092	0.020	0.004	0.227	0.007	0.018	0.002	0.008	23
Fulani_Cameroon	0.001	0.224	0.062	0.006	0.020	0.022	0.299	0.028	0.004	0.299	0.007	0.021	0.002	0.005	41
Fulani_Mbororo	0.002	0.027	0.062	0.004	0.007	0.034	0.647	0.022	0.002	0.168	0.008	0.015	0.001	0.002	13
Tupuri	0.002	0.382	0.036	0.009	0.046	0.003	0.004	0.031	0.005	0.420	0.007	0.048	0.002	0.005	22
Baka	0.001	0.013	0.013	0.005	0.009	0.001	0.003	0.006	0.003	0.210	0.005	0.012	0.002	0.718	48
Bakola	0.002	0.014	0.011	0.007	0.008	0.001	0.002	0.006	0.007	0.258	0.012	0.005	0.003	0.664	42
Bedzan	0.001	0.011	0.014	0.008	0.010	0.002	0.004	0.021	0.005	0.427	0.006	0.009	0.003	0.479	17
Mvae	0.002	0.021	0.032	0.010	0.022	0.003	0.003	0.008	0.012	0.808	0.011	0.016	0.003	0.050	24
Ngumba	0.002	0.019	0.031	0.019	0.018	0.002	0.006	0.009	0.008	0.772	0.014	0.018	0.004	0.079	27
Zime	0.002	0.045	0.028	0.022	0.017	0.002	0.003	0.018	0.007	0.740	0.022	0.026	0.004	0.066	30
Bamoun	0.002	0.035	0.027	0.015	0.019	0.002	0.011	0.020	0.007	0.811	0.012	0.024	0.001	0.015	31
Banen	0.003	0.017	0.029	0.031	0.021	0.003	0.004	0.020	0.011	0.817	0.010	0.015	0.003	0.016	25
Bafia	0.002	0.029	0.026	0.014	0.018	0.002	0.005	0.013	0.010	0.824	0.011	0.023	0.004	0.019	30
Lemande	0.002	0.026	0.020	0.015	0.014	0.003	0.004	0.007	0.006	0.862	0.009	0.012	0.004	0.015	26
Batanga	0.004	0.028	0.024	0.035	0.015	0.002	0.005	0.011	0.010	0.806	0.011	0.017	0.002	0.032	20
Iyassa	0.004	0.024	0.032	0.033	0.015	0.012	0.003	0.023	0.009	0.792	0.011	0.011	0.003	0.029	37

Table S9. Proportion of AACs from the Africa Structure Run at K=14

Population Name	Mbugu	Chadic	Cushitic	Eastern Bantu	Nilo Saharan	Saharan /Dogon	Fulani	Non-Bantu NK	S.African Khoesan /Mbuti	Niger Kordofanian	Sandawe	Central Sudanic	Hadza	W.Pygmy	n
Fang	0.002	0.022	0.030	0.012	0.021	0.002	0.004	0.012	0.009	0.803	0.011	0.009	0.004	0.059	19
Mabea	0.001	0.023	0.021	0.010	0.018	0.003	0.003	0.011	0.012	0.800	0.010	0.013	0.002	0.072	13
Yambassa	0.001	0.034	0.025	0.013	0.023	0.003	0.005	0.010	0.012	0.826	0.016	0.017	0.002	0.013	17
Tikar_South	0.005	0.036	0.024	0.020	0.015	0.002	0.007	0.018	0.013	0.785	0.010	0.035	0.003	0.026	21
Tikar_North	0.003	0.038	0.030	0.014	0.018	0.002	0.006	0.030	0.012	0.771	0.014	0.043	0.002	0.018	13
Ntumu	0.002	0.030	0.023	0.019	0.013	0.004	0.003	0.005	0.007	0.848	0.010	0.007	0.006	0.024	11
Bulu	0.002	0.033	0.025	0.018	0.018	0.002	0.004	0.014	0.010	0.801	0.012	0.021	0.003	0.036	22
Eton	0.005	0.033	0.019	0.008	0.029	0.003	0.015	0.025	0.003	0.787	0.008	0.026	0.005	0.034	4
Wimbum	0.002	0.023	0.038	0.015	0.020	0.002	0.003	0.019	0.008	0.812	0.018	0.026	0.003	0.011	15
Batie	0.001	0.029	0.019	0.028	0.019	0.002	0.004	0.034	0.012	0.803	0.010	0.027	0.001	0.011	16
Ewondo	0.001	0.045	0.017	0.015	0.013	0.001	0.002	0.006	0.004	0.827	0.010	0.014	0.001	0.043	3
Kanuri	0.004	0.342	0.056	0.014	0.058	0.004	0.024	0.034	0.006	0.390	0.010	0.047	0.003	0.007	31
Biaka	0.001	0.019	0.014	0.010	0.009	0.002	0.002	0.024	0.015	0.167	0.007	0.022	0.001	0.706	23
Mbum	0.002	0.092	0.015	0.005	0.020	0.002	0.005	0.019	0.005	0.589	0.006	0.234	0.002	0.004	13
Yakoma	0.003	0.133	0.019	0.016	0.031	0.003	0.004	0.036	0.011	0.584	0.015	0.093	0.003	0.048	6
Gbaya	0.001	0.090	0.029	0.010	0.033	0.001	0.004	0.016	0.006	0.643	0.003	0.151	0.002	0.012	15
Laka	0.001	0.102	0.022	0.007	0.015	0.002	0.006	0.035	0.006	0.533	0.005	0.254	0.002	0.008	33
Ngambaye	0.002	0.170	0.022	0.009	0.025	0.003	0.006	0.021	0.004	0.514	0.005	0.213	0.002	0.006	30
Kaba	0.001	0.098	0.026	0.011	0.015	0.003	0.004	0.025	0.004	0.541	0.006	0.259	0.002	0.006	27
Bulala	0.001	0.438	0.099	0.005	0.155	0.008	0.015	0.029	0.005	0.185	0.008	0.044	0.002	0.006	15
Kanembou	0.002	0.345	0.179	0.008	0.081	0.016	0.025	0.018	0.008	0.256	0.006	0.049	0.002	0.006	5
Sara_various	0.001	0.158	0.035	0.008	0.036	0.003	0.006	0.024	0.004	0.548	0.007	0.162	0.002	0.007	27
Baluba	0.001	0.018	0.011	0.024	0.022	0.002	0.002	0.055	0.009	0.772	0.020	0.040	0.002	0.024	6
Kongo	0.003	0.029	0.023	0.059	0.013	0.003	0.003	0.018	0.007	0.795	0.015	0.015	0.002	0.017	17
Barega	0.001	0.040	0.022	0.031	0.023	0.001	0.003	0.058	0.048	0.686	0.055	0.015	0.003	0.014	4
Mbuti	0.001	0.010	0.020	0.004	0.033	0.001	0.002	0.026	0.606	0.055	0.010	0.012	0.004	0.216	13
Tutsi/Hutu	0.002	0.014	0.186	0.144	0.039	0.003	0.002	0.022	0.011	0.526	0.012	0.019	0.002	0.017	8
Dinka	0.002	0.253	0.054	0.007	0.552	0.001	0.003	0.015	0.004	0.062	0.007	0.031	0.003	0.006	17
Nuer	0.003	0.234	0.078	0.008	0.547	0.002	0.004	0.015	0.006	0.068	0.007	0.018	0.002	0.008	18
Shilluk	0.006	0.224	0.062	0.007	0.518	0.003	0.004	0.007	0.005	0.109	0.006	0.039	0.005	0.005	15
Nyimang	0.003	0.255	0.154	0.010	0.393	0.011	0.011	0.011	0.022	0.086	0.008	0.028	0.001	0.006	12
Hadza	0.002	0.013	0.071	0.054	0.012	0.003	0.003	0.007	0.005	0.149	0.011	0.009	0.657	0.005	63
Sandawe	0.001	0.014	0.223	0.017	0.012	0.006	0.004	0.008	0.006	0.090	0.600	0.011	0.004	0.005	51
Iraqw	0.002	0.016	0.767	0.038	0.012	0.002	0.004	0.011	0.006	0.065	0.042	0.014	0.008	0.012	46
Burunge	0.002	0.014	0.394	0.192	0.015	0.004	0.003	0.013	0.012	0.153	0.152	0.017	0.011	0.018	22
Fiome	0.003	0.026	0.669	0.083	0.022	0.005	0.007	0.011	0.009	0.082	0.045	0.020	0.011	0.008	22
Mbugu	0.320	0.021	0.392	0.042	0.021	0.003	0.003	0.007	0.004	0.160	0.009	0.010	0.002	0.004	22
Maasai	0.002	0.049	0.468	0.065	0.159	0.008	0.004	0.009	0.011	0.175	0.025	0.011	0.007	0.009	36
Akie	0.006	0.022	0.602	0.021	0.049	0.007	0.005	0.009	0.011	0.139	0.106	0.008	0.005	0.008	23
Datog	0.002	0.044	0.632	0.042	0.086	0.004	0.007	0.014	0.009	0.105	0.020	0.022	0.006	0.008	54
Turu	0.002	0.023	0.329	0.150	0.017	0.005	0.007	0.013	0.015	0.274	0.134	0.011	0.008	0.012	32
Sukuma	0.002	0.015	0.111	0.255	0.021	0.003	0.004	0.005	0.015	0.515	0.014	0.022	0.012	0.006	10

Table S9. Proportion of AACs from the Africa Structure Run at K=14

Population Name	Mbugu	Chadic	Cushitic	Eastern Bantu	Nilo Saharan	Saharan /Dogon	Fulani	Non-Bantu NK	S.African Khoesan /Mbuti	Niger Kordofanian	Sandawe	Central Sudanic	Hadza	W.Pygmy	n
Gogo	0.001	0.016	0.221	0.192	0.017	0.002	0.004	0.017	0.019	0.374	0.108	0.013	0.007	0.007	13
Mbugwe	0.004	0.017	0.253	0.265	0.016	0.003	0.004	0.012	0.012	0.349	0.031	0.019	0.005	0.011	21
Rangi	0.002	0.022	0.299	0.251	0.013	0.006	0.004	0.014	0.013	0.289	0.059	0.013	0.005	0.010	36
Sambaa	0.021	0.014	0.200	0.230	0.016	0.005	0.004	0.007	0.009	0.452	0.017	0.016	0.002	0.007	18
Pare	0.025	0.015	0.160	0.203	0.014	0.003	0.004	0.007	0.010	0.516	0.020	0.014	0.003	0.007	23
Gabra	0.002	0.042	0.765	0.006	0.056	0.037	0.007	0.007	0.010	0.038	0.009	0.014	0.002	0.005	17
Borana	0.004	0.041	0.734	0.016	0.054	0.024	0.006	0.012	0.008	0.058	0.017	0.016	0.003	0.007	32
Wata	0.003	0.060	0.732	0.008	0.065	0.007	0.010	0.015	0.006	0.059	0.012	0.015	0.003	0.007	6
Rendille	0.001	0.043	0.758	0.005	0.072	0.033	0.005	0.010	0.004	0.046	0.004	0.013	0.002	0.004	28
El_Molo	0.001	0.047	0.532	0.011	0.249	0.006	0.004	0.014	0.007	0.089	0.019	0.012	0.002	0.007	16
Yaaku	0.001	0.025	0.624	0.036	0.180	0.004	0.005	0.009	0.006	0.084	0.012	0.008	0.002	0.005	19
Maasai_Mumonyot	0.001	0.037	0.498	0.029	0.224	0.012	0.008	0.017	0.010	0.113	0.028	0.009	0.007	0.007	12
Maasai_Il'gwesi	0.003	0.037	0.505	0.063	0.201	0.006	0.007	0.012	0.004	0.122	0.013	0.014	0.005	0.007	21
Dorobo	0.002	0.044	0.490	0.062	0.187	0.004	0.006	0.008	0.010	0.116	0.038	0.015	0.007	0.011	10
Tugen	0.002	0.037	0.324	0.011	0.500	0.005	0.008	0.012	0.004	0.060	0.007	0.018	0.005	0.007	22
Samburu	0.001	0.049	0.580	0.025	0.203	0.012	0.002	0.012	0.009	0.069	0.012	0.015	0.003	0.007	18
Marakwet	0.003	0.015	0.314	0.014	0.563	0.002	0.003	0.006	0.005	0.043	0.011	0.006	0.007	0.007	14
Sengwer	0.001	0.030	0.325	0.012	0.530	0.004	0.004	0.007	0.005	0.052	0.010	0.010	0.004	0.007	21
Okiek	0.002	0.017	0.502	0.017	0.293	0.003	0.009	0.008	0.007	0.097	0.026	0.008	0.005	0.006	22
Nandi	0.001	0.027	0.311	0.058	0.353	0.005	0.006	0.010	0.005	0.185	0.009	0.018	0.002	0.009	11
Sabaot	0.001	0.032	0.203	0.067	0.391	0.003	0.004	0.017	0.008	0.215	0.018	0.025	0.008	0.008	20
Turkana	0.002	0.107	0.248	0.050	0.364	0.004	0.004	0.022	0.012	0.125	0.013	0.033	0.005	0.010	26
Pokot	0.001	0.039	0.306	0.013	0.543	0.003	0.003	0.007	0.008	0.050	0.009	0.008	0.003	0.007	23
Maasai Ilchamus	0.002	0.031	0.459	0.032	0.307	0.013	0.006	0.016	0.006	0.090	0.012	0.016	0.004	0.007	27
Luo	0.002	0.045	0.044	0.236	0.094	0.002	0.005	0.013	0.013	0.501	0.011	0.020	0.004	0.010	28
Bantu_Kenya	0.002	0.024	0.069	0.222	0.093	0.006	0.004	0.043	0.022	0.476	0.013	0.013	0.002	0.012	11
Kikuyu	0.007	0.034	0.315	0.163	0.073	0.004	0.009	0.020	0.013	0.306	0.021	0.023	0.002	0.010	22
Luhya	0.002	0.026	0.048	0.272	0.097	0.003	0.006	0.008	0.010	0.484	0.018	0.017	0.003	0.008	17
Burji	0.001	0.025	0.710	0.030	0.034	0.016	0.010	0.019	0.017	0.068	0.021	0.024	0.005	0.021	24
Konso	0.005	0.036	0.731	0.012	0.033	0.015	0.002	0.015	0.017	0.041	0.031	0.024	0.007	0.031	14
Beta_Israel	0.003	0.022	0.681	0.006	0.027	0.170	0.006	0.009	0.014	0.030	0.009	0.011	0.002	0.009	17
Cape Mixed Ancestry	0.002	0.008	0.253	0.017	0.006	0.264	0.007	0.018	0.262	0.128	0.009	0.011	0.003	0.011	39
San	0.001	0.005	0.030	0.003	0.003	0.002	0.001	0.003	0.873	0.015	0.005	0.003	0.006	0.050	6
!Xun/Khoe	0.004	0.007	0.017	0.003	0.011	0.002	0.002	0.011	0.777	0.089	0.010	0.013	0.004	0.049	8
Xhosa	0.004	0.012	0.028	0.035	0.010	0.013	0.004	0.005	0.412	0.454	0.010	0.007	0.002	0.006	28
Venda	0.014	0.017	0.015	0.036	0.006	0.014	0.004	0.005	0.192	0.670	0.005	0.009	0.001	0.011	13
Bantu_South	0.001	0.016	0.021	0.070	0.011	0.002	0.003	0.024	0.208	0.621	0.007	0.007	0.002	0.007	8
Max	0.320	0.591	0.767	0.272	0.563	0.551	0.647	0.373	0.873	0.862	0.600	0.259	0.657	0.718	

References

S1. S. A. Miller, D. D. Dykes, H. F. Polesky, *Nucleic Acids Res* 16, 1215 (Feb 11, 1988).
S2. http://www.tau.ac.il/medicine/NLGIP/catalog.htm.
S3. http://research.marshfieldclinic.org/genetics/GenotypingData_Statistics/humanDiversityPanel.asp.
S4. N. A. Rosenberg *et al.*, *Science* 298, 2381 (Dec 20, 2002).
S5. N. A. Rosenberg *et al.*, *PLoS Genet* 1, e70 (Dec, 2005).
S6. N. A. Rosenberg *et al.*, *PLoS Genet* 2, e215 (Dec, 2006).
S7. M. P. Epstein, W. L. Duren, M. Boehnke, *Am J Hum Genet* 67, 1219 (Nov, 2000).
S8. N. A. Rosenberg, *Ann Hum Genet* 70, 841 (Nov, 2006).
S9. N. Patterson, A. L. Price, D. Reich, *PLoS Genet* 2, e190 (Dec, 2006).
S10. http://www.graphviz.org.
S11. J. Reynolds, B. S. Weir, C. C. Cockerham, *Genetics* 105, 767 (Nov, 1983).
S12. N. Takezaki, M. Nei, *Genetics* 178, 385 (Jan, 2008).
S13. F. Balloux, N. Lugon-Moulin, *Mol Ecol* 11, 155 (Feb, 2002).
S14. http://www.evolution.genetics.washington.edu/phylip.html.
S15. M. Slatkin, *Genetics*, 457 (1995).
S16. D. B. Goldstein, A. Ruiz Linares, L. L. Cavalli-Sforza, M. W. Feldman, *Proc Natl Acad Sci U S A* 92, 6723 (Jul 18, 1995).
S17. D. Paetkau, L. P. Waits, P. L. Clarkson, L. Craighead, C. Strobeck, *Genetics* 147, 1943 (Dec, 1997).
S18. R. D. Page, *Comput Appl Biosci* 12, 357 (Aug, 1996).
S19. M. K. Gonder, H. M. Mortensen, F. A. Reed, A. de Sousa, S. A. Tishkoff, *Mol Biol Evol* 24, 757 (Mar, 2007).
S20. J. Z. Li *et al.*, *Science* 319, 1100 (Feb 22, 2008).
S21. Y. S. Chen *et al.*, *Am J Hum Genet* 66, 1362 (Apr, 2000).
S22. S. A. Tishkoff *et al.*, *Mol Biol Evol* 24, 2180 (Oct, 2007).
S23. A. M. Valdes, M. Slatkin, N. B. Freimer, *Genetics* 133, 737 (1993).
S24. P. O. Lewis, D. V. Zaykin. (2001).
S25. M. Nei, *Molecular Evolutionary Genetics*. (Columbia University Press, New York, 1987).
S26. M. Kimura, T. Ohta, *Genetics* 75, 199 (1973).
S27. Z. A. Szpiech, M. Jakobsson, N. A. Rosenberg, *Bioinformatics* 24, 2498 (Nov 1, 2008).
S28. B. S. Weir, *Genetic data analysis II: Methods for discrete population genetic data.*, (Sinauer Associates, Sunderland, MA, ed. 2nd 1996).
S29. http://www.ethnologue.com.
S30. H. M. Cann *et al.*, *Science* 296, 261 (Apr 12, 2002).
S31. S. Ramachandran *et al.*, *Proc Natl Acad Sci U S A* 102, 15942 (Nov 1, 2005).
S32. O. Francois, S. Ancelet, G. Guillot, *Genetics* 174, 805 (Oct, 2006).
S33. http://genepath.med.harvard.edu/~reich/Software.htm.
S34. http://pritch.bsd.uchicago.edu/structure.html.
S35. J. K. Pritchard, M. Stephens, P. Donnelly, *Genetics* 155, 945 (Jun, 2000).
S36. J. S. Friedlaender *et al.*, *PLoS Genet* 4, e19 (Jan, 2008).
S37. D. Falush, M. Stephens, J. K. Pritchard, *Genetics* 164, 1567 (Aug, 2003).

S38. M. Jakobsson, N. A. Rosenberg, *Bioinformatics* 23, 1801 (Jul 15, 2007).

S39. http://rosenberglab.bioinformatics.med.umich.edu/distruct.html.

S40. S. S. Shapiro, M. B. Wilk, *Biometrika* 52, 591 (1965).

S41. R. R. Sokal, F. J. Rohlf, *Introduction to Biostatistics*. (Freeman and Company, San Francisco, ed. 2nd, 1987), pp. xii, 363

S42. SAS. (SAS Institute Inc, Cary, NC, 2004).

S43. C. Ehret, *A Comparative Historical Reconstruction of Proto-Nilo-Saharan.* . Sprache und Geschichte in Afrika. SUGIA-Supplements (Rüdiger Köppe Verlag, Cologne, 2001), vol. 12, pp. 633.

S44. C. Ehret, *Reconstructing Proto-Afroasiatic (Proto-Afrasian) : vowels, tone, consonants, and vocabulary*. University of California publications in linguistics v. 126 (University of California Press, Berkeley, 1995), pp. xvii, 557.

S45. P. Newman, *Journal of African Languages.* 5, 5 (1977).

S46. M. L. Bender, *Anthropological Linguistics* 13, 165 (1971).

S47. C. Ehret, *Sprache und Geschichte in Afrika* 8, 7 (1987).

S48. C. Ehret, in *In Hot Pursuit of Language in Prehistory,* J. Bengtson, Ed. (John Benjamins, Amsterdam, 2008), pp. 140-160.

S49. C. Ehret, *Studies in African Linguistics* 22, 211 (1991).

S50. C. Ehret, *The historical reconstruction of Southern Cushitic phonology and vocabulary*. Kölner Beiträge zur Afrikanistik 5. Bd. (Reimer, Berlin, 1980), pp. 407.

S51. K. Williamson, R. Blench, in *African Languages: an introduction,* B. Heine, D. Nurse, Eds. (Cambridge University Press, Cambridge, 2000), pp. 11-42.

S52. http://www.history.ucla.edu/ehret/.

S53. Y. Bastin, A. Coupez, M. Mann, *Continuity and divergence in the Bantu languages : perspectives from a lexicostatistic study*. Annalen. Menswetenschappen = Annales. Sciences humaines (Musée royal de l'Afrique centrale, Tervuren, Belgium, 1999), pp. vi, 225.

S54. W. A. A. Wilson, in *in The Niger-Congo Languages,* J. Bendor-Samuel, Ed. (University Press of America, Lanham, MD, 1989), pp. 81-104.

S55. T. Naden, in *in Niger-Congo Languages* J. Bendor-Samuel, Ed. (University Press of America, Lanham, MD, 1989), pp. 140-168.

S56. K. A. Klieman, *"The Pygmies were our compass" : Bantu and Batwa in the history of west central Africa, early times to c. 1900 C.E.* Social history of Africa (Heinemann, Portsmouth, NH, 2003), pp. xxxiv, 253.

S57. D. E. Saxon, in *in The Archaeological and Linguistic Reconstruction of African History,* C. Ehret, M. Posnansky, Eds. (University of California, Berkeley, 1982), pp. 39-56.

S58. R. G. Armstrong, *The Study of West African Languages* (Ibadan University Press, Ibadan, 1964), pp. 74.

S59. T. L. E. Wilson, Ph.D. diss, University of California at Los Angeles (1980).

S60. R. Vossen, *The Eastern Nilotic Languages* (Dietrich Reimer Verlag, Berlin, 1982), pp. 512.

S61. R. Thelwall, in *in The Archaeological and Linguistic Reconstruction of African History,* C. Ehret, M. Posnansky, Eds. (University of California, Berkeley, 1982), pp. 39-56.

S62. C. Ehret, in *Culture History in the Southern Sudan*, J. Mack, P. Robertshaw, Eds. (British Institute in Eastern Africa, Nairobi, 1983), pp. 19-48.
S63. B. Heine, *The Kuliak Languages*. (EAPH, Nairobi, 1976), pp. x, 79.
S64. J. Distefano, Ph.D. diss., University of California (1985).
S65. D. E. Saxon, Ph.D diss., University of California (1980).
S66. R. Thelwall, Ph.D. diss., New University of Ulster (1981).
S67. C. Ehret, in *Archaeology of Early Northeastern Africa*, M. Chlodnicki, K. Kroeper, Eds. (Poznan Archaeological Museum, Studies in African Archaeology, Poznan, Poland, 2006), vol. 9, pp. 1019-1055.
S68. A. Militarev, in *Time Depth in Historical Linguistics*, C. Renfrew, A. McMahon, L. Trask, Eds. (McDonald Institute for Archaeological Research, Cambridge, 2000), vol. 1, pp. 267-307.
S69. C. Ehret, *Ufahamu* 26, 54 (1997).
S70. C. Ehret, in *The Archeological and Linguistic Reconstruction of African History*, C. Ehret, M. Posnansky, Eds. (University of California Press, Berkeley, 1982), pp. 158-181.
S71. S. M. Embleton, Ed., *Statistics in Historical Linguistics*, (Studienverlag Brockmeyer, Bochum, Germany, 1986), pp. viii, 194.
S72. E. T. Wood *et al.*, *Eur J Hum Genet* 13, 867 (2005).
S73. J. H. Greenberg, *Studies in African Linguistic Classification* (Compass Press Bradford, CT, 1955).
S74. C. Ehret, *The civilizations of Africa : a history to 1800*. (University Press of Virginia, Charlottesville, 2002), pp. xii, 480.
S75. C. Ehret, S. O. Keita, P. Newman, *Science* 306, 1680; author reply 1680 (Dec 3, 2004).
S76. L. Lim, Ph.D. diss., Brown University (1992).
S77. T. Güldemann, M. Stoneking, *Annu Rev. Anthropol.* 37, 93 (2008).
S78. B. Sands, in *Language, identity and conceptualization among the Khoisan*, M. Schladt, Ed. (Rudiger Kupper Verlag, Koln, 1998), vol. Bd 15, pp. 266 - 283.
S79. E. D. Elderkin, *Sprache und Geschichte in Afrika* 4, 67 (1982).
S80. M. Kimmel *et al.*, *Genetics* 148, 1921 (Apr, 1998).
S81. J. P. King, M. Kimmel, R. Chakraborty, *Mol Biol Evol* 17, 1859 (Dec, 2000).
S82. M. C. Campbell, S. A. Tishkoff, *Annu Rev Genomics Hum Genet* 9, 403 (2008).
S83. C. Ehret, *An African Classical Age : Eastern and Southern Africa in World History, 1000 B.C. to A.D. 400*. (University Press of Virginia, Charlottesville, 1998), pp. xvii, 354.
S84. S. H. Ambrose, *Archaeology and linguistic reconstructions of history in Eastern Africa*. C. Ehret, M. Posnansy, Eds., The Archaeological and Linguistic Reconstruction of African History (University of California Press, Berkeley, 1982), pp. xii, 299.
S85. R. Scozzari *et al.*, *Am J Hum Genet* 65, 829 (Sep, 1999).
S86. O. Semino, A. S. Santachiara-Benerecetti, F. Falaschi, L. L. Cavalli-Sforza, P. A. Underhill, *Am J Hum Genet* 70, 265 (Jan, 2002).
S87. L. L. Cavalli-Sforza, A. Piazza, P. Menozzi, *History and Geography of Human Genes*. (Princeton University Press, Princeton, 1994).
S88. P. Forster, S. Matsumura, *Science* 308, 965 (May 13, 2005).

S89. M. Jakobsson *et al.*, *Nature* 451, 998 (Feb 21, 2008).
S90. C. Ehret, *The International Journal of African Historical Studies* 34, 5 (2001).
S91. J. L. Newman, *The Peopling of Africa*. (Yale University Press, New Haven and London, 1997), pp. xiv, 235.
S92. M. Mous, in *Mixed languages: 15 case studies in language intertwining*, P. Bakker, M. Mous, Eds. (IFOTT, Amsterdam, 1994), pp. 175-201.
S93. C. Ehret, *Southern African Humanities* 30, 7 (2008).
S94. C. Ehret, in *African Historical Demography,* C. Fyfe, D. McMaster, Eds. (University of Edinburgh Centre of African Studies, Edinburgh, 1981), vol. 2, pp. 802.
S95. C. Ehret, in *Early Human Kinship: From Sex to Social Reproduction*, N. J. Allen, H. Callan, R. Dunbar, W. James, Eds. (Blackwell, Oxford, UK, 2008), pp. 200-231, 259-269.
S96. V. Cerny, A. Salas, M. Hajek, M. Zaloudkova, R. Brdicka, *Ann Hum Genet* 71, 433 (Jul, 2007).
S97. H. Y. Hassan, P. A. Underhill, L. L. Cavalli-Sforza, M. E. Ibrahim, *Am J Phys Anthropol* 137, 316 (Nov, 2008).
S98. R. E. Bereir *et al.*, *Eur J Hum Genet* 15, 1183 (Nov, 2007).
S99. C. Ehret, in *West African Linguistics: Studies in Honor of Russell G. Schuh,* P. Newman, L. Hyman, Eds. (Ohio State University, Columbus, 2006), pp. 56-66.
S100. V. Cerny *et al.*, *BMC Evol Biol* 9, 63 (Mar 23, 2009).
S101. V. Cerny *et al.*, *Hum Biol* 78, 9 (Feb, 2006).
S102. C. A. Mulcare *et al.*, *Am J Hum Genet* 74, 1102 (Jun, 2004).
S103. C. Bailey, *Journal of the Economic and Social History of the Orient* 28, 20 (1985).
S104. Y. Lev, *International Journal of Middle East Studies* 19, 337 (1987).
S105. A. Levanoni, *Studia Islamica* 72, 121 (1990).
S106. B. J. Walker, *Near Eastern Archaeology* 62, 202 (1999).
S107. A. F. C. Holl, T. E. Levy, *The Biblical Archaeologist* 56, 166 (1993).
S108. T. E. Levy, A. F. C. Holl, *Journal of Anthropological Archaelogy* 21, 83 (2002).
S109. B. Lembezat, Ed., *Les populations païennes du Nord- Cameroun et de l'Adamaoua*, (Press Universitaires de France, Paris, 1961).
S110. C. Ehret, *Ethiopians and East Africans : the problem of contacts*. Nairobi Historical Studies (East African Publishing House, Nairobi, 1974), pp. 95.
S111. S. H. Ambrose, *Sprache und Geschichte in Afrika* 7.2, 11 (1986).
S112. S. Feierman, *The Shambaa Kingdom*. (University of Wisconsin Press, Madison, 1974).
S113. S. G. Thomason, *Contact Languages: A wider perspective*. S. G. Thomason, Ed., (John Benjamins Publishing Company, Amsterdam/Philadelphia, 1997), pp. xiii, 506.
S114. R. Blackburn, *In the land of milk and honey: Okiek adaptations to their forests and neighbors*. S. Kent, Ed., Cultural diversity among twentieth-century foragers: an African perspective (Cambridge University Press, Cambridge, 1996), pp. xiv, 344.
S115. L. Cronk, *Ethnology* 41, 27 (2002).
S116. S. A. Tishkoff *et al.*, *Nat Genet* 39, 31 (Jan, 2007).

S117. L. Cronk, *From Mukogodo to Maasai: Ethnicity and Cultural Change in Kenya*. (Westview Press, Boulder CO, 2004).
S118. C. Ehret, in *in Semino-Hamitic Festschrift for*, A. B. Dolgopolsky, G. Jungraithmayr, G. Takacs, Eds. (Dietrich Reimer, Berlin, 2008), pp. xii, 378.
S119. P. Lieberman, R. McCarthy, *Expedition* 49, 15 (2007).
S120. C. Ehret, in *in Origins: The Story of the Emergence of Humans and Humanity in Africa* G. Blundell, Ed. (Double Storey, Cape Town, 2006), pp. 45,47.
S121. S. Bahuchet, *La rencontre des agriculteurs : les Pygmées parmi les peuples d'Afrique centrale*. Histoire d'une civilisation forestière (Peeters, Paris, 1993), pp. 173
S122. S. Bahuchet, *Pygmées de Centrafrique : études ethnologiques, historiques et linguistiques sur les Pygmées "Ba. Mbenga" (aka/baka) du nord-ouest du bassin congolais*. Etudes pygmées 3 (SELAF, Paris, 1979), pp. 179
S123. H. Terashima, *African Study Monographs* 28, 7 (2003).
S124. L. L. Cavalli-Sforza, *Proc Natl Acad Sci U S A* 94, 7719 (Jul 22, 1997).
S125. S. Wang *et al.*, *PLoS Genet* 3, e185 (Nov, 2007).
S126. K. Hunley *et al.*, *PLoS Genet* 4, e1000239 (Oct, 2008).
S127. E. M. Belle, G. Barbujani, *Am J Phys Anthropol* 133, 1137 (Aug, 2007).
S128. W. E. Welmers, *African language structures*. (University of California Press, Berkeley, 1974), pp. xi, 488
S129. E. J. Parra *et al.*, *Am J Hum Genet* 63, 1839 (Dec, 1998).
S130. A. Salas *et al.*, *Am J Phys Anthropol* 128, 855 (Dec, 2005).
S131. J. M. Lind *et al.*, *Hum Genet* 120, 713 (Jan, 2007).
S132. M. W. Smith *et al.*, *Am J Hum Genet* 74, 1001 (May, 2004).
S133. E. J. Parra *et al.*, *Am J Phys Anthropol* 114, 18 (Jan, 2001).
S134. http://www.slavevoyages.org/tast/index.faces.
S135. G. T. Nurse, J. S. Weiner, T. Jenkins, *The Peoples of Southern Africa and Their Affinities*. (Oxford University Press, New York, 1985), pp. 409.
S136. J. S. Marais, *The Cape Coloured People, 1652-1937* (Witwatersrand University Press, Johannesburg, 1968), pp. 296.
S137. P. Mellars, *Proc Natl Acad Sci U S A* 103, 9381 (Jun 20, 2006).
S138. S. McBrearty, A. Brooks, *Journal of Human Evolution* 39, 453 (2000).
S139. C. S. Henshilwood *et al.*, *Science* 295, 1278 (Feb 15, 2002).